Responses to referee comments of second (revised) manuscript version

Markus Schartau et al.

Correspondence to: Markus Schartau (mschartau@geomar.de) and Phil Wallhead (philip.wallhead@niva.no)

General comments by the referee:

The authors performed a very good work in revising the manuscript. The big issue of the neglected sequential data assimilation schemes has been widely resolved. Also, the manuscript appears more balanced with regard to presenting examples of the authors' own work and those by others.

I still see some minor issues, which should be resolved before I can recommend the publication of the manuscript. Unfortunately, the manuscript still shows a clear preference against sequential data assimilation methods. The major disadvantage of the manuscript seems to lie in the fact that the group of authors does not include any scientist who applies sequential data assimilation. To this end, the apparent goal of being a fully comprehensive review and synthesis is not reached. This, however should not prevent the publication, because the manuscript is otherwise a very comprehensive, but also extremely long, review.

Author's response: We very much appreciate the time and effort that the referee has put into the evaluation of our revised manuscript. We regret that she/he has still the impression of a bias against sequential methods, but we do feel that the manuscript has achieved a level of balance that is reasonable for a review paper. Each of the referee's comments below have been addressed and we have made modifications where considered necessary.

Main comments

***Comment* 1**:

Abstract, last sentence: As I already commented in my first review, the last sentence of the abstract - the recommendation to balance level of sophistication of the model and the data assimilation treatment - is not a clear result of the study. The authors only responded that from literature review they had the 'general impression' and hence 'felt' that it would be helpful to report their impression to the readers. For me, this is pretty surprising response, because, as the statement itself, the response is not clearly based on scientific insight but rather represents an opinion. On the other hand, Section 6.2 does in fact discuss the model complexity in a scientific way. Thus, I can still only recommend to reformulate the last sentence in a way that it is not just a statement of opinion (which is not suitable for an abstract of a scientific paper), but shows that it is a result of literature review and studies on model complexity.

Author's response: Perhaps we were too careful (in our response) in expressing what we have learned from reading a large number of papers and the phrase "this was a general impression" does not actually reflect the insight we gained. We apologise for being vague in our response text.

The last sentence in the abstract includes a recommendation that is hardly controversial. The sentence follows after stating: "Our review discloses many facets of parameter identification, as we found many commonalities between the objectives of different approaches, but scientific insight differed between studies." In this context (logical sequence) a reader will understand that this recommendation is an overall outcome of our extensive literature review and not just an opinion. It is partially backed up with citations in Sect. (6.2). It is confirmed by our finding that latest studies that involve highly sophisticated data assimilation methods for parameter estimation (also along with state estimation) do not comply with latest developments in modeling phytoplankton growth and plankton interaction, e.g. as described in Sect. (3). Conversely, studies that focused on improving mathematical representations of plankton dynamics have often neglected aspects of parameter identification.

The above described development within the scientific community is evident and can be tested by screening through the literature we refer to in our review. The problem is addressed and discussed in our last paragraph (Sect. 10.3). In conclusion, inference with respect to model development and parameter identification could be further improved if knowledge and expertise of the different scientific communities would be better merged (i.e. communicated). It is for these above reasons that we do not want to follow the referee's request to rephrase the last sentence of the abstract.

***Comment* 2**:

In the new section 2.2.2 on sequential methods, the authors state that sequential data assimilation approached can be expedient in cases that assimilating all data at different times is 'computationally impractical'. This statement is misleading, because sequential schemes can be used (and actually have been used) also in cases where a variational scheme could still be applied (There are various cases in the literature). Even more, a sequential scheme can also have a 4D-component, when a smoother method is applied. Rather, it depends on the scientific question whether a sequential data assimilation method results in the required result and the manuscript itself provides examples for these cases. In this respect also the statement that sequential data assimilation method '...break the large integration problem ... into a number of smaller problems...' (page 6, lines 69-70) is incorrect. The sequential methods reformulate the data assimilation problem into a sequential form. This, however, does not necessarily 'break' it. Furthermore, the sequential methods are not necessarily a 'sequential approximation' (page 6, line 74). Please revise the first paragraph of Section 2.2.2 accordingly.

Author's response: Section (2.2.2) begins with: "In some problems, assimilating all the data at once from all available sampling times can be computationally impractical". This is a natural introductory sentence given the previous section, which introduces the posterior density (Eq. 6) and the likelihood (Eq. 7) as the basis of theoretically optimal estimates. If we had an "ideal" computational algorithm we would simply apply it to evaluate the posterior density or the likelihood over the space of model parameters.

As we see it, from a theoretical viewpoint (appropriate to a Theoretical Background section), the best motivation for sequential methods is to make computationally practical approximations of the likelihood or posterior for cases involving large and/or stochastic models. When we consider a stochastic model, Eq. (7) becomes a very high dimensional integral over the dynamical model state at all grid points and all times. But even if the integral in Eq. (7) is feasible (e.g. for a deterministic model) it may still be desirable to avoid having to redo it every time new data come in, or to avoid having to rerun over the entire simulation period for each new trial parameter value. So we go on to say: "This is particularly likely for models with stochastic dynamics ($\boldsymbol{\eta} \neq 0$ in Eq. 1), if the data are clustered in time, or if model states need to be repeatedly updated as new data come in. In such cases a sequential approach can be expedient. "

The wording is concise, but it is carefully chosen. Nowhere in the manuscript do we state or imply that sequential methods are a kind of "second choice" option for when variational methods become computationally impractical. That is not our opinion, and we have been careful to avoid such implications in the manuscript. Indeed, we do not consider "variational methods" to be the sole alternative to sequential methods (for example, MCMC sampling of the posterior probability is neither a sequential nor a variational method). We define "variational DA" in a broad sense in Sect. (2.2.3) ("Within the marine ecosystem modelling community, the term "variational DA" is often used more broadly to refer to all non-sequential methods that involve the minimisation of a cost function, whether or not this is based on a probability model"). Our theoretical ideal (probabilistic) method requires accurate evaluation of the posterior probability (Eq. 6) and/or the likelihood (Eq. 7). Minimization of a cost function can in general provide only a posterior mode estimate, and in practice only an approximation of this. Therefore, variational DA is not a theoretically ideal method. To be clear, neither sequential nor variational methods are ideal, nor do they cover all approaches. They both involve approximations that may be more or less appropriate depending on the particular problem. We do not imply that sequential and variational methods cannot both be applied to the same problem (indeed, this may be a useful approach to ensure robust scientific conclusions).

It is true that sequential methods can have a 4D component if a sequential smoother algorithm is applied. However, a sequential smoother is still functioning as a computational aid, breaking the large integration/sampling problem defined by Eqs. (6) and (7) into a set of smaller, more feasible integration/sampling problems by subsetting observations with respect to sampling time (e.g. see Evensen and van Leeuwen, 2000, section 2a). If we could accurately and efficiently evaluate Eqs. (6) and (7) by some other means, we would not need a sequential smoother algorithm. And if a smoother algorithm is not decomposing the problem by subsetting with respect to sampling time, then it would not be a "sequential" smoother. Our next sentence reads: "The basic idea is to break the large integration problem defined by Eq. (7) into a number of smaller problems by sequentially assimilating observations in subsets defined by sampling time."

The reviewer seems to give a negative connotation to the word "break" here, which was certainly not our intention. For us, the strategy of breaking a large, intractable problem into a set of smaller, tractable smaller problems is surely an intelligent one with positive connotations. The large integral in Eq. (7) is broken into smaller pieces in the sense that x in Eq. (7) has dimension $N_g \times N_s \times N_t$ where N_g is the number of spatial grid points, N_s, is the number of model state variables, and N_t is the number of time steps (between each of which a stochastic perturbation may be applied). The dimension of x_j^t in Eq. (8) is only $N_g \times N_s$. So the problem is far from solved, but it has certainly been broken into smaller pieces. We avoid saying that the

problem is "reformulated", because for us this word may imply some fundamental change in the statistical inference problem, either in the error model or the dynamical (process) model, or in the estimation method.

In our view it is essential that readers appreciate the fundamental statistical basis of data assimilation algorithms (which is not necessarily their historical development). We do this by linking the sequential forecast/analysis equations (Eqs. 8 and 9) and the application of Bayes' Theorem to the model parameters (Eq. 6) (cf. Evensen and van Leeuwen, 2000, sections 2a,b, also see Evensen, 2009, and Jazwinksi, 2007). We want readers to understand that by adopting a sequential method they do not necessarily need to change their statistical model or estimation method (e.g. posterior mode vs. posterior mean, Bayesian vs. maximum likelihood). In our view, the models (dynamical and statistical), and estimation methods should be selected to suit the scientific problem, then the data assimilation algorithm should be selected to suit the models and estimation methods. We should not in general choose a model/estimation method to suit an algorithm, just as we should not choose a problem to suit a model/estimation method (of course there are exceptions and compromise is necessary in the real world).

Regarding the final sentence: "If the sequential approximation or 'filter' is accurate, it should approximate the posterior distribution defined by Eqs. (6 and 7), when all data have been assimilated by the end of the assimilation period."
Here we have modified it to: "If the sequential algorithm is accurate, it should approximate the posterior parameter distribution defined by Eqs. (6 and 7) at times where all available data have been assimilated." This allows us to avoid specialising to the case of a filter, which as the reviewer noted is not appropriate for an introductory paragraph.

***Comment* 3**:

Page 7, lines 49-53: Here the (Extended) Kalman filter is discussed. Actually, these filters are not relevant for large-scale systems like in ocean biogeochemical modeling, because of model nonlinearities and because of the high-dimension of the models, which makes it impossible to store the full state error covariance matrix. This is know to most researchers for more than a decade and accordingly studies use variants of the ensemble Kalman filter, which can cope with high-dimension and partly with the nonlinearities. I recommend to remove the lines discussing the (extended) Kalman filter.

Author's response: We stress that this is not a review of data assimilation techniques applicable to large-scale systems. It is a review of parameter identification methods for marine ecosystem models of all scales, from laboratory through mesocosm to regional and global scale models, including box models and 1D vertical models of station time series. We think it is appropriate to keep these two sentences because: i) they help orient a reader in the literature and complete the overview by linking back to the historical origins of sequential methods, ii) given that we are not concerned only with large-scale models, we should not assume that the covariance matrix will exceed storage capacities, and iii) it is not clear that ensemble algorithms are the only reasonable approaches to limit storage requirements. There may be problems where a non-ensemble Kalman filter may be applicable to parameter estimation, perhaps after applying suitable transformations to deal with nonlinearity (e.g. using a SEEK approach to reducing the rank of the covariance matrix, see Nerger et al., 2005).

***Comment* 4**:

Page 7, lines 74-81: The text mentiones that there are 'powerful mathematical tools' for variational DA and that the adjoint methods are 'extremely efficient'. Here, the authors should be more specific as the statements are too superficial. What mathematical tools are meant; why are they 'powerful'? Further, what is 'extremely efficient' and compared to which methods is this the case? Actually, given the importance of the adjoint method, I would recommend that the authors state here in the main text, what the adjoint method actually is.

Author's response: All relevant information is given in the text already (Sect. 5.3.3 and Appendix C,). Section (2.2.3) informs about the advantages and drawbacks. However, we slightly rearranged and modified the second paragraph, so that it should become clearer: "In any case, there are some powerful mathematical tools developed for variational DA that can be applied to minimise cost functions. Adjoint methods allow the gradient of the cost function with respect to all fitted parameters to be computed in an extremely efficient manner, see Lawson et al. (1995), and Appendix (C). This is particularly useful when dealing with a large number of fitted parameters (high-dimensional Θ) of a computationally expensive model (e.g. Tjiputra et al., 2007). The application of the adjoint method helps reducing the number of model runs to provide access to joint posterior mode and maximum likelihood estimates."

***Comment* 5**:

Sections 2.3 and 2.4: These sections actually describe further aspects of the theoretical data assimilation background. I recommend to change them into sub-subsections of Section 2.2. Section 2.3 could also be merged with Section 2.2.3, because the cost function discussed here is only relevant for variational methods, which are discussed in Section 2.2.3. When Section 2.4 is changed in to sub-sub-section of 2.2 also its first sentence 'We close this section' would be reasonable since 2.2 is concerned with DA methods.

Author's response: The detailed structure was refined many times while being profoundly discussed among the lead authors of Sect. (2). On balance we realise that the present structure is entirely consistent:

> Sect. (2.1) reviews **Statistical model formulation**
>
> Sect. (2.2) reviews **Estimation Methods**
>
> Sect. (2.3) is a simple **worked example**
>
> Sect. (2.4) considers **terminology** (and closes Sect. 2 **Theoretical Background**)

Section (2.3) does not actually provide further methodology, but rather provides a worked example. The cost functions in Sect. (2.3) are not quite only relevant to variational methods; they are also relevant to other non-sequential methods such as MCMC sampling.

***Comment* 6**:

Last lines of Section 3.2 and Section 3.5: Section 3.5 seems to repeat some aspects on model formulations that account for acclimation dynamics which are already mentioned in Section 3.2. I recommend to focus 3.2 clearly on the aspect of limitation to avoid the redundancy.

Author's response: We thank the referee for spotting this. To avoid a repetition we now focus in Sect. (3.2) on the light and nutrient limitation aspect and modified Sect. (3.5) accordingly, with focus on the dependency between growth and cellular acclimation. The respective part in Sect. (3.2) now reads: "... .The third approach involves combinations of light- and nutrient limitation that resolve interrelations between cell quota, N-uptake and the photoacclimation state of the algae (e.g., Armstrong, 2006, see Sect. 3.5). Whether the first, second or third approach is considered can be expected to affect estimates of the associated parameter values."

And the modified Sect. (3.5): "More complex growth dependencies are described with models that consider intracellular acclimation dynamics (e.g., Geider et al., 1998; Pahlow, 2005; Armstrong, 2008; Wirtz and Pahlow, 2010). In these models, photoautotrophic growth rates become dependent on cell quota, e.g. usually normalised to carbon biomass (N:C), and the amount of synthesised Chl*a* per cell. With such approaches, the changes of the mass distribution of phytoplankton C and N, as well as the cellular Chl*a* content, have to be explicitly resolved in the model. One advantage is that these models are more sensitive to variations in light conditions and nutrient availability. The respective equations involve physiological parameters that are related but not identical to those of classical N- or P-based growth models, which impedes a direct comparison of older estimates of growth parameters with values currently used in models with acclimation processes resolved."

***Comment* 7**:

Page 12, line 96: The texts cites '(Simon and Bertino, 2012, Fig. 1)'. I can only guess that 'Fig. 1' does actually refer to Figure 1 of the manuscript and not Fig. 1 of Simon and Bertino (2012). This guess is based on the fact that I didn't find any other place where the manuscript refers to Fig. 1. Actually, in the current form, the figure is nothing more than a pure illustration because the only connection with the main text is that it refers to the figure. As for the other figures, please describe in the text what is the particular result shown in Fig. 1.

Author's response: To avoid ambiguity we have modified the sentence to: "It is yet unclear whether such extra flexibility is generally necessary, but it has been demonstrated that the choice of transformation can strongly affect estimates of plankton ecosystem fluxes (Evans, 2003) and that a good choice can improve parameter estimation in twin experiments (see Fig. 1 and Simon and Bertino, 2012)."

We acknowledge that there is not much discussion of Fig. (1), but it is not a pure illustration because it is clearly cited as an example of how a good choice of parameter transformation can improve parameter estimation in twin experiments. The particular result should be already clear from the above sentence. Note that other figures from variational studies (e.g. Figs. 3

and 4) receive a similarly terse treatment in the main text, as is necessary to limit length given the breadth of the manuscript's scope (treating studies at all scales).

***Comment* 8**:

Page 13, lines 14-16: It is stated 'Neglected correlation may result in parameter estimates that are less efficient... and more strongly correlated'. Please provide a reference for this statement.

Author's response: The sentence has been modified to: "Neglected correlation may result in parameter estimates that are less efficient (higher variance) and more strongly correlated (e.g. see example in Sect. 5.4)."

Note that such results may be well anticipated from classical results of linear regression analysis (where the generalised least squares estimator is more efficient than the ordinary least squares estimator when the true errors are correlated).

***Comment* 9**:

Page 13, lines 67-68: It is stated 'To our knowledge no application has yet incorporated prior correlations between parameters'. This statement does actually ignore that ensemble-based sequential DA schemes naturally include correlations between parameters, if the model dynamics yield them.

Author's response: Both sequential and non-sequential approaches can yield correlations between posterior parameter values (Bayesian paradigm) and between parameter estimates (frequentist paradigm). The statement clearly concerns the **prior** correlations between parameters. An ensemble of stochastic model simulations will certainly generate correlations between **state variables** (whether or not the parameters are estimated by sequential or non-sequential methods). But we are not aware of any study that has assumed **a priori** correlations between **parameters**.

***Comment* 10**:

page 16, lines 17-19: 'The determination of parameter uncertainties has many facets, getting to the core of discussions of Bayesian and frequentists approaches...'. This statement cannot be understood unless the reader already knows what it actually means. One needs to read the full section 5 to get the idea which facets are meant and what the Bayesian and frequentist approaches are. Please revise the text so that readers don't need to speculate what the authors actually mean to say.

Author's response: We understand the referee's concern and moved sentences from the first paragraph of Sect. (5.1) to this introduction part. The first paragraph of Sect. (5) now reads: "The determination of parameter uncertainties has many facets, getting to the core of discussions between Bayesian and frequentist approaches and interpretations (e.g., Efron, 1986; Cox, 2005; Lele and Dennis, 2009). Depending on the estimator, uncertainties in the combination of parameter values may either disclose a credible region of a random distribution of parameter values (Bayesian interpretation) or they mark a confidence region that should include the true value with a certain nominal probability of e.g. 95% (frequentist interpretation). The latter

means that different data sets would yield different confidence regions and e.g. 95% of those regions are expected to include the true "fixed" value."

***Comment* 11**:

Section 5 in general: Unfortunately, the authors missed to include uncertainty estimates from ensembles methods. Please also discuss it for completeness.

Author's response: This is a valuable comment. How the posterior distribution, in principle, can be resolved has already been mentioned in Sect. (2.2.2). However, we agree with the referee and think that it should be picked up again here. In this respect we also find the work of Weir et al. (2013) to provide good information with examples of a twin experiment. We have modified the paragraph: "In the case of classical BEs no tolerance limit Δ_J is explicitly prescribed. Instead, some efficient sampling of $(\Theta, J(\Theta))$, or directly of the posterior $p(\Theta \mid \boldsymbol{y})$, is applied. Sequential methods provide approximations of the posterior parameter distribution once all data have been assimilated. These approximations differ, depending on how Eqs. (6) and (7) are sampled and evaluated, as discussed in Sect. (2.2.2). A helpful overview with some comprehensible examples (of four different methods and three different ensemble sizes) is given by Weir et al. (2013). BE methods that do not rely on sequential approaches may also be applied and credible regions are then simply inferred from selective (acceptance/rejection) sampling schemes in a MCMC approach, e.g. Metropolis-Hastings algorithm (Metropolis et al. (1953; Hastings 1970). MCMC methods for the derivation of credible regions are also used for ML estimation problems (e.g., Smith and Yamanaka, 2007a). The main point is that here the data are assumed fixed.

A fundamentally different approach to the BE methods is to repeat parameter optimisations many times but with data subsamples or resample data sets. Large data sets are split up into... "

***Comment* 12a**:

Section 6.2: The section contains the statements: 'the appropriate degree of model complexity in any given situation is both one of the most important, and one of the least well defined' (p 20, lines 49-50) 'there exists a fundamental trade-off between simplicity and complexity' (p20, l54-55) 'the extra degrees of freedom can lead to the introduction of compensatory errors at the assimilation site' (p20, l67-68) 'an extra flexibility may lead to very different model solutions with only small variations in the assimilated data' (p20, l69-71) All these claims appear to be results from scientific studies rather than the authors' opinion. Accordingly supporting references are required to make the claims valid for a scientific paper.

Author's response: As a matter of style, we feel that excessive citation can in some cases detract from readability without providing any real additioinal help to the reader. Considering each of the instances in question:

1) "Of the many factors that affect the ability of a biogeochemical model to reproduce and predict observations, the appropriate degree of model complexity in any given situation is both one of the most important, and one of the least well defined."

This is a general and uncontroversial statement which serves to introduce Sect. (6.2). It does not require a supporting reference, in our view. It is followed by:

2) "This is because there exists a fundamental trade-off between simplicity and complexity."
Again, a general and uncontroversial statement, which is immediately explained in the two following sentences: "Simple models have the advantage of being easier to understand, and with fewer parameters they should also be better constrained (both before and after optimisation). Nonetheless, simplification requires a degree of abstraction, and it can sometimes be difficult to draw parallels with the complexities of the observed system."

Regarding the next two

3) "If insufficient observations are available, the extra degrees of freedom can lead to the introduction of compensatory errors at the assimilation site, which could then increase uncertainty at other locations."

4) "Similarly, an extra flexibility may lead to very different model solutions with only small variations in the assimilated data, also leading to increased uncertainty in model predictions."
Here we agree that references will be helpful; we have modified to: "If insufficient observations are available, the extra degrees of freedom can lead to the introduction of compensatory errors at the assimilation site, which could then increase uncertainty at other locations, as illustrated by Xiao and Friedrichs (2014b). Similarly, for small changes in the assimilated data an extra flexibility may lead to very different model solutions, also leading to increased uncertainty in model predictions (e.g. Xiao and Friedrichs, 2014a)."

***Comment* 12b**:
page 20, line 75-76: What makes the review by Johnson and Omland (2004) 'useful'?

Author's response: It is our opinion that the review is "useful". The word was included because it helps to communicate our thinking to the reader.

***Comment* 12c**:
Why is cross-validation 'most practical' and perhaps 'most general'? The particular expressions are opinions of the authors. It would be preferable for a scientific paper if the authors focus on facts.

Author's response: The full sentence reads: "One of the most practical (if not the most general) techniques is cross-validation, as described in the previous section." This is not such a strong statement (because we say "One of..."). It also directs the reader to the previous section where we give a description: "This is the principle of cross-validation, in which an optimised

model is tested in terms of its ability to reproduce data that were not included in the calibration phase. This is achieved by excluding a subset of the original calibration dataset, for later use in model evaluation." In our view, the fact that the method can be described in plain language in two sentences itself supports the statement that this is a practical and general method. In fact the practicality and generality of cross-validation is not controversial. Hastie et al. (2009) begin their section on 7.10 Cross-Validation with the sentence: "Probably the simplest and most widely used method for estimating prediction error is cross-validation." Given that this important text is not cited elsewhere and provides an excellent discussion of cross-validation (in our opinion) we have decided to cite it here. Our sentence is modified to: "One of the most practical and general techniques is cross-validation, as described in the previous section (see also Hastie et al., 2009, section (7.10) for an excellent discussion in a general statistical context). By looking at ..."

***Comment* 13**:

page 21, line 1: I recommend to start a new paragraph at 'A perhaps more intuitive....'

Author's response: This is a good suggestion. We changed it accordingly.

***Comment* 14**:

page 21, lines 68-69: It is stated 'Among models with a similar score, the simplest should be favoured'. Here, I again recommend to rephrase the statement to be scientific, which excludes 'favours'. Scientifically, the optimal model choice seems to be that one with the least parameters and minimum score or score within a certain threshold from the minimum.

Author's response: Actually this sentence is not correct, and begs the question of a threshold for defining "similar". The simplest standard approach is to choose the model with the lowest AIC (or AICc or BIC). However, this is not a robust approach when we have multiple models that achieve "similar" AIC scores; a better approach in general is to weight the models according to their AIC and perform multimodel inference (Burnham and Anderson, 2002). We have replaced this sentence and the preceding one with the following: "The AIC and BIC can be used to select a single model with the lowest score, or preferably to provide individual model weightings for multimodel inference (Burnham and Anderson, 2002), although it appears that this latter has so far seen little application to planktonic ecosystem models."

***Comment* 15**:

Section 7.1.4: Given the fact that Section 7.1.3 already discussed studies considering time- or space-varying parameters, the first sentence of the section reads quite odd. Please rephrase it.

Author's response: Section (7.1.4) (now Sect. 7.5, following a correction to the section numbering) is intended to be a concluding discussion for Sect. (7) as a whole, with the first sentence acting as a summary. This was not clear, so the sentence which reads "A variety of approaches have been explored for DA with parameters varying in space or time or both."

is now replaced with

"As shown in this section, a variety of approaches have been explored for DA with parameters varying in space or time or both. We conclude the section by considering what might be learnt from these types of studies."

***Comment* 16**:

Section 8: This sections appears to be overly detailed compared to the treatment of others aspects in the manuscript. At the same time it is too short to really understand details. E.g. from reading the text, I could not really understand the meaning of he 'alignment operator' (pages 25/26). How can the emulator equal the model (page 26, line 2)? this seems to ignore the presence of the alignment operator matrices $A_l i$. However, given the small number of references in Section 8, its length is not consistent with the current relevance of the methods. Thus, I recommend to shorten the section to a concise overview of the dynamic and statistical emulators. One clear possibility for shortening is also the example in Section 8.2 (page 26, right column), which is too detailed and incomplete at once. While for the model itself it is referred to a publication, the manuscript lists explicit parameter values, which is of no use without knowing the model equations. Further, the example is only concerned with a 0D case, which appears to be trivial in particular as the authors intent to discuss emulators in the context of high-dimensional models. The application in 3D cases of higher dimension, where the method could be most useful, appears to be extremely difficult.

Author's response: Emulator approaches are a relatively new development and thus the number of available publications is still limited. It seems, however, that emulator approaches will gain more and more importance in the future - particularly as model complexity increases due to increased computer power, which hinders systematic data assimilation with full models. We thus believe that this upcoming development needs a detailed elaboration. We again went critically through the section and decided to add few sentences to Sect. (8.1) for further clarification of the alignment operator. Along with this, we reformulated page 26 line 2 and we agree that the original formulation was misleading. We also found places where we could shorten text in Sect. (8.2).

Although simple, the example of the statistical emulator is not at all trivial. The example is very illustrative, because it gives an idea of the similarity between the contours (here of RMS error) of the model and the emulators, for a given training set. Whether a 0D- or a (computationally expensive) complex 3D model is applied is irrelevant in this context. We agree that the level of detail provided in the example description may not be appropriate. We therefore decided to shorten the description of the example considerably, and we have moved relevant details to the caption of Fig. (7).

In Sect. (8.3) we revised text, mainly to come up with a slightly shortened version.

***Comment* 17**:

Section 9: The authors stated in their response that they shortened this section. However, this doesn't seem true as in the orig-

inal version the Section spread over about 4.5 pages and now (in the document version 'author response version 1', it's again 4.5 pages. Particularly long is subsection 9.3. While its title suggests that the section discusses 'Parameters relevant for global ocean BGC modelling', the section does almost exclusively discuss the parameter 'b' of the power law for particle flux. Here also an example from Kriest et al. (2016) is included, which is described in quite some detail. This again seems to be too detailed, even more as the conclusion appears to be that the value of 'b' is 'well identifiable' (page 39, line 76) (similarly in line 62 for a different case). To this end, the main result appears to be that the value of 'b' can be determined but is specific for each model. I wonder, why so much space is used to describe this result.

Author's response: Following the attempts of our first response we removed original paragraphs, then restructured and rewrote paragraphs with less detail. However, we then learned that we should consider examples with sequential approaches as well (Ridgwell et al., 2007 and Simon et al., 2015). Eventually, we realised that we had forgotten to refer to Tjiputra et al. (2007), a study where an adjoint model had been applied. These new paragraphs (with new information included) turned out to compensate for the reduction of text elsewhere in Sect. (9).

We screened the individual sections and removed sentences or phrases where possible. But in general, we think that the level of detail is appropriate for Sects. (9.1), (9.2), and (9.4). Sect. (9.3) is of central importance, since therein we address those parameters that are relevant for determining dissolved inorganic tracer distributions on long time scales and large spatial scales (sinking velocity of detritus and remineralisation rate) and we discuss how they are linked to parameterisations of particle export flux (e.g. when described with power law of depth). Note that the focus is on dissolved inorganic tracer distributions for model calibration, which is justified by the data availability discussed in Sect. (9.2). Further, this section, and its emphasis on "b" is motivated by the importance of this parameter for atmospheric pCO_2 (Kwon et al., 2009; see first paragraph of this subsection). This parameter has therefore received much attention in the global biogeochemical modelling community. Because it has been popular in global models, we think it is justified to have a closer look at its implicit assumptions. The fourth paragraph of Sect. (9.3) describes an example of how climatological data impose different constraints on parameters that act on the surface and those of particle export, and thus provides a link to Sect. (5) of the paper; this time from a global perspective.

To summarise, in the first paragraph of Sect. (9.3) it is clarified that not all ecosystem model parameters that are important in a local or mesocosm context, or on short time scales, are also relevant for tracer distribution in the global ocean. We address computational constraints, and implications for parameter identification in Sect. (9.1), data constraints (Sect. 9.2), and state-of-the-art of identifying the biogeochemical parameters in global models, run on long time scales (Sect. 9.3). Section (9.4) finally closes this review by summarising the relevance of model sensitivities for future projections. As such, we think the Sect. (9) is comprehensive and concise.

***Comment* 18**:

Page 31, line 17: Please add a reference to the assessment report of the IPCC.

Author's response: We added a reference.

***Comment* 19**:

Page 31, lines 94-95: It is stated: 'In BGC models the conservation of mass can be essential, in particular for large-scale or global ocean simulation'. Actually, the conservation requirement is not resulting from a model being large-scale or global, but it results from the scientific question to be considered. In the current form, the text implies that all data assimilation applications of sequential methods with large-scale or global models are wrong because they don't conserve the mass. This is certainly not true and the authors contradict their own statements in Section 2.2. Please reformulate the statement.

Author's response: The sentence in question concludes a sequence of three sentences: "Whether the DA approach conserves mass and/or energy is relevant in this respect, depending on the scientific problem addressed. Some ecosystem model applications may not critically depend on mass conservation, e.g. when simulating plankton growth to act as food source in regional simulations of fish stock size and recruitment. In BGC models the conservation of mass can be essential, in particular for large-scale or global ocean simulations."

No mention is made here of sequential methods, and nowhere in the present manuscript do we state that sequential methods necessarily violate mass conservation. We only state in Sect. (1.4) that stochastic models (which do not necessarily imply sequential methods) may violate mass conservation, and then in Sect. (4.5) we discuss how stochastic noise may be injected without violating mass conservation. Further, the second sentence in the above explicitly acknowledges that there may be problems where mass conservation may not be essential, and the final sentence only states that mass conservation can be essential. However we agree with the reviewer regarding the model vs. scientific question, and have therefore modified the final sentence to read: "In BGC applications the conservation of mass can be essential, in particular for large-scale or global ocean applications."

***Comment* 20**:

Page 32, right column, lines 2-3: I already recommended in my first review to mentioned here what the dynamics and statistical emulators are. Unfortunately, the authors just replied that they don't see a need for this, because they defined these emulators in Sections 8.1 and 8.2. To this end, I like to remind the authors about the fact that this is the summary section. Usually, one doesn't expect that readers will read the whole paper but many readers will focus on the introduction and summary (which are already quite long in this manuscript). Thus, it would just help readers if the authors would add one or two sentences shortly mentioning that a dynamical emulator is a computationally cheap approximation of the model operator, while a statistical emulator simulates the output from inputs in a statistical way based on a prior training with independent input/output sets. This should be possible in a very short way so that the overall length of the manuscript is not significantly changed.

Author's response: We understand and follow the referee's suggestion. We like the referee's phrase but are not sure whether

we could/should adopt it directly. However, we added three sentences: "The dynamical emulator is a simpler representation of a full model operator that is computationally expensive, thereby approximating the underlying model dynamics. A statistical emulator interpolates model output from a set of training runs with different values assigned to the parameter vector. Based on the derived statistics it can be applied to approximate unknown model output for other input parameters."

***Comment* 21**:

Fig. 6: It's written 'Geographic extent of the two sub-domains'. I can only guess that the colors in the plot distinguish the two sub-domains. Unfortunately, this is never described.

Author's response: The beginning of the caption is now clarified to read: "Geographic extent of the two sub-domains giving the optimal calibration in the split-domain calibration study of Hemmings et al. (2004), shown here in yellow and green. Also shown are the distributions of the sites used from the calibration set to obtain the parameter vectors for each sub-domain and the sites used for cross-validation. Biogeochemical provinces defined by Longhurst ...".

***Comment* 22**: Typos

Author's response: We thank the referee for spotting those typos. We corrected all six of them.

References

Evensen, G.: The ensemble Kalman filter for combined state and parameter estimation, IEEE Control Systems, 29, 83–104, 2009.

Evensen, G. and van Leeuwen, P.: An Ensemble Kalman Smoother for Nonlinear Dynamics, Monthly Weather Review 128, 1852-1867, 2000.

Hastie, T., R. Tibshirani, and J. Friedman (2009), The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed., Springer, New York.

Jazwinski, A. H.: Stochastic processes and filtering theory, Courier Corporation, 2007.

Nerger, L., Hiller, W., and Schroter, J.: A comparison of error subspace Kalman filters, Tellus 57A, 715-735, 2005.

Reviews and syntheses: Parameter identification in marine planktonic ecosystem modelling

Schartau Markus[1], Wallhead Philip[2], Hemmings John[3,4], Löptien Ulrike[1], Kriest Iris[1], Krishna Shubham[1], Ward Ben A.[5], Slawig Thomas[6], and Oschlies Andreas[1]

[1]GEOMAR Helmholtz Centre for Ocean Research Kiel, Germany
[2]NIVA, Norwegian Institute for Water Research, Bergen, Norway
[3]Wessex Environmental Associates, Salisbury, United Kingdom
[4]now at Met Office, Exeter, United Kingdom
[5]University of Bristol, School of Geographical Sciences, Bristol, United Kingdom
[6]Christian-Albrechts-Universität zu Kiel, Department of Computer Science, Kiel, Germany

Correspondence to: Markus Schartau (mschartau@geomar.de) and Phil Wallhead (philip.wallhead@niva.no)

Abstract. To describe the underlying processes involved in oceanic plankton dynamics is crucial for the determination of energy and mass flux through an ecosystem and for the estimation of biogeochemical element cycling. Many planktonic ecosystem models were developed to resolve major processes so that flux estimates can be derived from numerical simulations. These results depend on the type and number of parameterisations incorporated as model equations. Furthermore, the values assigned to respective parameters specify a model's solution. Representative model results are those that can explain data, therefore data assimilation methods are utilised to yield optimal estimates of parameter values while fitting model results to match data. Central difficulties are 1) planktonic ecosystem models are imperfect and 2) data are often too sparse to constrain all model parameters. In this review we explore how problems in parameter identification are approached in marine planktonic ecosystem modelling.

We provide background information about model uncertainties and estimation methods, and how these are considered for assessing misfits between observations and model results. We explain differences in evaluating uncertainties in parameter estimation, thereby also discussing issues of parameter identifiability. Aspects of model complexity are addressed and we describe how results from cross-validation studies provide much insight in this respect. Moreover, approaches are discussed that consider time and space dependent parameter values. We further discuss the use of dynamical/statistical emulator approaches, and we elucidate issues of parameter identification in global biogeochemical models. Our review discloses many facets of parameter identification, as we found many commonalities between the objectives of different approaches, but scientific insight differed between studies. To learn more from results of planktonic ecosystem models we recommend finding a good balance in the level of sophistication between mechanistic modelling and statistical data assimilation treatment for parameter estimation.

1 Introduction

The growth, decay, and interaction of planktonic organisms drive the transformation and cycling of chemical elements in the ocean. Understanding the interconnected and complex nature of these processes is critical to understanding the ecological and

biogeochemical function of the system as a whole. The development of biogeochemical models requires accurate mathematical descriptions of key physiological and ecological processes, and their sensitivity to changes in the chemical and physical environment. Such mathematical descriptions form the basis of integrated dynamical models, typically composed of a set of differential equations that allow credible computations of the flux and transformation of energy (light) and mass (nutrients) within the ecosystem (U.S. Joint Global Ocean Flux Study Planning Report Number 14, *Modeling and Data Assimilation*, 1992).

Generalised mechanistic descriptions of how energy is absorbed and how mass becomes distributed in an ecosystem already exist, such as dynamic energy budget models (Kooijman, 1986) or the metabolic theory of ecology (Brown et al., 2004). But these theories still have limitations, and include incompatible assumptions (van der Meer, 2006). So far no fundamental ecophysiological principle has been further exacted beyond the conservation of mass. A consistent theme running through most ecosystem models is the determination of mass flux of certain biologically important elements, such as nitrogen, phosphorus, iron and carbon (N, P, Fe and C). Nonetheless, the precise details of how mass is transformed and allocated within an ecosystem is far from being established. For this reason, we find a large variety of plankton ecosystem models that differ in their number of state variables as well as in their parameterisation of individual physiological and ecological processes.

1.1 Mass flux induced by plankton dynamics

Dynamical marine, as well as limnic, ecosystem models usually start from a description of the build-up of biomass by photoautotrophic organisms (phytoplankton) as these take up dissolved nutrients from the water column and exploit light energy by photosynthesis. Phytoplankton biomass, as a product of primary production, is subsequently removed by natural mortality (cell lysis due to starvation, senescence, and viral attack), predation by zooplankton, and vertical export away from surface ocean layers via sinking of single or aggregated cells and of fecal pellets. Parameterisations of these three loss processes can be interlinked e.g. grazing of phytoplankton aggregates by large copepods. Depending on the trophic levels considered in a model, the predation among different zooplankton types (e.g. between herbivores, carnivores or omnivores) can be explicitly parameterised. Mortality and aggregation of phytoplankton cells and the excretion of organic matter (fecal pellets) by zooplankton act as primary sources of dead particulate organic matter (detritus) that can be exported to depth via sinking. Exudation by phytoplankton and bacteria can be a major source of labile dissolved organic matter that represents diverse substrates for remineralisation. The transformation of particulate and dissolved organic matter back to inorganic nutrients is parameterised as hydrolysis and remineralisation processes. Often hydrolysis and remineralisation are assumed to be proportional to the biomass of heterotrophic bacteria, which is considered in many models. Heterotrophic bacteria remain unresolved in some models where microbial remineralisation is parameterised only as a function of concentration and quality of organic substrates.

At some level most models include a parameterisation to account for the net effect of higher trophic levels that are not explicitly resolved. This is usually formulated as a closure flux back to nutrient pools and whose rates simply depend on the biomass of the highest trophic level resolved. These closure assumptions ensure mass conservation while neglecting the actual mass loss to higher trophic levels like fish, which would be subject to fish movements and changes in biomass on multi-annual rather than seasonal time scales. Every marine planktonic ecosystem model can thus be described as a simplification

of the dynamics inherent to a system of nutrients, phytoplankton, zooplankton, detritus, dissolved organic matter, and possibly bacteria.

In many cases marine ecosystem models are embedded in an existing physical ocean model setup that simulates environmental conditions, advection and mixing of the biological and chemical state variables. Feedbacks from the ecosystem model states on physical variables can be relevant (e.g., Murtugudde et al., 2002; Oschlies, 2004; Löptien et al., 2009; Löptien and Meier, 2011) but are rarely considered in current marine biogeochemical studies.

1.2 Parameters of plankton ecosystem models

Amongst the most influential model approaches to study the nitrogen flux through such a marine plankton ecosystem at a local site was proposed by Fasham et al. (1990). Their model involves 27 parameters and they stressed the invidious situation of finding a reliable ecosystem model solution by choosing parameter values that are uncertain or unknown. Laboratory measurements, as well as ship-based experiments with field samples, can provide information about the range of typical values for some parameters, for example the maximum growth rate of photo-autotrophs or the maximum ingestion rate of herbivorous plankton. Other model parameters are extremely difficult to measure, like exudation rates of dissolved organic carbon by phytoplankton or by bacteria. Another difficulty is that parameter values from laboratory experiments are often specific with respect to plankton species, temperature, and light conditions. Their values may not be directly applicable for ocean simulations where parameter values need to be representative for a mixture of different plankton species in a continuously varying physical environment. For example, for a natural composition of diverse phytoplankton cells that all differ in their genotypic- and phenotypic characteristics, we may expect values of some model parameters to follow a distribution rather than having a single fixed value.

In practice, there are always some fixed model parameters that need to be assigned values, whether they describe the behaviour of fixed plankton functional types or the distributions of traits in a stochastic community. In the end, it is the choice of these parameter values that determines a specific model solution of any ecological- or biogeochemical model setup.

1.3 The vital role of observational data

Model solutions of interest are typically those that can simulate and explain complex data. Model calibration, which can be considered a form of data assimilation (DA), is the process by which model parameter values are inferred from the observational data. Optimal parameter values are regarded as those that generate model results that match observations (data-model misfit) but are also in accordance to the range of values known e.g. from experiments or from preceding DA studies. To determine optimal parameter estimates we have to account for uncertainties in data and in model dynamics as well, which is specified by an error model. Parameter estimates are thus conditioned by a) the dynamical model equations, b) the data, c) our prior knowledge about the range of possible parameter values, and d) the underlying error model (Evans, 2003).

Situations can occur where model results that are compared with data are insensitive to variations of some parameters. Values of those parameters remain unconstrained by the available data, which is a problem of parameter identifiability. The availability (type and number) of data thus places limitations on the number of model parameters whose values become identifiable,

and values of some parameters may never be fully constrained. This in turn sets restrictions on the complexity of plankton interactions that can be unambiguously confined during ecosystem model calibration (Matear, 1995). Choosing appropriate model complexity is ambiguous and is still subject to discussion (e.g., Franks, 2002; Denman, 2003; Fulton et al., 2003; Anderson, 2005; Le Queré, 2006; Friedrichs et al., 2007; Franks, 2009; Kriest et al., 2012; Ward et al., 2013), a situation which sustains large differences in the level of complexity of current plankton models.

1.4 Inferences from data assimilation

Much of the literature on DA in oceanography is focussed on state estimation (e.g., Allen et al., 2003; Natvik and Evensen, 2003; Dowd, 2007; Nerger and Gregg, 2008; van Leeuwen, 2010). In these studies, the primary objective is to improve hindcasts, nowcasts, or forecasts of time-dependent variables such as chlorophyll *a* (Chl*a*). However, many of the DA methods originally developed for state estimation have more recently been adapted to estimate static parameters, especially for stochastic models where random noise is injected into the model dynamics. Stochastic noise offers a plausible way to represent model error, but it should be noted that it can lead to violations of mass conservation unless it is injected in certain ways (e.g. by perturbing growth rate parameters). Deterministic plankton ecosystem models guarantee mass conservation and have a longer tradition in parameter estimation for marine ecosystem models, although they imply a less explicit treatment of model error. To identify and gradually eliminate model deficiencies it can be helpful to analyse model state and flux estimates while mass conservation is imposed as a strong constraint. The optimisation of only parameter values assures that simulation results remain dynamically and ecologically consistent, which is comparable with those DA approaches in physical oceanography that produce dynamically and kinematically consistent solutions of ocean circulation (e.g., Wunsch and Heimbach, 2007; Wunsch et al., 2009).

Thorough reviews of common DA methods applied in marine biogeochemical modelling are given by Robinson and Lermusiaux (2002) and by Matear and Jones (2011). Dowd et al. (2014) provide a helpful and up-to-date overview of mainly sequential DA approaches where state estimation is combined with parameter estimation. Gregg et al. (2009) and Stow et al. (2009) discuss how the success of DA results of marine ecosystem models have been evaluated in the past and how model performance can be generally assessed. Fundamentals on DA that include aspects relevant to marine ecosystem and biogeochemical modelling are explained in Wikle and Berliner (2007) and in Rayner et al. (2016).

In our review we primarily focus on topics related to parameter identification, thereby including basic aspects of DA. Parameter identification in marine planktonic ecosystem modelling is a wide field and we do not attempt to discuss differences between various DA tools or techniques. We rather put emphasis on models, including parameterisations of ecosystem processes, statistical (error) models, model uncertainties, and structural complexity. We adopt and explain mathematical notation that is often used for DA studies in operational meteorology and oceanography. On the one hand we provide background information that should facilitate intelligibility when studying DA literature. On the other hand we like to elucidate typical objectives and common problems when simulating a marine planktonic system. In this manner we hope to support a mutual understanding between ecologically/biogeochemically and mathematically/statistically motivated studies.

The paper starts with some theoretical background information (Sect. 2), introducing mathematical notation and depicting prevalent assumptions that are typically made for parameter identification analyses and model calibration (Sect. 2.1). We then branch off from DA theory and discuss the parameters typically dealt with in plankton ecosystem models. In Sect. (3) we disentangle major differences between approaches to parameterising photoautotrophic growth and briefly discuss simple but common parameterisations of plankton loss rates. In this context we also address the utilisation of data from laboratory and mesocosm experiments. Error models are described in order to elucidate error assumptions made in previous ecosystem modeling studies (Sect. 4). This is followed by a description of different approaches to specify uncertainties in parameter values (Sect. 5). An example of parameter estimation with simulations of a mesocosm experiment connects aspects of Sect. (3) with the theoretical considerations of Sect. (5). Thereafter, model complexity is jointly addressed together with cross-validation in Sect. (6), followed by a review of space-time variations in marine ecosystem model parameters (Sect. 7). Emulator, or surrogate-based, approaches are briefly explained and exemplified (Sect. 8) before we discuss parameter estimation of large-scale and global biogeochemical ocean circulation models (Sect. 9). Finally, we summarise the insights that we gained on parameter identification in Sect. (10), and we will briefly address prospects of some marine ecosystem model approaches that could improve parameter identification.

2 Theoretical background

The term parameter identification is used broadly to describe parameter estimation problems, including the specification of uncertainties in parameter estimates and model parameterisations. It involves the following procedures:

a) Parameter sensitivity analyses: the evaluation of how model results change with variations of parameter values.

b) Parameter estimation: the calibration of model results by adjusting parameter values in light of the data.

c) Parameter identifiability analyses: the specification of parameter uncertainties in order to reveal structural model deficiencies and shortages in data availability/information.

All three aspects are interrelated and should not be viewed as mutually exclusive procedures. For example, before starting with parameter estimation it is helpful to include information from a preceding sensitivity analysis, e.g. selecting only parameters to which model results are sensitive to. Likewise, an identifiability analysis complements the sensitivity analysis by providing information about error margins and possible ambiguities of optimal parameter estimates.

2.1 Statistical model formulation

2.1.1 Model states, parameters, and dynamical model errors

The prognostic dynamical equations of a marine ecosystem model can be expressed as a set of difference equations:

$$\boldsymbol{x}_{i+1} = M\left[\boldsymbol{x}_i, \theta_e, \boldsymbol{f}_i, \boldsymbol{\eta}_i(\theta_\eta)\right] \tag{1}$$

with index i representing a particular time step (i.e. t_i). The model state vector $\boldsymbol{x}_i$ has dimension $N_x = N_g \times N_s$ where N_g is the number of spatial grid points and N_s is the number of model state variables (e.g. phytoplankton biomass). The dynamical model operator M is typically at least a nonlinear function of the earlier state $\boldsymbol{x}_i$, a set of ecosystem parameters θ_e describing rate constants and coefficients in the dynamical model, and a set of time and space dependent forcings and boundary conditions $\boldsymbol{f}_i$. If the ecosystem model is coupled "online" with a physical ocean model, $\boldsymbol{f}_i$ includes both physical model forcings (e.g. wind stress) and ecosystem model forcings (e.g. surface short-wave irradiance). If the physics is coupled "offline", $\boldsymbol{f}_i$ includes ecosystem model forcings and physical model outputs (e.g. seawater temperature).

For stochastic dynamical models, M also depends on random noise variables or dynamical model errors $\boldsymbol{\eta}_i$ while for deterministic models we have $\boldsymbol{\eta}_i = 0$. These errors are described by distributional parameters θ_η, e.g. location and scale parameters of a probability density function. Dynamical model errors usually enter the dynamics additively, multiplicatively, or as time/space-dependent corrections to $\boldsymbol{f}$ or θ_e. They may represent the individual or combined effects of errors in forcings, boundary conditions, random variability in model parameters, and structural errors in both the physical transport model (e.g. due to limited spatial resolution) and the biological source-minus-sink terms (e.g. due to aggregation of species into model groups). In the geophysical DA community, error models that explicitly account for dynamical model errors (noise) are often termed *weak constraint* models, while those that assume a deterministic model are termed *strong constraint* (Sasaki, 1970; Bennett, 2002, page 25).

2.1.2 True states and kinematic model errors

To relate the dynamical model output of Eq. (1) to observations, it is helpful to first consider how it may relate to a conceptual and hypothetical true state $\boldsymbol{x}^t$, which is then imperfectly observed. In this respect we must also consider the averaging scales. In marine ecosystem modelling there is almost always a large discrepancy between the spatio-temporal averaging scales of the model, that define the meaning of the "concentrations" in $\boldsymbol{x}$, and the averaging scales of the observations from in-situ sampling or remote sensing. For example, the spatial averaging scale of a model may be defined by a model grid cell of size 10 km in the horizontal and 10 m in the vertical, while the averaging scale of the observations might be the 10 cm scale, e.g. of a Niskin bottle sample. Even with a perfect model, data from finescale observations may diverge from model output due to unresolved sub-grid scale variability induced by fluid structures such as eddies and fronts, forming patches of high next to low concentrations e.g. of nutrients or organic matter.

A general relationship between the true state and model state can be expressed as:

$$\boldsymbol{x}^t = T\left[\boldsymbol{x}, \boldsymbol{\zeta}(\theta_\zeta)\right] \tag{2}$$

where T is a truth operator, and $\boldsymbol{\zeta}$ is a set of random variables described by distributional parameters θ_ζ. We will refer to the $\boldsymbol{\zeta}$ as *kinematic* model errors because they are associated with the model state, while the *dynamical* model errors $\boldsymbol{\eta}$ in Eq. (1) act to perturb the model dynamics. The true values of the kinematic model errors therefore define the potential discrepancy between the target true state and a hypothetical ideal model output (i.e. with the "true" values of the parameters and, if applicable, also with the "true" values of the dynamical model errors).

How we interpret and specify Eq. (2) depends on the spatio-temporal averaging scales chosen to define the true state $\boldsymbol{x}^t$, which in turn depends on the objectives of the modelling study. One approach is to define these averaging scales as equal to or larger than the shortest space and time scales that are fully resolved by the model. Kinematic model errors $\boldsymbol{\zeta}$ may then represent the integrated effects of the various dynamical sources of model error, if these are not already accounted for by dynamical model errors $\boldsymbol{\eta}$ in Eq. (1). Alternatively, the true state can be defined over scales smaller than those resolved by the model, possibly at the scales of the observations. This may lead to a simpler model for observational error (see below), but now the $\boldsymbol{\zeta}$ must account for the unresolved scales, in addition to any error effects in the model dynamics otherwise not accounted for. With stochastic dynamical models ($\boldsymbol{\eta} \neq 0$), the true state is usually defined on the scales of the model and assumed to coincide with the model output for some (θ_e, $\boldsymbol{\eta}$), such that no kinematic error model is needed.

2.1.3 Data and observational errors

The observation vector $\mathbf{y}$ can be related to the true state via:

$$\boldsymbol{y} = O\left[\boldsymbol{x}^t, \boldsymbol{\epsilon}(\theta_\epsilon)\right] \tag{3}$$

where O is the generalized observation operator and $\boldsymbol{\epsilon}$ is a set of random *observational* errors described by distributional parameters θ_ϵ and accounting for uncertainties associated with the usage and interpretation of the data. These include at least the random measurement error due to, for example, instrument noise. In addition they may include a contribution from *representativeness* error due to finescale variability, if $\boldsymbol{x}^t$ is defined as an average over larger scales than those of the observations (see above). Alternatively, if the observations are preprocessed into estimates on the larger scales of $\boldsymbol{x}^t$, there may be an *undersampling* error component due to inexhaustive coverage of the raw samples. The observation operator O may also contribute to $\boldsymbol{\epsilon}$, for example if the model output needs to be interpolated from the model grid to the data coordinates, or if O includes conversion factors such as chlorophyll *a*-to-nitrogen (Chl*a*:N) ratios.

The simplest possible example of an observational error model assumes additive Gaussian errors. Equation (3) then becomes:

$$\begin{aligned} \boldsymbol{y} &= H\left(\boldsymbol{x}^t\right) + \boldsymbol{\epsilon} \\ &\longrightarrow \boldsymbol{\epsilon} = \boldsymbol{y} - H\left(\boldsymbol{x}^t\right) \end{aligned} \tag{4}$$

where H accounts for interpolation and units conversion and $\boldsymbol{\epsilon} \sim G(0, \mathbf{R})$ is Gaussian distributed with mean zero and covariance matrix $\mathbf{R}$. This may be a reasonable error model for most physical variables and chemical concentrations with ranges well above zero (e.g. dissolved inorganic carbon or total alkalinity in the open ocean). However, many nutrients and plankton biomass variables may vary close to their lower bounds of zero, and display positive skew in their observational errors. For such variables, a lognormal observational error model may be more appropriate:

$$\begin{aligned} \boldsymbol{y} &= H\left(\boldsymbol{x}^t\right) \circ \exp\left(\tilde{\boldsymbol{\epsilon}} - \frac{\tilde{\boldsymbol{\sigma}}^2}{2}\right) \\ &\longrightarrow \tilde{\boldsymbol{\epsilon}} = \log(\boldsymbol{y}) - \log\left(H\left(\boldsymbol{x}^t\right)\right) + \frac{\tilde{\boldsymbol{\sigma}}^2}{2} \end{aligned} \tag{5}$$

where $\circ$ denotes element-wise multiplication and $\tilde{\boldsymbol{\sigma}}^2$ denotes the variance in logarithmic space. The bias correction term ($\tilde{\boldsymbol{\sigma}}^2/2$) ensures unbiased errors, but is frequently neglected in practice. The various options and challenges of defining an appropriate error model are discussed in detail in Sect. (4).

2.2 Estimation methods

2.2.1 Basic probabilistic approaches

We now consider how to estimate uncertain parameters Θ given the data $\boldsymbol{y}$, where Θ includes all biological parameters θ_e and possibly distributional parameters (θ_η,θ_ζ,θ_ϵ). There are basically two probabilistic approaches for doing this: Bayesian estimation and maximum likelihood estimation. In the Bayesian approach, we treat the parameters as random variables, and choose parameter values on the basis of their 'posterior probability' i.e. the conditional probability density of the parameter values given the data $p(\Theta \mid \boldsymbol{y})$. The posterior probability is computed using Bayes' theorem:

$$p(\Theta \mid \boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid \Theta) \cdot p(\Theta)}{p(\boldsymbol{y})} \propto p(\boldsymbol{y} \mid \Theta) \cdot p(\Theta) \tag{6}$$

where $p(\boldsymbol{y} \mid \Theta)$ is the likelihood and $p(\Theta)$ is the unconditional or 'prior' distribution of the parameter values. The proportionality follows in Eq. (6) because the probability of the data $p(\boldsymbol{y})$, otherwise known as the "evidence" for the model, is independent of the parameter values.

In general the likelihood can be expressed as an integral over probabilities conditioned on particular values of the model state and true state:

$$p(\boldsymbol{y} \mid \Theta) = \int\int p(\boldsymbol{y} \mid \boldsymbol{x}^t, \Theta) \cdot p(\boldsymbol{x}^t \mid \boldsymbol{x}, \Theta) \cdot p(\boldsymbol{x} \mid \Theta) \, \mathrm{d}\boldsymbol{x}^t \, \mathrm{d}\boldsymbol{x} \tag{7}$$

where the conditional probabilities $p(\boldsymbol{y} \mid \boldsymbol{x}^t, \Theta)$, $p(\boldsymbol{x}^t \mid \boldsymbol{x}, \Theta)$, and $p(\boldsymbol{x} \mid \Theta)$ are specified by the chosen models for observational error (Eq. 3), kinematic model error (Eq. 2), and dynamical model error (Eq. 1) respectively. In practice we are unlikely to require such a complex expression for numerical evaluation; aggregation of error terms and redundancy between kinematic and dynamical model error usually allows simplifications.

The Bayesian approach encourages us to explicitly quantify our prior knowledge about the parameter values through the prior $p(\Theta)$. In marine ecosystem modelling, we are unlikely to ever consider cases of complete parameter ignorance, where a parameter value could possibly switch sign or get incredibly large. Every parameter is expected to have a value that falls into a credible range, otherwise the associated parameterisation would be difficult to defend. In some cases, when broad uniform or "uninformative" priors are assumed, it may not be necessary to specify exact limits of these distributions as the analyses may become insensitive to these limits once the range becomes sufficiently broad. There are inherent difficulties with the concept of "ignorance" priors: for example, a flat prior distribution over ϕ will correspond to an informative prior for some function $g(\phi)$ (see Cox and Hinkley, 1974 for further discussion). In any case, trying to minimise the impact of prior distributions is rather defeating the object of Bayesian estimation, which explicitly aims to synthesise information from new data with prior information from previous analyses.

Once the likelihood is formulated and a prior distribution is prescribed, classical Bayes estimates (BEs) may be computed from posterior mean or posterior median values of Θ. Assuming the statistical assumptions are correct, these estimators will minimise the mean square error or mean absolute error respectively of the parameter estimate $\widehat{\Theta}$ (e.g., Young and Smith, 2005). To obtain BEs can be computationally expensive, requiring sophisticated techniques to sample efficiently from the posterior distribution (e.g. by Markov Chain Monte Carlo, MCMC, methods). An alternative Bayesian estimator, very widely used in geosciences, is the joint posterior mode or maximum a posteriori (MAP) estimator (e.g., Kasibhatla, 2000; Bocquet, 2014), given by maximising the posterior probability $p(\Theta \mid \boldsymbol{y})$ as a function of Θ. Such estimates are more computationally feasible in large problems where the search for the maximum of the posterior (or the minimisation of its negative logarithm) can be greatly accelerated by techniques such as the variational adjoint (Bennett, 2002, Chapter 4).

In maximum likelihood (ML) estimation we seek the parameter values $\widehat{\Theta}_{\text{ML}}$ that maximise the probability of the data given the parameter set, i.e. $p(\boldsymbol{y} \mid \Theta)$. When considered as a function of Θ, this probability is called the *likelihood* of the parameter values $L(\Theta \mid \boldsymbol{y})$ because it is strictly a probability of the data, not of the parameter values. Indeed, in ML estimation we do not need to consider the parameter values as random variables at all; rather they are considered as fixed, unknown constants. For this reason the '|'s are sometimes replaced by ';'s to emphasise that, in a non-Bayesian context, the likelihood is not a conditional probability in the sense of one set of random variables dependent on another (e.g., Cox and Hinkley, 1974). In the ML approach, no prior information on the parameter values is used except possibly to define upper or lower plausible limits or allowed ranges for the parameter search (Young and Smith, 2005).

Historically, Bayesian methods (Bayes, 1763; Bayes and Price, 1763) predate ML methods of Fisher (1922) by some margin. Fisher introduced ML methods partly to avoid problems in defining prior ignorance (see above) but also to avoid the noninvariance property of Bayesian estimators (Hald, 1999). This property means that given the BE of one parameter $\widehat{\phi}_{\text{B}}$, the corresponding BE of a nonlinear function of that parameter $g(\phi)$ is not simply given by plugging in the estimate ($\widehat{g}_{\text{B}} \neq g(\widehat{\phi}_{\text{B}})$), while for ML estimates the invariance property does hold ($\widehat{g}_{\text{ML}} = g(\widehat{\phi}_{\text{ML}})$). We will see an example of this in Sect. (2.3).

2.2.2 Sequential methods

In some problems, assimilating all the data at once from all available sampling times can be computationally impractical. This is particularly likely for models with stochastic dynamics ($\boldsymbol{\eta} \neq 0$ in Eq. 1), if the data are clustered in time, or if model states need to be repeatedly updated as new data come in. In such cases a sequential approach can be expedient. The basic idea is to break the large integration problem defined by Eq. (7) into a number of smaller problems by sequentially assimilating observations in subsets defined by sampling time. The method comprises a consecutive sequence of two major steps, a forecast- and an analysis step respectively. ~~If the sequential approximation or 'filter' is accurate, it should approximate the posterior distribution defined by Eqs. and 7), when all data have been assimilated by the end of the assimilation period.~~ If the sequential algorithm is accurate, it should approximate the posterior parameter distribution defined by Eqs. (6 and 7) at times where all available data have been assimilated

To see how this works, suppose we know the probability density $p(\boldsymbol{x}_j^t \mid \boldsymbol{y}_{1:j}, \Theta)$ of the true state at sampling time t_j (possibly an initial condition) for a given value of the uncertain parameters Θ and given all the previously assimilated observations $\boldsymbol{y}_{1:j}$

(possibly null). The probability density at sampling time t_{j+1} is given by the forecast density:

$$p(\boldsymbol{x}_{j+1}^t \mid \boldsymbol{y}_{1:j}, \Theta) = \int p(\boldsymbol{x}_{j+1}^t \mid \boldsymbol{x}_j^t, \Theta) \cdot p(\boldsymbol{x}_j^t \mid \boldsymbol{y}_{1:j}, \Theta)\, \mathrm{d}\boldsymbol{x}_j^t \tag{8}$$

In general this integral can be approximated by an ensemble of Monte Carlo simulations, sampling an initial condition from $p(\boldsymbol{x}_{j+1}^t \mid \boldsymbol{y}_{1:j}, \Theta)$ and then running the model to the next sampling time t_{j+1} (possibly including stochastic dynamical noise, and possibly accounting for kinematic model error). Next, in the analysis step, the new observations are assimilated by applying Bayes' theorem:

$$p(\boldsymbol{x}_{j+1}^t \mid \boldsymbol{y}_{1:(j+1)}, \Theta) \propto p(\boldsymbol{y}_{j+1} \mid \boldsymbol{x}_{j+1}^t, \Theta) \cdot p(\boldsymbol{x}_{j+1}^t \mid \boldsymbol{y}_{1:j}, \Theta), \tag{9}$$

which again can be approximated e.g. by Monte Carlo sampling. The forecast and analysis steps can then be repeated until all the data are assimilated. Note that Eq. (9) assumes conditional independence of the observations, allowing us to write $p(\boldsymbol{y}_{j+1} \mid \boldsymbol{x}_{j+1}^t, \Theta)$ instead of $p(\boldsymbol{y}_{j+1} \mid \boldsymbol{x}_{j+1}^t, \boldsymbol{y}_{1:j}, \Theta)$. This amounts to assuming that the observational errors are independent between sampling times (Evensen, 2009), which may not be strictly true if sampling is frequent and if there is a noticeable contribution from representativeness/undersampling, or from errors in conversion factors (see Sect. 2.1.3).

Once the predictive filtering densities $p(\boldsymbol{x}_{j+1}^t \mid \boldsymbol{y}_{1:j}, \Theta)$ have been approximated for all sampling times (t_j with $j = 1, \ldots, N_t$), these can be used to approximate the likelihood in Eq. (7), since:

$$\begin{aligned} p(\boldsymbol{y} \mid \Theta) &= \prod_{j=1}^{N_t} p(\boldsymbol{y}_j \mid \boldsymbol{y}_{1:j-1}, \Theta) \\ &= \prod_{j=1}^{N_t} \int p(\boldsymbol{y}_j \mid \boldsymbol{x}_j^t, \boldsymbol{y}_{1:j-1}, \Theta) \cdot p(\boldsymbol{x}_j^t \mid \boldsymbol{y}_{1:j-1}, \Theta)\, \mathrm{d}\boldsymbol{x}_j^t \\ &= \prod_{j=1}^{N_t} \int p(\boldsymbol{y}_j \mid \boldsymbol{x}_j^t, \Theta) \cdot p(\boldsymbol{x}_j^t \mid \boldsymbol{y}_{1:(j-1)}, \Theta)\, \mathrm{d}\boldsymbol{x}_j^t \end{aligned} \tag{10}$$

For j=1 in Eq. (10) we have a set of zero members and $p(\boldsymbol{y}_j \mid \boldsymbol{y}_{1:j-1}, \Theta) = p(\boldsymbol{y}_1 \mid \Theta)$. The third line of Eq. (10) again assumes conditional independence of the observations and the final integral can in general be approximated using the predictive ensembles (see Jones et al., 2010; Dowd, 2011; Dowd et al., 2014). This procedure can be repeated for different values of Θ and combined with Eq. (6) to assess posterior probability.

Alternatively, $p(\Theta \mid \boldsymbol{y})$ can be calculated from a single application of the filter using a 'state augmentation' approach whereby the parameters Θ are appended to the vector $\boldsymbol{x}$ as additional state variables with zero dynamics. In practice, random parameter noise may need to be added to avoid filter degeneracy, such that this approach may be considered a separate estimation method (Dowd, 2011). However, if such ad hoc noise can be avoided, or if the parameters are in fact assumed to vary stochastically, then the augmented-state filter at the end of the assimilation interval should approximate the theoretical Bayesian posterior for this time. For other times, a 'smoother' algorithm would be required. A further benefit of the augmented-state filter is that the parameter estimates for intermediate time periods may show temporal pattterns that expose deficiencies in the model formulation and provide useful information for model development (e.g., Losa et al., 2003).

The various types of filter differ essentially in terms of how the integrals in Eqs. (8) and (9) are approximated. Particle filters (van Leeuwen, 2009) use Monte Carlo sampling for both steps while the Ensemble Kalman Filter (Evensen, 2003, 2009) uses Gaussian and linear approximations for the analysis step, enabling the use of smaller ensembles but at the cost of lower accuracy in strongly nonlinear/non-Gaussian problems. The (Extended) Kalman Filter applies when the model dynamics are (quasi-) linear and both model and observational errors are Gaussian. These conditions allow both integrals to be evaluated analytically, but appear to be rarely applicable to parameter estimation in marine ecosystem models. For reviews of sequential approaches the reader is referred to Dowd et al. (2014) for marine biogeochemical modelling and to Bertino et al. (2003) for oceanography in general.

2.2.3 Variational methods

At present there appears to be some ambiguity regarding the term "variational" in the context of DA. It is sometimes used to describe approaches explicitly based on control theory or "inverse methods" that may not include explicit assumptions on error distributions and where cost functions are defined a priori, rather than being derived from statistical or probabilistic models. However, a distribution-free approach seems difficult to recommend in general for marine ecosystem model parameter estimation, given the strong nonlinearity, non-Gaussianity, and relatively weak data constraint often encountered in such problems. Within the marine ecosystem modelling community, the term "variational DA" is often used more broadly to refer to all non-sequential methods that involve the minimisation of a cost function, whether or not this is based on a probability model.

In any case, there are some powerful mathematical tools developed for variational DA that can be applied to minimise cost functions. ~~and thereby provide access to joint posterior mode and maximum likelihood estimates.~~ Adjoint methods allow the gradient of the cost function with respect to all fitted parameters to be computed in an extremely efficient manner, see Lawson et al. (1995), and Appendix (C). This is particularly useful when dealing with a large number of fitted parameters (high-dimensional Θ) of computationally expensive models (e.g., Tjiputra et al., 2007). The application of the adjoint method helps reducing the number of model runs to provide access to joint posterior mode and maximum likelihood estimates.

Pelc et al. (2012) provide useful theoretical background for different 4DVar approaches (four-dimensional, in space and time, variational approaches) and show how this adjoint method can be used to estimate ecosystem model parameters jointly with a large number of initial condition parameters. See also Bennett (2002) for an introduction to variational DA and adjoint methods in physical oceanography.

However, it can be disadvantageous to employ a search algorithm that relies too much on local gradients (e.g. from an adjoint model) to minimise the cost function, because this may result in finding a local minimum rather than the global minimum that defines the MAP or ML estimate (Vallino, 2000). This issue appears to be frequently encountered in marine ecosystem modelling applications, and should be expected as a product of strong nonlinearity and weak data/prior constraint. For such cases, a non-local approach such as simulated annealing, following Hurtt and Armstrong (1996, 1999) or a microgenetic algorithm, following Schartau and Oschlies (2003), may be preferable, at least during an initial period of the search before the broader region of the global minimum is located (Ward et al., 2010). The main drawback of these non-local search algorithms is that they tend to require a larger number of model runs (at least order of 10^3) to have a good chance of accurately locating

the global minimum, although they may yet provide meaningful improvements to prior parameter estimates for order of 100 runs (Mattern and Edwards, 2017).

2.2.4 Recent approaches

Much recent interest has focused on combined state and parameter estimation, whereby model parameters Θ are estimated together with a true state $\boldsymbol{x}^t$ (e.g., Simon and Bertino, 2012; Fiechter et al., 2013; Parslow et al., 2013; Weir et al., 2013; Dowd et al., 2014). In the Bayesian approach, model parameters and system state are both random variables. We can therefore apply Bayes' Theorem to the composite random variable $\Psi = (\Theta, \boldsymbol{x}^t)$ and decompose the prior as $p(\Psi) = p(\boldsymbol{x}^t \mid \Theta) \cdot p(\Theta)$ to obtain an expression for the joint posterior:

$$p(\boldsymbol{x}^t, \Theta \mid \boldsymbol{y}) \propto p(\boldsymbol{y} \mid \boldsymbol{x}^t, \Theta) \cdot p(\boldsymbol{x}^t \mid \Theta) \cdot p(\Theta) \tag{11}$$

This equation has so far been applied to stochastic dynamic models with no kinematic model error (cf. Fiechter et al., 2013; Parslow et al., 2013). Equation (6) can be recovered from Eq. (11) by integrating (marginalising) both sides over $\boldsymbol{x}^t$.

In some other recent studies emphasis is put on "hierarchical" error models (Zhang and Arhonditsis, 2009; Parslow et al., 2013; Wikle et al., 2013). Here, the traditional model parameters are replaced with stochastic processes over time and/or space, and parameter identification focuses on the *hyperparameters* that describe the stochastic processes (e.g. means, variances, autocorrelation parameters). This is essentially similar to the case of parameter estimation for a stochastic dynamical model (Sect. 2.2.2) and fits into the general formulation in Sect. (2.1), if we treat the stochastic parameters as additional state variables with dynamical model errors $\boldsymbol{\eta}$. The hyperparameters could in principle be estimated by ML, sometimes referred to as an "empirical Bayesian" approach (Cox and Hinkley, 1974), but it appears that computational tractability may favour the "hierarchical Bayesian" approaches (e.g., Zhang and Arhonditsis, 2009), which may also make use of sequential Monte Carlo methods (e.g., Jones et al., 2010; Parslow et al., 2013).

Another important initiative is the estimation of hyperparameters of the kinematic error model along with the ecosystem parameters (Arhonditsis et al., 2008). The posterior of the kinematic model error provides an estimate of the model discrepancy, introduced by Kennedy and O'Hagan (2001) and originally referred to as model inadequacy. The model discrepancy is defined as the model error for the "true" values of the model parameters, i.e. the unknown values of the parameters for which the model best represents $\boldsymbol{x}^t$. Estimates of model discrepancies may thus provide useful diagnostics for model skill assessment and development.

2.3 From statistical model to cost function

The choice of a suitable estimation method for marine ecosystem model parameters should be mainly based on the availability of relevant prior information, as well as on the basic error assumptions (Eqs. 1, 2, 3). Once the error model and estimation method have been chosen, we can derive the probability densities and cost functions that can be used for parameter estimation.

As a simple but common example, consider a deterministic model with no model error and data with additive Gaussian observational errors, Eq. (4), with known covariance matrix $\mathbf{R}$. We wish to use a total of N_y data, summing over all data types,

to estimate N_Θ parameters by Bayesian estimation. A survey of the literature might lead us to model the prior distribution of Θ as Gaussian with a mean Θ^b and covariance matrix $\mathbf{B}$. From Eq. (6) the posterior density is proportional to a product of the likelihood and the prior density:

$$p(\Theta \mid \boldsymbol{y}) \propto \frac{1}{\sqrt{(2\pi)^{N_y} \det \mathbf{R}}} \cdot \exp\left[-\frac{1}{2}\, \boldsymbol{d}^T \mathbf{R}^{-1} \boldsymbol{d}\right] \cdot \frac{1}{\sqrt{(2\pi)^{N_\Theta} \det \mathbf{B}}} \cdot \exp\left[-\frac{1}{2} \boldsymbol{\Delta}_\Theta^T\, \mathbf{B}^{-1} \boldsymbol{\Delta}_\Theta\right] \tag{12}$$

where the data-model residual $\boldsymbol{d}$ is defined by $\boldsymbol{d} = \boldsymbol{y} - H(\boldsymbol{x})$ (see $\boldsymbol{\epsilon}$ in Eq. 4). The deviation from the prior is $\boldsymbol{\Delta}_\Theta = \Theta - \Theta^b$. A MAP or joint posterior mode estimate of Θ can then be obtained by minimising the cost function $J(\Theta) = -2\, \log p(\Theta \mid \mathbf{y})$ + constant, given by:

$$J(\Theta) = \boldsymbol{d}^T \mathbf{R}^{-1} \boldsymbol{d} + \boldsymbol{\Delta}_\Theta^T\, \mathbf{B}^{-1} \boldsymbol{\Delta}_\Theta \tag{13}$$

where constant terms (since independent of Θ) have been dropped.

Alternatively, nonnegativity constraints on the variables and parameters may lead us to prefer the lognormal observational error model. Likewise, we can assume lognormal priors for the parameters. In this case the posterior density becomes:

$$p(\Theta \mid \boldsymbol{y}) \propto \frac{1}{\sqrt{(2\pi)^{N_y} \det \tilde{\mathbf{R}}\, \prod_j y_j}} \cdot \exp\left[-\frac{1}{2}\, \tilde{\boldsymbol{d}}^T \tilde{\mathbf{R}}^{-1} \tilde{\boldsymbol{d}}\right] \cdot \frac{1}{\sqrt{(2\pi)^{N_\Theta} \det \tilde{\mathbf{B}}\, \prod_l \Theta_l}} \cdot \exp\left[-\frac{1}{2} \tilde{\boldsymbol{\Delta}}_\Theta^T\, \tilde{\mathbf{B}}^{-1} \tilde{\boldsymbol{\Delta}}_\Theta\right] \tag{14}$$

where the data-model residuals and parameter corrections on the transformed scale are defined by $\tilde{\boldsymbol{d}} = \log(\boldsymbol{y}) - \log(H(\boldsymbol{x})) + \frac{\tilde{\boldsymbol{\sigma}}^2}{2}$ and $\tilde{\boldsymbol{\Delta}}_\Theta = \log(\Theta) - \log(\Theta^b) + \frac{(\tilde{\boldsymbol{\sigma}}^b)^2}{2}$. A MAP estimator of Θ is then obtained by minimising:

$$J(\Theta) = \tilde{\boldsymbol{d}}^T \tilde{\mathbf{R}}^{-1} \tilde{\boldsymbol{d}} + 2 \sum_{l=1}^{N_\Theta} \log(\Theta_l) + \tilde{\boldsymbol{\Delta}}_\Theta^T\, \tilde{\mathbf{B}}^{-1} \tilde{\boldsymbol{\Delta}}_\Theta \tag{15}$$

The MAP or posterior mode estimator of $\log(\Theta)$ is equivalent here to the posterior median estimate and is obtained by maximising $p(\log(\Theta) \mid \boldsymbol{y})$. This leads to a cost function given by Eq. (15) without the second term, $2\sum_{l=1}^{N_\Theta} \log(\Theta_l)$ (cf., Fletcher, 2010). Due to the noninvariance property of Bayesian estimates, the exponent of the MAP estimator of $\log(\Theta)$ will generally differ from the MAP estimator of Θ. By contrast, ML estimates are obtained by minimising the cost functions without any of the prior terms (second terms in Eq. 13, second and third terms in Eq. 15). In each case the same ML estimator for Θ is obtained whether we use Θ or $\log(\Theta)$, as expected from the invariance property of ML estimates.

2.4 Remarks on data assimilation terminology

We close this section with some cautionary remarks about different terminology that the reader may encounter the literature. First, many DA papers and textbooks start by assuming a certain cost function, based on variational or optimal control theory,

rather than deriving it from a probabilistic treatment as herein (e.g., Le Dimet and Talagrand, 1986; Bennett, 2002; Fletcher, 2010). These studies tend to refer to MAP estimates obtained by minimising cost functions such as Eq. (13) as "weighted least squares estimates". However, any analogy with regression analysis is stretched because these estimates are fundamentally dependent on, and potentially biased by, the assumed prior distributions. Second, many DA papers and textbooks use the term "likelihood" to refer to the posterior probability $p(\Theta \mid \boldsymbol{y})$ in Eq. (6), and the term "maximum likelihood estimators" although modifiers such as "(Bayesian)" (Jazwinski, 2007, page 156) or "(posterior)" (Tarantola, 2005, page 40) are sometimes added. This obscures the fact that posterior mode estimators, like all BEs, are dependent on assumed prior distributions. Maximum likelihood avoids this dependence, but in doing so tends to be unsuitable for high-dimensional parameter estimation in the partially-observed systems typically encountered in oceanography and geophysics.

3 Typical parameterisations of plankton models and their parameters

Deviant parameter estimates of a model may point towards a deficiency in model structure, forcing, or in boundary conditions. Estimates of the effectively same parameters may turn out to be different within dissimilar plankton ecosystem models, even if those models may have been calibrated with the same data and although they possibly share an identical physical (environmental) setup. To understand why parameter estimates can be different it is helpful to unravel some of the basic differences between major parameterisations that describe growth and loss rates of phytoplankton.

A crucial element of most plankton ecosystem models is the description of phytoplankton growth as a function of light, temperature, and nutrient availability. How growth of algae is parameterised is relevant and the associated parameter values affect timing and intensity, e.g. of a phytoplankton bloom in model solutions.

3.1 Differences between maximum carbon fixation and maximum growth rate

The build up of phytoplankton biomass depends on how much of the available nutrients can be utilised and how much energy can be absorbed from sun light. Under nutrient-replete and light-saturated conditions, the carbon fixation (gross primary production, GPP) reaches a (temperature dependent) maximum rate, described as a parameter (P_m^C) with unit d^{-1}. For models that do not resolve mass flux of carbon explicitly, P_m^C is substituted by a maximum growth rate (μ_m) to express the phytoplankton's maximum assimilation rate of nitrogen (N), or of phosphorus (P). The maximum GPP and the maximum growth rate are interrelated and in principle one can be derived from the other (Smith, 1980). In reality, maximum C-fixation, maximum N- or P-assimilation, and cell doubling rates are highly variable. This requires at least cellular C, N and Chl*a* to be explicitly resolved, (linking for example, intracellular nutrient allocation to photoacclimation Shuter, 1979; Laws et al., 1983; Pahlow, 2005; Armstrong, 2006).

In practice an analogy between P_m^C and μ_m is often assumed in N- or P- based biogeochemical models (assuming fixed stoichiometric elemental C:N:P ratios for algal growth). The parameter P_m^C or μ_m is typically multiplied with a dimensionless temperature function (f_T) (e.g., Arrhenius, 1889; Eppley, 1972), allowing for temperature induced changes of metabolic rates. The actual potential maximum rate ($P_m^C \cdot f_T$ or $\mu_m \cdot f_T$) is then reached at some prefixed reference or optimum temperature

accordingly. In early N-based plankton modelling studies (e.g., Evans and Parslow, 1985; Fasham et al., 1990; Doney et al., 1996) the maximum growth rate was mainly adopted from Eppley (1972). In subsequent DA studies this maximum rate was either subject to optimisation (e.g., Fasham and Evans, 1995; Spitz et al., 2001) or it was kept fixed because then parameter values of the limitation functions could be better identified (Matear, 1995; Fennel et al., 2001).

3.2 Combining parameterisations of light- and nutrient limitation

In many marine ecosystem models two separate limitation functions are combined: one that expresses the photosynthesis versus light relationship (P-I curve) and another that describes the dependence between ambient nutrient concentrations and nutrient uptake. The two functions are similar in their characteristics, starting from zero (no light or no nutrients) and approaching saturation at some high light and at replete nutrient concentration. Three approaches are generally found in marine ecosystem models to limit algal growth by photosynthesis and nutrient uptake. The first is to apply Blackman's law (Blackman, 1905), assuming that growth is reduced by the most limiting factor, either by light or by nutrient availability (e.g., Hurtt and Armstrong, 1996; Oschlies and Garçon, 1999; Klausmeier and Litchman, 2001). The second is to multiply both limitation functions (e.g., Evans and Parslow, 1985; Fasham et al., 1990; Follows et al., 2007). The third approach involves ~~more complex representations of growth~~ combinations of light- and nutrient limitation that ~~, as it accounts for~~ resolve interrelations between cell quota, N-uptake and the photoacclimation state of the algae (e.g., Armstrong, 2006, see Sect. 3.5). ~~Here, photoautotrophic growth depends on the cellular C:N (or N:C) ratio and the mass distribution of phytoplankton C and N has to be explicitly resolved in the model.~~ Whether the first, second or third approach is considered can be expected to affect estimates of the associated parameter values.

3.3 Photosynthesis as a function of light (P-I curve)

In a P-I curve the level of increase from low to high irradiance is specified by the initial slope parameter (the maximum of the first derivative of the P-I curve with respect to light), also referred to as photosynthetic efficiency (α_{phot}) (Smith, 1936; Jassby and Platt, 1976; Cullen et al., 1992; Baumert, 1996). Photosynthetic efficiencies were derived from P-I measurements, for example by Platt and Jassby (1976), Peterson et al. (1987), and Platt et al. (1992) and their mean values were used for many N-based models (e.g., Fasham et al., 1990; Sarmiento et al., 1993; Doney et al., 1996; Oschlies and Garçon, 1999). Published measurements of α_{phot} were typically normalised to Chl*a* concentrations. In case of N- or P-based models careful considerations are then needed with respect to the phytoplankton's cellular Chl*a* content, which can vary by a factor of ten and more. Values of α_{phot} were found to vary by a factor of three (Côté and Platt, 1983) during a three month period, which can be attributed to changes in phytoplankton community structure as well as to photoacclimation. Platt and Jassby (1976) reported an even larger variational range over a one year period, from α_{phot} = 0.03 to 0.63 mg C (mg Chl*a*)$^{-1}$ h^{-1} W^{-1} m^2 within the upper ten meters.

3.4 Algal growth and nutrient limitation

Typical parameterisations of growth limitation by nutrient availability (ambient nutrient concentrations) are expressed with the half-saturation constant (K_s) of a classical Monod equation (Monod, 1942, 2012). Another approach is to parameterise limitations of the nutrient uptake rate, described with a parameter referred to as nutrient affinity (α_{aff}) (Aksnes and Egge, 1991). The affinity based parameterisation may also be applied to describe nutrient-limited growth, assuming that the rates of nutrient uptake and growth are balanced. In this case both parameters (K_s and α_{aff}) can be interpreted as being interrelated $\alpha_{\text{aff}} = \mu_m \cdot f_T \,/\, K_s$. However, α_{aff} is derived from mechanistic considerations that are fundamentally different from former interpretations of K_s of a Monod equation (Pahlow, 2005; Armstrong, 2008; Pahlow and Oschlies, 2013; Fiksen et al., 2013). For comparison between estimates of α_{aff} it is important to know whether this parameter describes limitation of growth or of nutrient uptake. The description of nutrient limited growth with the Monod equation, thereby retrieving values for K_s from measurements, had been discussed in the past (e.g., Eppley et al., 1969; Falkowski, 1975; Burmaster, 1979; Droop, 1983). This discussion regained attention during recent years and the sole application of the Monod equation is currently viewed as a considerable drawback when simulating plankton growth under transient (unbalanced growth) conditions (Flynn, 2003; Smith et al., 2009; Franks, 2009; Smith et al., 2014, 2015).

3.5 Algal growth and intracellular acclimation

More complex growth ~~inter~~dependencies ~~between light and nutrient limitation~~ are ~~resolved by~~ described with models that ~~account for~~ consider intracellular acclimation dynamics (e.g., Geider et al., 1998; Pahlow, 2005; Armstrong, 2008; Wirtz and Pahlow, 2010). In these models, photoautotrophic growth rates become dependent on cell quota, e.g. usually normalised to carbon biomass (N:C), and the amount of synthesised Chl*a* per cell. With such approaches, the changes of the mass distribution of phytoplankton C and N, as well as the cellular Chl*a* content, have to be explicitly resolved in the model. One advantage is that these models are more sensitive to variations in light conditions and nutrient availability. ~~These approaches~~ The respective equations involve physiological parameters that are related but not identical to those of classical N- or P-based growth models, which impedes a direct comparison of older estimates of growth parameters with values currently used in models with acclimation processes resolved.

3.6 Losses of phytoplankton biomass

Parameterisations of phytoplankton cell losses involve lysis (starvation and/or viral infection), the aggregation of cells together with all other suspended matter, and grazing by zooplankton. Exudation and leakage are processes of organic matter loss that occur while the physiology of the algae is functional. Cell lysis, exudation and leakage are usually expressed as a single rate parameter and this loss of organic matter is assumed to be proportional to the phytoplankton biomass.

Parameterisations of phytoplankton losses due to the process of coagulation and sinking of phytoplankton and detrital aggregates are basically derived from the principle theory of coagulation. The application of coagulation theory to simulate phytoplankton aggregation is well established for models that resolve size classes of particles (of phytoplankton cells and de-

tritus) explicitly (Jackson, 1990). But the representativeness of simplifications (e.g. reduction to two size classes) assumed for model simulations remains an open task (e.g., Ruiz et al., 2002; Burd and Jackson, 2009). Aggregation parameters in marine ecosystem models are often assumed to represent the combination of a collision rate and the probability of two particles sticking together after collision (e.g. stickiness of algal cells). These two parameters, collision rate and stickiness, are multiplied with each other to yield a final aggregation rate. They are therefore difficult to estimate separately. Unless prior information can be used their estimates are always collinear, which suggests to estimate their product instead (as done in example in Sect. 5.4).

A common problem is to find constraints that allow for a clear distinction between phytoplankton losses due to the export of aggregated cells and the loss because of grazing. Both processes can be responsible for the drawdown of phytoplankton biomass, and data that cover the onset, peak and decline of a bloom are needed for a possible distinction. How the complex nature of predator-prey interaction is parameterised remains a critical element of plankton ecosystem models. Compared to the approaches that describe algal growth an even larger number of different parameterisations exist for grazing (Gentleman et al., 2003). Experimental data of grazing rates and collections of field data of zooplankton abundance are therefore of great value.

Elaborate analyses of meso- and microzooplankton biomass, grazing and mortality rates were done by Buitenhuis et al. (2006, 2010). For their two studies they compiled an extensive database with laboratory and field measurements. With their data syntheses they could derive parameter values for simulations with a global ocean biogeochemical model. Furthermore, independent field data, not used to derive the meso- and microzooplankton parameter values, were considered for assessing the performance of their model on global scale. Their work reflects the large effort that can be dedicated to this topic for achieving reliable simulation results of zooplankton grazing.

The explicit distinction between zooplankton size classes, like meso- and microzooplankton, was bypassed in Pahlow et al. (2008). Their model allows for omnivory within a the zooplankton community, which is resolved by introducing adaptive food preferences. These preferences are treated as trait (property) state variables that adapt to the relative availability of different prey. This reduces the number of parameters needed to describe a variety of different behaviour in grazing responses. Field data from three ocean sites in the North Atlantic were used by Pahlow et al. (2008) for calibrating their plankton model. They conducted a two-step approach for parameter optimisation. First they optimised parameter values so that depths and dates of minimum and maximum observed values become well respresented by their model at all three sites. In a second step they refined their parameter estimates by minimising weighted data-model residuals. After parameter optimisation they identified distinctive complex patterns between zooplankton grazing and plankton composition for the three simulated ocean sites. Besides their phytoplankton grazing losses it turned out that their optimal estimates of photo-acclimation and maximum C-fixation (α_{phot}, P_m^C) agree with those values derived from model calibrations with laboratory data.

3.7 Constraining simulations of algal growth with laboratory and mesocosm data

Parameter values of acclimation models have typically been adjusted to explain laboratory measurements (Geider et al., 1998; Flynn et al., 2001; Pahlow, 2005; Armstrong, 2006; Smith and Yamanaka, 2007a; Pahlow and Oschlies, 2009; Wirtz and Pahlow, 2010). So far, there is a limited number of experimental studies whose data were used to calibrate these acclimation

models (Laws and Bannister, 1980; Terry et al., 1983, 1985; Healey, 1985; Flynn et al., 1994; Anning et al., 2000). Model calibrations were usually done by tuning parameter values so that model solutions provide a qualitative good fit to the laboratory data. In many cases the parameter adjustments relied on the researchers' experience and intuition, sometimes accounting for prior parameter values obtained from preceding model analyses (e.g., Flynn et al., 2001). Analyses of parameter uncertainties of recent acclimation models are often lacking. Most laboratory modelling studies had put emphasis on the physiological mechanistic model behaviour while error assumptions for quantitative data-model comparison were hardly considered.

Explicit error assumptions for parameter optimisations and for comparisons of acclimation model results with laboratory data were introduced by Armstrong (2006) and by Smith and Yamanaka (2007a). In both studies additive uncorrelated Gaussian observational errors were assumed and optimised results of different model versions had been compared. Armstrong (2006) applied a "simulated annealing" algorithm (Metropolis et al., 1953) to fit his optimality-based model version to the data of Laws and Bannister (1980). The same data were used to also fit the model of Geider et al. (1998) and he evaluated the likelihood ratio of the two ML estimates, to discuss and underpin the improved performance of his refined acclimation parameterisations. Smith and Yamanaka (2007a) also compared the performance of two acclimation models, of Geider et al. (1998) and Pahlow (2005) respectively. Optimal parameter values for the two model versions were obtained with the MCMC method, minimising the misfit between model results and data of the Flynn et al. (1994) experiment. Apart from mechanistic considerations, Smith and Yamanaka (2007a) concluded that the models of Pahlow (2005) and Geider et al. (1998) were describing the assimilated data equally well, since both cost function minimum values were comparable. However, the simulated N:C and Chl*a*:N ratios of the model proposed by Pahlow (2005) were in much better agreement with observations during the exponential growth phase, which remained undifferentiated by their error model (assuming C, N and Chl*a* data to be independent). Different considerations for error models will be addressed hereafter in Sect. (4).

To collect diverse data that fully resolve onset, peak and decline of an algal bloom at ocean sites is difficult to achieve. Data derived from remote sensing, e.g. Chl*a* concentration and primary production rates, provide limited information to explain relevant differences between processes described before, like N-utilisation, fixation and release of C, and synthesis and degradation of Chl*a*. Mesocosm experiments that enclose a large volume of a natural plankton and microbial community can be helpful in this respect, if they provide a good temporal resolution of the exponential growth phase as well as of the post-bloom period. Vallino (2000) highlighted the benefits of using mesocosm data to test plankton ecosystem models, as done before by Baretta-Bekker et al. (1994, 1998). One advantage is that mesocosms are, apart from the surface, closed systems and measurements of inorganic nutrients, dissolved and particulate organic matter should, in principle, add up to approximately constant concentrations of total nitrogen and total phosphorus. Total carbon concentrations may only vary due to air-sea gas exchange. By design these experiments often integrate valuable series of joint and parallel measurements, yielding detailed data from various scientist with different expertise (e.g., Williams and Egge, 1998; Riebesell et al., 2008; Guieu et al., 2014). Drawbacks are uncertainties in initial conditions and also the representativeness of mesocosm data to reflect the real dynamics in the ocean is subject to discussion (e.g., Watts and Bigg, 2001). In spite of these limitations, simulations of mesocosms or of enclosures experiments (e.g. with large carboys deployed in the field) have helped to identify credible model parameter values and assess model performance. This is particularly true for tracing microbial dynamics (Van den Meersche et al., 2004; Lignell et al.,

2013) or for details in the composition and fate of particulate organic carbon and nitrogen (POC and PON) (Schartau et al., 2007; Joassin et al., 2011).

In contrast to laboratory measurements, data from mesocosm experiments reflect some natural variability of the plankton community, mainly captured by replicate mesocosms. The availability of measurements from replicate mesocosms is also helpful when defining error models that specify the statistical treatment of the data used for parameter estimation.

4 Error models

Error models define our assumptions about uncertainties and the statistical relationships between observed data, the true state, model output, model inputs (forcings and initial/boundary conditions), and model parameters. Here we review error models that have been applied to address the various sources of uncertainty in marine ecosystem models and consider their implications for parameter identification. An explicit treatment of each source of uncertainty may not be necessary but we do recommend to reflect on how these uncertainties can be accounted for when modelling plankton dynamics and biogeochemical cycles.

4.1 Uncertainty in observations

The simplest and most common models for observational error assume that the observational errors ϵ are: i) additive normal, ii) constant variance between samples, and iii) independent between samples and variable types. Such models are also commonly used to represent aggregated errors accounting for both observational and kinematic model error (see Sect. 2.1); we will refer to these as *residual* errors.

The additive normal assumption (i) is straightforward but also restricted, as it does not capture three common characteristics of some ecosystem data such as Chl*a* concentrations: 1) larger values tend to have larger errors, 2) values cannot be negative, and 3) the error distribution has positive skew. Characteristic (1) may be captured by scaling the standard error with modelled values (e.g., Hurtt and Armstrong, 1996) or with observed values (e.g., Harmon and Challenor, 1997), while characteristic (2) can be resolved using truncated error distributions (e.g., Hooten et al., 2011). All three characteristics together can be captured by gamma distributions (Dowd, 2007) or power-normal distributions whereby normality is assumed on a power-transformed scale (Freeman and Modarres, 2006). The power-normal family includes log-normal (e.g., Hemmings et al., 2003) and square-root normal models (e.g., Fasham and Evans, 1995).

For power-normal, gamma, or proportional error assumptions we have the difficulty that the variance on the original scale approaches zero at low values. This may be unrealistic, at least in regard to instrumental noise. In normal models this problem can be addressed by adding a constant term to the variance (Schartau et al., 2001; Schartau and Oschlies, 2003) or standard deviation (Vallino, 2000). Another difficulty is that transform-normal models may require unbiasing factors when assuming unbiased errors on the original scale (e.g. $\exp -\tilde{\sigma}^2/2$ for the log-transform). More flexible models may be obtained by e.g. fitting the power transform parameter (Box and Cox, 1964), assuming generalised Gaussian distributions (Tarantola, 1987; Evans, 2003), or using 'anamorphic' transformations (Bertino et al., 2003; Simon and Bertino, 2012). It is yet unclear whether such extra flexibility is generally necessary, but it has been demonstrated that the choice of transformation can strongly af-

fect estimates of plankton ecosystem fluxes (Evans, 2003) and that a good choice can improve parameter estimation in twin experiments (see Fig. 1 and Simon and Bertino, 2012).

The validity of the constant variance assumption (ii) may be improved by a scale transformation, although the transformation that best normalises the error distribution (see above) may not best promote the homogeneity of variance. Spatiotemporal variations in the error variance may naturally occur, for example due to seasonal modulations of the unresolved variability and hence the representativeness error component. Accounting for this variation should improve parameter estimates and uncertainty assessment (cf., Hemmings and Challenor, 2012), but in applications this has rarely been attempted (Hemmings et al., 2003; Dowd, 2007).

In some contexts e.g. mesocosms, the error covariance matrix might be estimated from experimental replicates prior to fitting the model (Sect. 5.4). In problems where sampling is sparse and/or when the model error contribution is large, the error variances may not be estimable from data alone (Evans, 2003). Here the variances may instead be parameterised and estimated jointly with the ecosystem model by Bayesian or ML estimation, which has been done in few studies (Hurtt and Armstrong, 1996, 1999; Stock et al., 2005; Malve et al., 2007; Lignell et al., 2013).

The assumption of independent errors between samples and variable types (iii) can be invalidated in cases where contributions from representativeness error or kinematic model error are large, or where the data have been derived by interpolation or application of a regression model. Neglected correlation may result in parameter estimates that are less efficient (higher variance) and more strongly correlated (e.g. see example in Sect. 5.4). Pre-averaging the data is somewhat helpful to promote independence (and normality, via the Central Limit Theorem), but might also remove some of the informative variability. One common *ad hoc* intervention in the cost function is to scale the residual error variance with the sample size of each data type, to avoid biasing the fit in favour of better-sampled variables (e.g., Schartau and Oschlies, 2003; Friedrichs et al., 2007). More formal treatments have fitted parameterisations of the error correlations jointly with the ecosystem model (e.g., Stock et al., 2005; Arhonditsis et al., 2008).

Whatever the assumptions of the observational/residual error model, it is possible to test their validity using the assimilated data, either by analysing the residuals and performing lack-of-fit tests (Bennett, 2002, p43; Stock et al., 2005; Wallhead et al., 2014) or by comparing fit statistics with those obtained under alternative error models (using e.g. likelihood ratio tests, information theoretic or Bayesian criteria, see Sect. (6.2).

Finally, we caution that certain interpolated or derived data may strictly invalidate the observational error model, not only due to error correlation (see above), but also due to the introduction of *smoothing bias*. Data interpolated onto a model grid will tend to systematically underestimate true values where they are high and overestimate them where low; an effect that will be difficult to account for in the observational error model. In this situation parameter estimates can become biased towards values that suppress spatiotemporal variability in plankton dynamics. Similarly, if the data are derived from a regression model, these estimates may also "trim the peaks and fill the valleys", because in a regression model (e.g. $y = a_0 + a_1 p + \epsilon$, where p is some predictor data) there is always some part of the true variability that is included in the error term, and therefore subject to smoothing bias. In principle this could be avoided by including an inverted regression relationship in the operator O and assimilating the "raw" predictor or proxy data instead of the regression-based estimates.

4.2 Prior uncertainty in Θ

Prior uncertainty plays an important role in estimating model parameters. Typically, there is not enough information in the assimilated data to constrain all parameters of a biogeochemical model. The results may well be sensitive to the "error model of prior uncertainty". Prior uncertainty can be represented by prior probability densities in Bayesian approaches or plausible ranges in non-Bayesian approaches. To account for nonnegativity constraints, prior distributions typically include lognormal (Parslow et al., 2013), square-root normal (Gunson et al., 1999), or beta distributions (Dowd and Meyer, 2003), although normal distributions may yet be applicable for parameters that are well constrained above zero (Parslow et al., 2013). To our knowledge no application has yet incorporated prior correlations between parameters in Θ (i.e. off-diagonal terms in matrix $\mathbf{B}$ introduced in Sect. 2.3). This is surprising, given the fact that posterior uncertainty assessments consistently reveal strong correlations (e.g., Matear, 1995; Prunet et al., 1996; Fennel et al., 2001; Faugeras et al., 2003; Kreus and Schartau, 2015).

Quantifying the prior uncertainty in Θ is often difficult due to: 1) the existing diversity of model structure, functional forms used in the various parameterisations, and definitions of model state variables, and 2) the intrinsic variability between assimilated data sets in terms of taxonomic composition of the plankton community vs. (usually monospecific) laboratory cultures. As a result, it may not be advantageous to simply set the prior uncertainty in Θ_l as the posterior uncertainty from one previous study. A more common approach is to first gather best estimates of Θ_l from a series of previous studies that included parameterisations and state variable definitions sufficiently consistent with the present, and then treat these as unbiased data from which a prior distribution or plausible range can be determined.

When posterior uncertainty becomes unacceptably high, it can be reduced by reducing the prior uncertainty in Θ, and there are several strategies for doing this. First, we should incorporate further data, perhaps of a qualitative nature, into the prior constraints. For example, if it is known *a priori* that certain species or functional groups coexist in certain regions at certain times of the year, then any Θ resulting in competitive exclusion of one of these groups might be ruled out *a priori*. Another possibility within the Bayesian paradigm is to incorporate the subjective opinion of experts (O'Hagan, 2006). A second strategy is to model statistical structure in the prior parameter values, and thereby fill in missing prior parameter estimates for certain species included in the modelled species or groups. Examples here include the use of allometric scaling relationships with cell size (e.g., Edwards et al., 2012) and phylogenetic relationships derived from stochastic modelling of trait evolution (Bruggeman et al., 2009; Bruggeman, 2011). Third, we may seek to reduce the model complexity in terms of the number of free parameters, thereby removing poorly-constrained parameters and parameter correlations that may act to inflate the posterior uncertainty. This may be achieved using sensitivity analysis (e.g., Friedrichs, 2001; Garcia-Gorriz et al., 2003; Hemmings et al., 2003) or model selection criteria (e.g., Ward et al., 2013). A risk here is that parameter estimates and uncertainty assessment may be compromised if model selection uncertainty is not properly accounted for (Burnham and Anderson, 2002). Fourth, it may be possible to reformulate the model in such a way that the prior parameter uncertainty is reduced. For example, a hierarchical model in which parameters vary randomly over space (Zhang and Arhonditsis, 2009) or time (Parslow et al., 2013) may enable the use of stronger prior constraints on the distributional parameters describing this variability (i.e. the 'hyperparameters'). Similarly, a stochastic trait-based approach (e.g., Follows et al., 2007) may employ distributional parameter values that are

better known a priori than values for individual species or functional groups, although such a reduction in prior uncertainty has not yet been clearly demonstrated in the literature.

4.3 Uncertainty in initial conditions (ICs)

Dynamical marine ecosystem models are usually specified by differential equations that are first-order in time, and therefore require for solution one initial condition (IC) for each grid cell or spatial location in the model. These inputs are, in general, uncertain, and liable to impact the model output, at least during a transient relaxation period, or indefinitely if the uncertainty spans more than one basin of attraction of the dynamical system or if the model dynamics are chaotic (e.g., Huisman and Weissing, 1999).

In some cases it is possible to neglect IC error because of accurate measurements, or because a steady state (equilibrium or seasonal cycle) that is only sensitive to Θ can be assumed. Caution is required if neglecting IC uncertainty because initial concentrations are known to be small (e.g. in January); small absolute errors may be large relative errors that can still affect e.g. timing and magnitude of a spring bloom (Evans and Parslow, 1985).

In non-spatial (0D) models, IC errors have been modelled as both fixed parameters (e.g., Vallino, 2000) and as random variables (Bayesian parameters) with specified prior distributions (e.g., Arhonditsis et al., 2008). In mesocosm studies, ICs can play a critical role in determining the model trajectory, and can comprise a large proportion of the fitted parameters (e.g., Lignell et al., 2013). For spatial models, it seems necessary to limit the degrees of freedom of the IC uncertainty (Li et al., 2006), e.g. by using a Bayesian error model with spatial covariance in the prior (Smith et al., 2009; Pelc et al., 2012). To model IC uncertainty, Gaussian distributions are most often employed, often with a log transform to improve realism of the distributional form (see Sect. 4.1). For systems with strong physical control, it may be possible to limit IC uncertainty to only the physical variables, allowing this to generate biochemical uncertainty over an initial burn-in period (Natvik and Evensen, 2003; Simon et al., 2015).

4.4 Uncertainty in forcings and boundary conditions (BCs)

Marine ecosystem models are usually modulated by time and space-dependent environmental drivers (forcings) and boundary conditions that are not predicted by the model dynamics but are necessary inputs to determine the evolution of the model state variables. Studies have demonstrated the sensitivity of biogeochemical variables to errors in bottom-up forcings such as wind stress and vertical mixing (e.g Evans, 1988; Friedrichs et al., 2006; Béal et al., 2010; Sinha et al., 2010) and top-down forcings such as fishing (e.g., Heath, 2012). BC errors may have little impact on variables strongly controlled by internal dynamics at sufficient distance from the boundaries, but they may become critical if they affect internal system constraints such as the supply of limiting nutrients or fluxes of heat/salinity that drive internal circulation and stratification.

There are basically two approaches to modelling the effects of BC/forcing error: 1) to consider individual or net impacts on model dynamics as dynamical model errors ($\boldsymbol{\eta}$ in Eq. 1), thus requiring a stochastic model, or 2) to consider the net impacts on state variables as kinematic model errors ($\boldsymbol{\zeta}$ in Eq. 2), which may permit a deterministic model. The dynamical approach (Eq. 1) is arguably more realistic, more likely to generate realistic temporal correlations and cross-correlations, and accounts

for time and parameter-dependent variation in the form and correlation structure of the joint state variable probability density. It also allows individual error sources to be considered separately. However, approaches based on stochastic models can be computationally intensive and methodologically complex, and parameterising all individual sources of BC/forcing error poses a major challenge. Rather than attempting a comprehensive treatment, current approaches tend to restrict the dynamical noise to certain key sources such as the atmospheric forcing (Natvik and Evensen, 2003; Simon and Bertino, 2009) or surface irradiance and background light attenuation (Torres et al., 2006; Ciavatta et al., 2011), and/or they model the net effect of BC/forcing errors and structural errors synthetically as additive (e.g., Losa et al., 2003, 2004) or multiplicative (e.g., Dowd and Meyer, 2003; Weir et al., 2013) perturbations. It may be questioned to what extent the simple parameterisations used to describe these noise processes accurately describe the net or individual error sources, and it can be difficult to constrain the distributional parameters *a priori*, especially if the structural component is important. Hierarchical filtering methods may allow these "hyperparameters" to be estimated jointly with the other parameters (Jones et al., 2010) but these may incur a computational cost that is prohibitive for spatial models at present.

The kinematic approach (Eq. 2) offers an immediate computational saving because the integral over model error configurations (over $\boldsymbol{x}^t$ in Eq. 7) can usually be performed analytically, such that accounting for model error may amount to simply adding variance and correlation structure to the observational error covariance matrices. However, this may require a more complex parameterization of the error covariance that may still not properly capture seasonal or ecosystem parameter dependence (Stock et al., 2005; Arhonditsis et al., 2008; Zhang and Arhonditsis, 2009). Hemmings and Challenor (2012) demonstrated a Monte Carlo simulation approach to determine the variability of kinematic error variances due to BC/forcing error, but without accounting for correlations or θ_e-dependence. Note that with a deterministic model, and model error treated kinematically, the ecosystem parameters θ_e will likely be optimised to reproduce the ensemble-mean or ensemble-median behaviour of the true system. This may be convenient for future simulations, but it may also result in biases when using previous parameter estimates from laboratory experiments or stochastic model data assimilations to constrain the prior uncertainty in θ_e (see Sect. 4.2).

In either case, BC/forcing error models may fall short in describing potential errors in *phase*, like the timing of nutrient depletion. Model solutions that predict the right sequence of events (e.g. a plankton bloom) but with slightly wrong timing or spatial location, perhaps due to phase error in the atmospheric forcing or ocean circulation, may suffer a double penalty due to changes where none occur in the data and no change where the data do vary. DA may then "smooth out" the model variability in order to minimise this double penalty (Wallhead et al., 2006; Ravela et al., 2007). The problem of phase/timing error has received substantial attention in numerical weather forecasting and geophysical DA (e.g., Hoffman et al., 1995; Lawson and Hansen, 2005; Mittermaier, 2007; Ravela et al., 2007; Ziegeler et al., 2012) and has been highlighted as an issue for marine ecosystem models (Schartau and Oschlies, 2003; Friedrichs et al., 2006). A simple remedy is to average the data and model over larger spatio-temporal scales in the data assimilation (e.g., Schartau and Oschlies, 2003), but again this may remove informative variability and result in a $\widehat{\Theta}$ that is only suited to those larger scales. Wallhead et al. (2006) explored a more explicit approach assuming random time lags between the true state and model state i.e. kinematic model errors in phase, which can be expressed as $\boldsymbol{\zeta}(\theta_\zeta)$ in Eq. (2) (see Appendix A). This may improve the bias and variance of ecosystem parameter estimates compared to a simpler approach assuming only additive residual error (Wallhead et al., 2006, Table A).

For some problems, in particular for chaotic systems, the phase noise may be too intense or ill-defined to allow effective use of a parametric phase lag model. A better approach here might be to use a 'synthetic likelihood' (Wood, 2010), whereby the raw data and model output are replaced with a carefully chosen, informative set of phase-insensitive summary statistics (e.g. means, standard deviations, and lag correlations; cf., Heath, 2012). This approach could incorporate the comparison of modelled vs. observed Fourier spectra and cross-spectra/coherences (e.g., Powell et al., 2006). Whether the statistics e.g. of spectral slopes by themselves provide good constraint on ecological parameters should be tested since it may not be sufficient (Armi and Flament, 1985; Martin, 2003; Franks, 2005).

4.5 Uncertainty in model formulation and structure

Even with perfectly-known parameters, forcings and initial/boundary conditions, we would still not expect the modelled fluxes such as primary productivity and grazing to perfectly reproduce the true fluxes, or the state variables to perfectly follow the true variability. Aggregation of species into model functional groups, effects of finite spatial and temporal resolution, and inherent approximations in the flux parameterisations and model structure may all contribute to "structural error" in the model dynamics.

One promising approach to account for structural error is to add stochastic noise (dynamical model errors) to the ecosystem model parameters θ_e (see Sect. 7). This preserves mass conservation and may allow information on the temporal (e.g. seasonal) variability of species composition within functional groups to be utilised within the stochastic process parameters (e.g., Parslow et al., 2013). However, as with explicit treatments of BC/forcing error (see Sect. 4.4), a comprehensive treatment of all sources seems likely to result in an overparameterised error model and appears to be not yet attempted. An alternative (or complementary) approach is to treat the structural errors as synthetic dynamical or kinematic model errors, with one noise process for each state variable. Here it seems the challenges are to control mass conservation and to find some efficient way to constrain the distributional parameters *a priori* or *a posteriori*.

We note that some structural errors may impose persistent or intermittent biases in the model output that may not be amenable to a simple statistical description. For example, a succession in blooming phytoplankton species might extend or multiply the bloom periods in ways that are not "random" and that are difficult to reconcile with a single model functional group, even with stochastic parameters. Limited spatial resolution can also impose persistent biases that lead to poor extrapolation properties when we try to correct them by adjusting θ_e (Wallhead et al., 2013). In such cases, rather than elaborating the error models, effort might be better spent improving the explicit biological or spatial resolution of the model, or exploring implicit resolution techniques (e.g., Wirtz and Eckhardt, 1996; Merico et al., 2009; Wallhead et al., 2013).

An alternative approach might be to employ the tools of multimodel inference (Burnham and Anderson, 2002; Link and Barker, 2006). The idea here is to base inference of target parameters, states, and fluxes on a family of candidate models, each differing in structure and parameterisation, rather than on a single model. For example, we might be fairly certain about the form of the photosynthesis-irradiance (P-I) function in phytoplankton, but much less certain about the appropriate formulation of zooplankton grazing. Multimodel inference would allow the P-I parameter values and their uncertainties to be inferred on the basis of several candidate models, each assuming the same P-I function but different grazing parameterisations. The resulting

multimodel estimates and uncertainties would be less likely to be biased by a poor choice of grazing formulation than the inference premised on a single *a priori* formulation.

5 Posterior parameter uncertainties

The determination of parameter uncertainties has many facets, getting to the core of discussions of Bayesian and frequentist approaches and interpretations (e.g., Efron, 1986; Cox, 2005; Lele and Dennis, 2009). Depending on the estimator, uncertainties in the combination of parameter values may either disclose a credible region of a random distribution of parameter values (Bayesian interpretation) or they mark a confidence region that should include the true value with a certain nominal probability of e.g. 95% (frequentist interpretation). The latter means that different data sets would yield different confidence regions and e.g. 95% of those regions are expected to include the true "fixed" value.

In general, if we wish to make inference about uncertainties of parameter estimates ($\widehat{\Theta}$) we need some knowledge about the distributional shape of the posterior $p(\widehat{\Theta} \mid \boldsymbol{y})$ or of the likelihood $p(\boldsymbol{y} \mid \widehat{\Theta})$. Likewise, we can gather information about the parameter-cost function manifold in the vicinity of $(\widehat{\Theta}, J(\widehat{\Theta}))$. For this we may consider some threshold offset value Δ_J, which is an upper limit for the deviation from the minimum value $J(\widehat{\Theta})$. Such a limit may identify all cost function values that are insignificantly larger than $J(\widehat{\Theta})$. Large deviations from optimal estimates might be required for some parameters (components of $\widehat{\Theta}$) before the corresponding cost function values reach this threshold, while for other components only small variations are enough. Such tolerance limit defines an uncertainty region in parameter space:

$$\left\{\Theta : J(\Theta) - J(\widehat{\Theta}) \leq \Delta_J\right\} \tag{16}$$

Typical threshold values are defined as the α quantile of a parametric or nonparametric probability distribution.

For an unbiased ML estimator, the χ^2-distribution with the degree of freedom (df=$N_y - N_\Theta$) has been suggested for deriving a threshold value $\chi^2(\text{df}, \alpha)$ (e.g., Kuczera, 1990; Meeker and Escobar, 1995; Raue et al., 2009, 2011). But for nonlinear models the χ^2-distribution might be inappropriate and the α quantile of the actual distribution, $J(\Theta) - J(\widehat{\Theta})$, needs to be evaluated by other means (e.g., Raue et al., 2011). Furthermore, the degree of freedom (df) that specifies location and shape of the χ^2-distribution may not be representative. Only if error correlations have been correctly specified in J (see Sect. 4) and the asymptotic approximation (for large N_y) is applicable, then the correct degree of freedom is $N_y - N_\Theta$.The effective number of independent observations can be lower and the considered error correlations can be imprecise, for example when measurements like Chl*a* and carbon dioxide concentrations are negatively correlated during exponential growth but can then become positively correlated shortly after the peak of an algal bloom. We therefore expect the effective degree of freedom to be often lower than $(N_y - N_\Theta)$ and $\chi^2(\text{df}, \alpha)$ would therefore be an optimistic threshold, i.e. likely to underestimate the true range of uncertainty, unless the correct number of degrees of freedom is determined.

5.1 Confidence and credible regions

Uncertainty regions in parameter space can be determined basically in two different ways, either based on a Bayesian- or frequentist interpretations. ~~Depending on the estimator, uncertainties in the combination of parameter values may either disclose a credible region of a random distribution of parameter values (Bayesian interpretation) or they mark a confidence region that should include the true value with a certain nominal probability of e.g. 95% (frequentist interpretation). The latter means that different data sets would yield different confidence regions and e.g. 95% of those regions are expected to include the true "fixed" value.~~ According to the Bayesian interpretation a credible region is specified by conditional probability distribution of the true value given the data. For maximisations of the likelihood $p(\boldsymbol{y} \mid \Theta)$ it is often stated that credible and confidence regions are practically identical. Such interpretation is imprecise since the methods to confine either regions can be very different with respect to the underlying assumptions, e.g. MCMC versus bootstrap approaches.

In case of classical BEs no tolerance limit Δ_J is explicitly prescribed. Instead, an efficient sampling of $(\Theta, J(\Theta))$, or directly of the posterior $p(\Theta \mid \boldsymbol{y})$, is applied. Sequential methods can provide approximations of the posterior parameter distribution once all data have been assimilated. These approximations differ, depending on how Eqs. (6) and (7) are sampled and evaluated, as discussed in Sect. (2.2.2). A helpful overview with some comprehensible examples (of four different methods and three different ensemble sizes) is given by Weir et al. (2013). BE methods that do not rely on sequential approaches may also be applied and credible regions are then simply inferred from selective (acceptance/rejection) sampling schemes in a MCMC approach, e.g. Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970). MCMC methods for the derivation of credible regions are also used for ML estimation problems (e.g., Smith and Yamanaka, 2007a). The main point is that here the data are assumed fixed.

A fundamentally different approach to the ~~MCMC~~ BE methods is to repeat parameter optimisations many times but with data subsamples or resample data sets. Large data sets are split up into a series of subsamples that should be as independent as possible. Or many synthetic data sets are created by applying a random number generator to independently draw bootstrap samples (Efron, 1985; Efron and Tibshirani, 1986). This approach accounts for variable data and it mimics a repetition of an experiment or a repeated sampling at ocean sites. For each bootstrap data set $(\boldsymbol{y}^*)$ a corresponding optimum estimate $\widehat{\Theta}^*$ is obtained. A distribution of $\Delta_\Theta = \overline{\Theta^*} - \widehat{\Theta}^*$ can be derived from a series of optimisations with different bootstrap data sets. Furthermore, nonparametric density estimates of all $J(\widehat{\Theta}^*)$ can be derived and the α quantile can then be determined from the cumulative distribution of such probability density. For some situations a bootstrap approach with as few as ten resample data sets may suffice to highlight specific uncertainties in some model parameters (e.g., Schartau et al., 2007). But to ascertain confidence regions, much larger bootstrap sample sizes are typically needed (Efron and Tibshirani, 1986). In the end, both approaches, MCMC and bootstrap methods, require a large number of model evaluations, typically o(10^2) - o(10^4). The benefit is that skewed and contorted posteriors can be better resolved.

5.2 Profile likelihoods

An alternative to ensemble-based sequential, MCMC, and bootstrap methods for determining uncertainties of parameter estimates is the construction of 1D- or 2D profile likelihoods (Venzon and Moolgavkar, 1988). For a 2D profile likelihood an array of combinations of two parameters (Θ_m, Θ_n) is constructed. For every combination of parameter values (elements of the 2D array) a minimisation of $J(\Theta)$ is repeated while varying all other parameters $(\Theta_{l \neq m,n})$. This is done for all arrays with possible combinations of two parameters, which requires a large number of additional optimisations. The advantage is that uncertainty intervals $[\widehat{\Theta}_l - \boldsymbol{u}_l^-, \widehat{\Theta}_l + \boldsymbol{u}_l^+]$) can be well resolved for each component (l) of Θ, with lower and upper uncertainty limits possibly being different $(\boldsymbol{u}_l^- \neq \boldsymbol{u}_l^+)$. Unfortunately, the evaluation of a profile likelihood is impracticable for most marine ecosystem model applications, because of the associated computational costs. Parameter identifiability analyses based on profile likelihoods have been applied to problems where fast evaluations of $J(\Theta)$ were possible (e.g., Brun et al., 2001; Raue et al., 2009, 2011). Brun et al. (2001) evaluated confidence regions for three parameters (rate constants of production, respiration and water-air gas exchange) from profile likelihoods and they showed that the error margins of the parameter estimates can be much larger than those derived with e.g. a point-wise approximation of a posterior uncertainty covariance matrix, described in the following.

5.3 Point-wise approximations of posterior uncertainty covariance matrix

A single point in parameter space is identified by ML and MAP estimators, i.e $\widehat{\Theta}$ where the posterior $p(\Theta \mid \boldsymbol{y})$ has its maximum. Because of the computational costs we often find studies where parameter uncertainties of ecosystem models had been approximated point-wise in the immediate vicinity of $\widehat{\Theta}$. A common theory for deriving variance information of a ML estimate is based on the inverse of the Fisher information (Fisher, 1922; see also e.g., Fisher, 1934; Efron and Hinkley, 1978; Cao and Spall, 2010). The underlying assumption is that the likelihood $p(\boldsymbol{y} \mid \widehat{\Theta})$ is nearly normal shaped nearby its maximum, which is tantamount to a quadratic increase of $J(\Theta)$ as parameter values are varied around the estimate. Series expansions, like Taylor power series, around the estimate $\widehat{\Theta}$ can be applied to derive relevant properties of $J(\Theta)$ that are theoretically attributed to an uncertainty covariance matrix ($\mathbf{U}_\Theta$). Confidence regions for $\widehat{\Theta}$ can then be expressed in terms of approximations of $\mathbf{U}_\Theta$. For example, for some prescribed df an upper critical confidence level can be specified by the α quantile of a F-distribution (Marsili-Libelli et al., 2003):

$$\left\{\Theta : \left(\Theta - \widehat{\Theta}\right)^T \mathbf{U}_\Theta^{-1} \left(\Theta - \widehat{\Theta}\right) \leq N_\Theta \cdot F_{df}^{1-\alpha}\right\} \tag{17}$$

Confidence ellipsoids are described with Eq. (17), thus yielding symmetric uncertainty limits around $\widehat{\Theta}$, i.e. $\boldsymbol{u}_l = \boldsymbol{u}_l^- = \boldsymbol{u}_l^+$. With an approximation of $\mathbf{U}_\Theta$ a confidence interval for every single parameter can be described as $[\widehat{\Theta}_l \pm \boldsymbol{u}_l]$. The individual uncertainty limits can be computed as

$$\boldsymbol{u}_l = t_{\mathrm{df}}^{1-\alpha/2} \sqrt{\mathbf{U}_{\Theta_{ll}}}. \tag{18}$$

where $t_{\mathrm{df}}^{1-\alpha/2}$ is the two-tails Student's t-distribution for prescribed α and df (Marsili-Libelli et al., 2003). Two approaches to point-wise approximations of $\mathbf{U}_\Theta$ are found in ecological and ecosystem modelling studies. One approach uses first derivates

of the model's observation vector with respect to the parameters (Jacobian) whereas the other requires calculations of second derivatives of $J(\Theta)$ (Hessian).

5.3.1 Uncertainty covariances based on the Jacobian matrix

A first approach considers a linearisation (first order power expansion) of the model's observation vector $H(\boldsymbol{x})$ around the point estimate $\widehat{\Theta}$. As long as $H\left(\boldsymbol{x}(\widehat{\Theta})\right)$ is not subject to strong nonlinearities, its first derivatives (sensitivity) with respect to Θ can be used to estimate $\mathbf{U}_\Theta$. For an unbiased ML estimator the covariance matrix can be approximated as:

$$\mathbf{U}_\Theta = \frac{J(\Theta)}{\mathrm{df}} \cdot \left(\mathbf{H}_\Theta^T \mathbf{R}^{-1} \mathbf{H}_\Theta\right)^{-1} \tag{19}$$

with the Jacobian matrix $\mathbf{H}_\Theta(\widehat{\Theta})$, its transpose ($\mathbf{H}_\Theta^T$), and with the observational error covariance matrix $\mathbf{R}$ (e.g., Thacker, 1989; Kuczera, 1990; Omlin and Reichert, 1999; Brun et al., 2001; Omlin et al., 2001). The term $J(\Theta)$/df is added as an approximation of the residual variance of J, which should be considered unless $H(\boldsymbol{x})$ is in such good agreement with data so that the minimum of $J(\Theta)$ actually matches the exact degree of freedom, df. The rows of the Jacobian $\mathbf{H}_\Theta$ are the first derivatives with respect to the parameters $\nabla H(\boldsymbol{x})$, with $\nabla = (\partial/\partial\Theta_1, \partial/\partial\Theta_2, \ldots, \partial/\partial\Theta_{N_\Theta})$ being the Napla operator of first partial derivatives.

5.3.2 Uncertainty covariances based on the Hessian matrix

Another more common approach for a point-wise approximation of $\mathbf{U}_\Theta$ is derived from a Taylor expansion around $J(\widehat{\Theta})$. Since $\nabla J(\widehat{\Theta}) \approx 0$ in the minimum, the first order term of the Taylor expansion is negligible. The series expansion then approximates the distribution:

$$J(\Theta) - J(\widehat{\Theta}) \approx \frac{1}{2}\left(\Theta - \widehat{\Theta}\right)^T \boldsymbol{\mathcal{H}}_\Theta \left(\Theta - \widehat{\Theta}\right) \tag{20}$$

The matrix $\boldsymbol{\mathcal{H}}_\Theta$ is the Hessian whose elements are second derivatives of $J(\Theta)$ with respect to the parameters (e.g., Tziperman and Thacker, 1989; Matear, 1995):

$$\boldsymbol{\mathcal{H}}_\Theta = \left.\nabla^T \nabla J(\Theta)\right|_{\Theta=\widehat{\Theta}} \tag{21}$$

With the Taylor expansion in Eq. (20) we obtain an approximation of the local curvature of $J(\Theta)$ at point $\widehat{\Theta}$, also referred to as the *observed* Fisher information. Like in Eq. (19), but instead of using first derivatives of $H(\boldsymbol{x})$, a posterior uncertainty covariance of $\widehat{\Theta}$ is then approximated by computing the inverse of a Hessian matrix:

$$\mathbf{U}_\Theta = \frac{J(\Theta)}{\mathrm{df}} \cdot 2 \cdot \boldsymbol{\mathcal{H}}_\Theta^{-1} \tag{22}$$

Both approximations (Eqs. 19 and 22) yield, in principle, similar results for accurate ML estimates i.e. when the actual minimum of $J(\Theta)$ has been identified by the optimisation algorithm. In practice search algorithms can terminate at some distance from the actual minimum for numerical reasons, e.g. when the minimum is located in a flat valley of J and the imposed convergence criterion makes an algorithm terminate the search in the periphery of the valley. Marsili-Libelli et al. (2003) proposed

an approach where the accuracy of parameter estimates can be improved by minimising differences between the results of Eq. (19) and Eq. (22).

5.3.3 The Hessian: its approximation and inversion

Hessian matrices have often been approximated with a finite central differences approach for first and second derivatives of J with respect to ecosystem model parameters at the point-estimate $\widehat{\Theta}$ (e.g., Matear, 1995; Kidston et al., 2011; Kreus and Schartau, 2015). A critical issue of finite difference calculations of the Hessian's elements is the choice of an appropriate increment size ($\boldsymbol{\delta}$), which sets the distance of departure from the optimal parameter point estimate $\widehat{\Theta}$. Sometimes a compromise between resolving flat regions around $(\widehat{\Theta}+\boldsymbol{\delta}, J(\widehat{\Theta}+\boldsymbol{\delta}))$ and numerical precision has to be found (Kreus and Schartau, 2015). To approach a high accuracy of the Hessian approximation it is possible to consider a set of different increment sizes for the central differences approach, as given in Marsili-Libelli et al. (2003).

The problem of increment size reduces if first derivatives of J with respect to the parameters (gradient, ∇J) are readily obtained with an adjoint model, e.g. as used in a variational DA approach (Sect. 2.2.3). Adjoint versions of plankton ecosystem models have been constructed primarily to compute ∇J for an efficient search with gradient descent algorithms in the parameter-cost function manifold (e.g., Lawson et al., 1996; Fennel et al., 2001; Schartau et al., 2001; Spitz et al., 2001; Friedrichs, 2002; Faugeras et al., 2003; Zhao et al., 2005; Friedrichs et al., 2007; Xiao and Friedrichs, 2014a). To elucidate the nature of adjoint model developments is beyond the scope of this paper, but a brief summary about adjoint model developments is given in the Appendix (C). The advantage is that all elements of the Hessian can be approximated with finite differences of adjoint model results (e.g., Fennel et al., 2001; Friedrichs, 2002; Faugeras et al., 2003; Friedrichs et al., 2007; Kreus and Schartau, 2015).

Computations of the Hessian, Eq. (21), provide valuable identifiability information even if this matrix is not explicitly used to specify confidence regions of parameter estimates. For example, a decomposition of the Hessian matrix into its eigenvalues and the corresponding eigenvectors reveals which parameters are weakly constrained by the data or it helps to identify structural deficiencies of a model. The eigenvectors' components (l) represent the components of Θ. Components of those eigenvectors that belong to small eigenvalues indicate parameter combinations that are poorly constrained or cannot be estimated. In contrast, those eigenvectors that correspond with the largest eigenvalues show parameter combinations that are well constrained. The studies of Fennel et al. (2001) and Faugeras et al. (2003) are informative in this respect, because they provide insight into the range of characteristic eigenvalues and eigenvectors of 0D and 1D marine ecosystem models.

Ideally, every eigenvector would exhibit only one single component, meaning that values of every parameter can be estimated independently of the other parameters' values. In practice this is only the case for few parameters of a planktonic ecosystem model. Eigenvectors with two or more distinct components disclose those parameters whose estimated values are correlated and for which correlation coefficients can be explicitly derived (e.g., Matear, 1995; Prunet et al., 1996). Correlations between parameter estimates are referred to as collinearities. A useful collinearity index was introduced by Brun et al. (2001). Their index expresses how a change in J (or in $H(\boldsymbol{x})$), due to a shift in the value of one parameter can be entirely compensated by adjusting the value of another (correlated) parameter.

5.4 Parameter collinearities: an example with phytoplankton loss parameters

In Sect. (3.6) we discussed the difficulty of constraining parameters that determine loss rates of phytoplankton biomass due to grazing, aggregation or exudation and leakage or organic matter. With an example we illustrate typical uncertainties and collinearities in the estimation of phytoplankton loss parameters in the absence of explicit zooplankton observations like micro- and mesozooplankton abundance or grazing rates. Three parameters that affect the loss of phytoplankton biomass have been optimised together with other parameters. For this we assimilated five different types of daily mean observations of a mesocosm study (Engel et al., 2005; Delille et al., 2005) into a plankton ecosystem model with optimal nutrient allocation and photoacclimation (Pahlow, 2005), as mentioned in Sect. (3.5).

Details of the cost functions and the corresponding mapping from model results $\boldsymbol{x}$ to observations $H(\boldsymbol{x})$ are given in the Appendix (B). In our example we consider two cost functions, with and without covariances respectively (Eqs. B4 and B5). For both cost functions no prior information is included. As an error model we assume additive Gaussian errors, applying Eq. (4) in Sect. (2.1.3). A simulated annealing (SA) algorithm is first used to identify a best parameter estimate in the vicinity of the global cost function minimum. This point estimate is then used to derive error ellipses (confidence regions) according Eq. (22). These point-wise approximations of parameter uncertainties are finally incorporated to initialise the MCMC method that derives a credible region of posterior parameter uncertainties, based on an algorithm provided by Soetaert and Petzoldt (2010).

Figure (2) shows contours of $J(\widehat{\Theta}_m \pm \Delta_m, \widehat{\Theta}_n \pm \Delta_n; m,n = 1,2,3)$ around the optimum at $(\widehat{\Theta}_m, \widehat{\Theta}_n, \min(J))$, while all other parameters are fixed to their optimal estimates $(\widehat{\Theta}_{l \neq m,n})$. Each plot is thus a combination of two loss parameters: maximum grazing ($\Theta_1 = g_m$) and carbon loss rate ($\Theta_2 = \gamma_C$) on top (1a/b in Fig. 2); γ_C and aggregation parameter ($\Theta_3 = \Phi_{\text{agg}}$) in the middle (2a/b); Φ_{agg} and g_m on the bottom (3a/b). Results from MCMC (dots and asterisks) reveal similar collinearities between parameter combinations that involve g_m for the two cost functions (1a/b and 3a/b in Fig. (2). It means that g_m can only be estimated in combination with Φ_{agg} and γ_C. Only if Φ_{agg} and γ_C were known, then g_m could be identified in this mesocosm model setup with these available data types. We do not find such strong collinearity expressed between γ_C and Φ_{agg} and their estimates seem to be rather independent (2a/b of Fig. 2), given the mesocosm data.

Another peculiarity is that the ranges of the MCMC's posterior indicate larger uncertainties if the cost function without covariance information is applied (right side of Fig. 2), although model and data are identical. This behaviour is also resolved by the 95% confidence regions that are obtained with a point-wise approximation of error ellipses (lines). Furthermore, collinearities according to the error ellipses are smaller for the cost function with covariances compared to the case of independent data. Here, confidence regions of the error ellipses correspond well with the credible regions of the MCMC results. We stress that this may not be the general case and the good correspondence is likely attributable to the low dimension of the example looked at.

Overall, these results exemplify the uncertainty in constraining major loss parameters in the presence of grazing, if no explicit prior information about grazing rates or data of zooplankton biomass are available. Collinearities between grazing parameters and other phytoplankton biomass losses may be reduced by testing model performance against independent data, e.g. as done for the meso- and microzooplankton grazing in Buitenhuis et al. (2010). In cross-validation studies some combinations of

parameters that produce indistinguishable solutions for one experiment or for one ocean site are compared with data of another experiment or at another ocean site, which will be addressed in the following Sect. (6).

6 Cross-validation and model complexity

Good performance should be attributable to a model capturing the predominant plankton dynamics under varying conditions in different environments. Parameter values are often optimised for local ocean sites, but ideally, parameter estimates from one site should improve model performance at other locations as well. The generality of optimised models can be tested by cross-validating against independent data, providing a direct and effective test of predictive skill (Gregg et al., 2009).

6.1 Cross-validation

Parameter optimisations can often improve the fit of a model by selecting unrepresentative parameter values that serve only to compensate for misfits between data and model results. It is therefore essential to check whether the resultant 'optimised' model is giving the right answer for the correct reasons.

Xiao and Friedrichs (2014b), for example, found that while the optimisation of a range of NPZD models to satellite data tended to reduce model-data misfit, this was often achieved through the adoption of extremely unrealistic parameter estimates, sometimes being multiple orders of magnitude higher or lower than their best a priori estimates. The same authors (Xiao and Friedrichs, 2014a) showed that adding synthetic noise to assimilated satellite data led to the introduction of similar errors, and a significant deterioration of one model's predictive skill. The extreme parameter estimates were not representative for the system and the model performance turned out to be poor when the model was tested against independent data that were not used during the optimisation procedure.

This is the principle of cross-validation, in which an optimised model is tested in terms of its ability to reproduce data that were not included in the calibration phase. This is often achieved by excluding a subset of the original calibration dataset, for later use in model evaluation. For example, in a variational data assimilation exercise for the Arabian Sea, Friedrichs et al. (2006) repeated their optimisation a number of times, each time excluding data from a particular season. The calibrated models were then used to predict the system behaviour during the withheld season, with the resultant model-data misfit labelled the 'predictive cost function'.

The cross-validation approach has the advantage of testing one of the key attributes of marine biogeochemical models, namely their predictive skill. The technique is, however, not without its difficulties. The first issue is that it is important to ensure the test data are truly independent from the training data. In this regard, Friedrichs et al. (2006) took advantage of the highly seasonal nature of the Arabian Sea, but it would perhaps be less appropriate in regions with a less pronounced seasonal cycle, such as at the centre of a subtropical gyre. A potentially more serious problem occurs when researchers simply divide the available data at random, such that highly correlated data appear in the assimilated and the test data. Under such circumstances, the cross-validation would give no indication as to the ability of the model to predict independent data.

The potential to select unrealistic, compensatory, parameter values may not always be obvious, especially if good estimates of the 'true' (or at least sensible) values of the model parameters are not well known a priori. Such errors may, nonetheless, strongly impact the ability of a model to reproduce anything but the assimilated data. This issue appears to be a common theme in simple marine biogeochemical models calibrated to time-series data, as a number of studies (Fennel et al., 2001; Friedrichs et al., 2006; Ward et al., 2010) have found that parameter optimisation resulted in decreased predictive skill, relative to 'off-the-peg', prior parameterisations. A notable counterpoint to those studies is given by Oschlies and Schartau (2005), who found that simultaneous optimisation of an NPZD model at three time-series sites (Schartau and Oschlies, 2003) led to improved performance when the model was applied within a 3D simulation of the North Atlantic. On the one hand, it seems likely that this improvement was dependent on assimilating data from three highly dissimilar North Atlantic locations, which prevented the inclusion of compensatory errors that were highly specific to any one site (see also Xiao and Friedrichs, 2014a). On the other hand, in Schartau and Oschlies (2003) and in Oschlies and Schartau (2005) it is also stressed that the apparent improvement is associated with some ambiguous rapid nitrogen remineralisation pathway in their simple NPZD model, which can be incorrect in either simulations (1D and 3D), but with the same positive effect on primary production rates in the central North Atlantic.

6.2 Model performance as a function of model complexity

Of the many factors that affect the ability of a biogeochemical model to reproduce and predict observations, the appropriate degree of model complexity in any given situation is both one of the most important, and one of the least well defined. This is because there exists a fundamental trade-off between simplicity and complexity. Simple models have the advantage of being easier to understand, and with fewer parameters they should also be better constrained (both before and after optimisation). Nonetheless, simplification requires a degree of abstraction, and it can sometimes be difficult to draw parallels with the complexities of the observed system.

At the other end of the spectrum, a highly complex model can explicitly resolve more processes, allowing more detailed comparison with observations. As models become more complex, the number of degrees of freedom increases, and the calibrated model will generally be able to match the observations better than a simpler model. If insufficient observations are available, the extra degrees of freedom can lead to the introduction of compensatory errors at the assimilation site, which could then increase uncertainty at other locations, as illustrated by Xiao and Friedrichs (2014b). Similarly, for small changes in the assimilated data an extra flexibility may lead to very different model solutions, also leading to increased uncertainty in model predictions (e.g. Xiao and Friedrichs, 2014a) .

A range of statistical techniques are available to assess this trade off, and a useful review is given by Johnson and Omland (2004). One of the most practical (if not the most general) techniques is cross-validation, as described in the previous section (see also Hastie et al., 2009, section 7.10 for an excellent discussion in a general statistical context). ~~In their experiments~~ By looking at the effects of adding noise to assimilated remote sensing data, Xiao and Friedrichs (2014a) found that the most complex model they evaluated was also the most sensitive to the introduction of synthetic errors in the assimilated data (Fig. 3). They attributed this result to the extra degrees of freedom that could be 'fit to noise'. This is consistent with earlier findings

that model predictive skill deteriorates as complex models can become "overfit" to the data (i.e. too many parameters are fit to inadequate data) (Friedrichs et al., 2006, 2007; Ward et al., 2010).

Aside from directly assessing a model's predictive skill using cross-validation, a number of alternative approaches are available to identify the minimum number of model parameters that are supported by the available data. One of the simplest techniques (in terms of its applicability), is the Akaike Information Criterion (AIC, Akaike 1973). The AIC considers two opposing terms corresponding to the maximum log-likelihood of the parameters given the data ($\ln[L(\widehat{\Theta} \mid \boldsymbol{y})]$, measuring model data misfit) and a bias-correction factor, that increases with the number of free parameters (N_Θ).

$$\text{AIC} = -2 \ln\left[L\left(\widehat{\Theta}_p \mid \boldsymbol{y}\right)\right] + 2N_\Theta \tag{23}$$

Note that for a model fitted by least-squares, the log-likelihood can be approximated by the residual sum of squares (RSS), following Johnson and Omland (2004): $\ln[L(\widehat{\Theta}_p \mid \boldsymbol{y})] \approx -N_y/2 \cdot \ln(\text{RSS}/N_y)$, with N_y being the total number of observations. The AIC, and alternative techniques (weighted AIC, or Bayesian Information Criterion, BIC), seek to quantify the trade-off between bias and variance (e.g., Burnham and Anderson, 2004). Of a range of competing models, the one with the lowest AIC has the greatest empirical support.

A perhaps more intuitive approach is given by the Likelihood Ratio Test (LRT) for e.g. comparing so-called nested models, in which the simpler model is a special case of the more complex model, in the sense that $M_p = f_1$ is a special case of $M_{p+1} = f_1 + f_2$ where f_2=0. Like the AIC, the LRT aims to account for model complexity in the sense that it compares log-likelihoods:

$$\text{LRT} = J(\widehat{\Theta}_p) - J(\widehat{\Theta}_{p+q}) \tag{24}$$

with $J(\widehat{\Theta}) = -2 \ln[L(\widehat{\Theta} \mid \boldsymbol{y})]$ and index $p+q$ indicating the number of free parameters of the full model. An alternative simpler model (with p parameters) that is not significantly worse than the full model (with $p+q$ parameters) can be selected using this ratio. There is a clear analogy to Eq. (16) in Sect. (5). In other words, although having removed individual parameters (going from Θ_{p+q} to Θ_p) we may still have an increase in the data-model misfit that is tolerable or insignificant within some limit Δ_J. For nested models only, a value for Δ_J can be derived from a $\chi^2(\text{df} = q, \alpha)$ distribution. The respective degree of freedom (df) is then assumed to be equal to the difference in the number of free parameters between the full and the reduced model, which is q. For LRT with non-nested models an empirical, non-parametric distribution needs to be derived by other means instead, for instance using synthetic (or resample) data sets (e.g., Lewis et al., 2011).

The theory mentioned above is well described by Johnson and Omland (2004), and have already been applied in few ecosystem modelling studies (e.g., Crout et al., 2009; McDonald and Urban, 2010; Ward et al., 2013). The techniques for model selection have generally shown that more complex models are more vulnerable to over-tuning than simpler models. This appears to be because the number of uniquely identifiable parameters in marine biogeochemical models is often very low. Studies based on classic NPZD type models have typically found that the inclusion of as few as three to 15 parameters was supported by the assimilated data (Matear, 1995; Friedrichs et al., 2007; Ward et al., 2013; Löptien and Dietze, 2015). It should however be noted that these studies made use of only very limited datasets, and a higher level of complexity would likely be supported with the incorporation of more comprehensive datasets, especially those describing fluxes.

Ward et al. (2013) sequentially removed parameters from a relatively simple 2NPZD model to show that much of the model structure was redundant, with respect to the assimilated data, Fig. (4). They applied an F-score where the relative change in LRT is related to the relative change in parsimony (i.e. difference in the number of free parameters between the reduced and the full model divided by the degrees of freedom of the full model, df_{p+q}=N_y - $N_{\Theta_{p+q}}$):

$$\mathrm{F} = \left[\frac{\mathrm{LRT}}{J(\widehat{\Theta}_{p+q})}\right] \cdot \left[\frac{N_{\Theta_{p+q}} - N_{\Theta_p}}{\mathrm{df}_{p+q}}\right]^{-1} \tag{25}$$

As model complexity was reduced, model predictive skill was initially very slow to deteriorate, and J remained similarly low. The increased parsimony of the simpler models led to improved performance in terms of the LRT, and the AIC and Bayesian information criterion (BIC). Once all of the redundant components of the model were removed, removal of essential components led to a rapid increase in J, with an associated increase in the other metrics. The LRT selects the simplest model with an F-score below a variable threshold value. ~~The AIC and BIC select the model with the lowest overall score. Among models with a similar score, the simplest should be favoured.~~ The AIC and BIC can be used to select a single model with the lowest score, or preferably to provide individual model weightings for multimodel inference (Burnham and Anderson, 2002), although it appears that this latter has so far seen little application to planktonic ecosystem models.

7 Space-time variations in model parameters

Theoretical arguments, as well as results from cross-validations, have revealed problems with the portability of locally calibrated models (e.g., Hurtt and Armstrong, 1999; Friedrichs et al., 2007) and raise the question of how representative local estimates are if applied at larger scales. These limitations encourage estimators that allow spatial and/or temporal variations of parameter values.

For spatial or temporal variation to be useful we have to make sure that the corresponding parameter adjustments reflect changes in the actual underlying (real-world) dynamics. To assess whether this condition is met is a particularly challenging problem that has yet to be adequately addressed. Direct comparisons are needed between optimisations that allow variation in posterior parameter vectors and those that do not. In studies where direct comparisons are made, a common finding is a reduction in the model misfit to the assimilated data by allowing these kinds of variations, but this tells us little. A reduction of the cost function is expected, as a direct consequence of an effective increase in the number of adjustable parameters. As pointed out by Gregg et al. (2009), "skill assessment using assimilated data lacks the independence necessary for a comprehensive, objective evaluation". Studies where cross-validation is performed to test predictive skill are more informative. Switching between different parameter sets in time or for specific regions may not necessarily be a solution *per se* but may indicate where model refinements have to be investigated (Huret et al., 2007). From analyses of spatially- and temporally varying parameter estimates that improve predictive skill we can learn where and when particular model equations are limited in reproducing changes in plankton dynamics with fixed parameter values. Such analyses should provide important feedback information on revising these parameterisations.

7.1 Regional differences between parameter estimates

Satellite ocean colour data are widely used to investigate spatial differences in parameter estimates. In many cases, a local calibration method is applied where parameters are optimised separately to fit Chl*a* data for a number of pre-defined sites or regions spanning a domain of interest. For example, parameters of a 3D-NPZ model were optimised by Garcia-Gorriz et al. (2003) for January and June for two regions, the North- and South Adriatic basin in the Mediterranean Sea. They inferred comparable parameter vectors for the two regions during bloom conditions in January but considerable differences between the regionally optimised parameter sets emerged for June. Garcia-Gorriz et al. (2003) attributed this difference to unresolved variations in plankton composition and changes in biomass concentration between the two basins. Huret et al. (2007) performed a similar assimilation experiment for the Loire and Gironde river plumes in the Bay of Biscay. On the one hand, they found some similarities between parameter estimates for the two distinct river plumes for particular conditions during spring, suggesting the possibility of a common set of parameter values for both plume areas. On the other hand, the authors stressed their optimal parameter estimates to be based on data for a specific period and obtained excessively high Chl*a* concentrations in the Bay of Biscay for the entire simulation year when utilising the mean of parameter estimates for the two plume regions.

Pronounced regional and seasonal differences are not restricted to adjacent seas and coastal areas. Large scale studies for the North Atlantic have shown comparably strong regional differences between parameter estimates (Hemmings et al., 2003; Losa et al., 2004; Doron et al., 2013; Kuhn et al., 2015). A set of sites representing distinct latitude bands was considered for a one year calibration of a NPZ and a NPZD model in Hemmings et al. (2003). The annual cycle at locations on a five degree grid was simulated with variable parameter estimates of a NPZD model in Losa et al. (2004) and individual parameter estimates for thirteen provinces in the North Atlantic, pre-defined according to Longhurst (1995), were derived for a six-compartment 3D biogeochemical model in Doron et al. (2013). Kuhn et al. (2015) estimated NPZD model parameters for six 5×10 degree regions of the central North Atlantic. Despite the fact that these studies used different models, it is possible to compare some optimised parameters that are equivalent or closely related between all studies. However, little obvious consistency is seen in the spatial patterns between their estimates, although Doron et al. (2013) suggested some similarity between their estimates of phytoplankton maximum growth rate and zooplankton maximum grazing rate with those of Losa et al. (2004). Patterns of spatial variation in parameters are not easily validated as most parameters do not have well-observed equivalents in nature. Nevertheless, Losa et al. (2004) were able to document the plausibility of their posterior photosynthesis parameter values for the maximum phytoplankton growth rate (μ_m in Sect. 3.1) and intial slope of the P-I curve (α_{phot} in Sect. 3.3) by comparison with observational estimates of Platt et al. (1991). Six parameters were optimised in all and the posterior parameter fields were cross-validated in a 3D version of their model by comparing the output with an independent SeaWiFS chlorophyll data from 1997-2003 (Losa et al., 2006). The spatially-varying parameter set of Losa et al. (2004), obtained by assimilating Coastal Zone Color Scanner (CZCS) data for the period 1979-1985, was interpolated and extrapolated onto the spatial grid of the 3D model as shown for the two parameters relevant for phytoplankton growth, μ_m and α_{phot} respectively (Fig. 5). This enabled the model to simulate the seasonal patterns in SeaWiFS data much better than with a fixed prior parameter vector. An important caveat is that the calibration and validation data sets are essentially two realisations of the same emerging spatio-temporal patterns. To

demonstrate improved predictive skill attributable to its dynamics the model would be expected to resolve differences between the two independent data sets, given physical forcing data specific to each period.

7.2 Combining sites or regions

The presence of parameter variation between sites or regions for which a model was calibrated independently does not refute the existence of a common parameter vector with which the model could achieve similar results. Garcia-Gorriz et al. (2003) and Hemmings et al. (2003) performed alternative experiments in which regions were combined under a uniform parameter vector constraint, but did not include predictive skill tests for direct comparisons of the performance of spatially-varying and uniform parameter solutions. In other studies, sites have been combined without considering the alternative of allowing parameters to vary spatially. By optimising a 13-parameter model for locations of the Ocean Wheather Ship India (OWSI) and of the Bermuda Atlantic Time-series Study (BATS) simultaneously Hurtt and Armstrong (1999) found that it could capture the primary observed characteristics of the annual cycle at both sites, despite being unable to reproduce the cycle at BATS when calibrated at OWSI. As mentioned in the previous section, the approach of data assimilation over multiple sites has since been used by Schartau and Oschlies (2003) with some success in improving predictive skill of a 3D North Atlantic simulation (Oschlies and Schartau, 2005) based on a simultaneous three-sites calibration. A relatively complex global model with 45 adjustable parameters was similarly demonstrated to improve the predictive skill after assimilating time series data at five different calibration sites (Kane et al., 2011).

There is a clear advantage of combining sites or regions, in that it makes more data available to constrain parameters. It also creates a representative sample for the domain of interest, reducing the risk of over-fitting. In contrast, when assimilating data at a single site, Friedrichs et al. (2007) found it necessary to limit the number of adjustable parameters (to four or even less) to avoid portability problems. Use of a larger data set representing a wider diversity of ecosystem behaviour should support a greater number of parameters to be constrained, which would allow a model's true flexibility to be more fully exploited. However, there is a potential disadvantage of combining sites or regions, particularly over large spatial scales, in that the resultant parameter vectors may be less suitable for either region than parameter vectors obtained by local calibration.

Hemmings et al. (2004) introduced the idea of allowing provinces that are in a sense optimal for calibration to emerge during the data assimilation process. A sample of sites from the domain of interest is divided into two similarly distributed sets, one for calibration and the other for cross-validation. The objective is to find "the number and geographic scope of parameter vectors which allow the lowest possible cost of the calibrated model, with respect to the stations in the validation set, to be obtained". The method involves first performing a *whole-domain calibration* where parameters are optimised for all calibration sites, then recursively splitting the domain into two geographic provinces to investigate whether a better calibration can be achieved by optimising parameters for each one separately, a procedure referred to as *split-domain calibration*. The relative merits of ~~alternative~~ the calibration procedures are assessed by cross-validating the posterior parameter vector or vectors against sites from the validation set.

Application of the method to the North Atlantic data set used by Hemmings et al. (2003), with the same NPZ model and twelve adjustable parameters, resulted in the discovery of a two-parameter vector solution having a cross-validation misfit cost 25% lower than that for the single vector solution obtained for all calibration sites. The two sub-domains are shown in Fig. (6). The validation cost was also 24% lower than that obtained when the model was calibrated locally using individual sites. This is consistent with subsequent findings of Xiao and Friedrichs (2014b), where combining sites tends to reduce validation costs. Note that the validation scheme used by Hemmings et al. (2004) may not be able to discriminate well between skill associated with the model dynamics and that associated with the ability of the model to interpolate spatio-temporal patterns between the calibration sites shown in Fig. (6). This could be resolved by comparison with interpolated output from some purely empirical model fitted to the calibration data.

7.3 Spatially varying parameter estimates derived with Bayesian hierarchical modelling

Zhang and Arhonditsis (2009) proposed a Bayesian hierarchical formulation for calibrating aquatic biogeochemical models at multiple sites. In this framework, posterior parameter distributions can vary between sites but the sites share common prior distributions. Fiechter et al. (2013) used this approach to estimate parameter distributions for a 1D NPZD-iron model at two sites in the Gulf of Alaska. Non-informative prior distributions were employed for each parameter so the influence of the priors on the solution for each site was fairly weak. In a parallel Bayes' hierarchical modelling study for the same model, Leeds et al. (2013) assimilated satellite chlorophyll data at nine sites using a spatial Gaussian process model for the parameters with an anisotropic correlation matrix to allow for differences between along-shelf and cross-shelf dependence. The methods employed by Leeds et al. (2013) and Fiechter et al. (2013) seem promising because of their potential for rigorous treatment of uncertainty. However, in the absence of cross-validation experiments, their potential for improving the predictive skill of the models is not well evaluated at present.

7.4 Time-varying parameters

The idea of representing seasonal variation in part by temporal variations in the parameters has been examined in various studies (Losa et al., 2003; Brasseur et al., 2005; Dowd, 2006; Roy et al., 2012; Mattern et al., 2012, 2013a, 2014; El Jarbi et al., 2013; Melbourne-Thomas et al., 2013). In some cases, parameters are allowed to vary in space and in time (Tjiputra et al., 2007; Fan and Lv, 2009; Doron et al., 2013; Li et al., 2013). Cross-validation tests comparing the merits of varying and non-varying parameter solutions are mostly lacking, which prevents inferences being drawn about the superiority of these parameter variations for improving predictive skill. Temporal variation is handled naturally by adapting widely used sequential state estimation techniques to obtain parameter values along with state estimates.

Losa et al. (2003) applied a SIR particle filter to a model with 15 time-varying parameters in an assimilation of multi-year time series at the BATS site. The model was treated as a weak constraint with an additive system noise term that was uncorrelated between state variables. Mattern et al. (2013a) instead added noise to their two parameters in a 7-compartment 3D biogeochemical model of the Middle Atlantic Bight, with the advantage that the state evolution over each forecast step was true

to the model and correlated errors between state variables were represented. In both cases, the error model is highly subjective, yet it can have a major impact on the results. For instance, Losa et al. (2003) found the level of noise to be a critical factor affecting their solution. This motivated subsequent experiments in which additional time-varying parameters representing the noise level for each state variable were optimised (Brasseur et al., 2005). The posterior parameter trajectories thus obtained were not consistent with the earlier results. Despite the subjective characteristics of the system noise, the solution of Losa et al. (2003) improved the model prediction of unassimilated bacteria data. The necessity of time-variation in the parameters for achieving this is unclear, since no alternative results for static parameter solutions were analysed.

In a more recent BATS assimilation study with a simpler NPZD model, El Jarbi et al. (2013) did compare the performance of time-varying and static parameter solutions. Rather than employing a sequential method, they opted to solve the optimal control problem, i.e. to find parameter trajectories that minimise a cost function for the complete time period. An annual periodicity constraint on posterior parameter trajectories was introduced to allow the calibrated model to be also applied for time periods beyond the range of observations. Optimal periodic parameters were obtained using a two-year data set and validated against independent data for the following three-year period. In cross-validation tests, this solution was shown to improve predictive skill over the static parameter solution of Rückelt et al. (2010). Their results suggest that the time-varying parameter model may capture some aspects of the inter-annual variability, which would indicate dynamical skill.

Mattern et al. (2014) compared the predictive skill of versions of their two-parameter model with time-varying and static parameter solutions. Here, the time-varying solution was obtained using an alternative, emulator-assisted sequential data assimilation scheme. Their cross-validation experiments show a modest improvement in the ability to predict the annual cycle with time-varying parameters. Ability to predict the inter-annual variability was not tested and the achievability of similar predictive skill by purely empirical representations of the annual cycle derived from the observational data is not ruled out.

An experiment allowing both time and space variation in biogeochemical parameters that includes cross-validation is presented by Simon et al. (2015). Performance is compared against that of a model with constant spatially uniform parameters specified a priori but not against static and/or uniform parameter solutions to the data assimilation problem. The study employed an Ensemble Kalman Filter approach for combined parameter and state estimation in a coupled model of the North Atlantic and Arctic Oceans. Estimates for 4 model parameters that varied spatially and seasonally over the domain were obtained by assimilating satellite chlorophyll data for 2008 and 2009 and applied to the estimation of chlorophyll in 2010. A slight improvement was seen in 2010 chlorophyll relative to that for the prior parameter simulation. This suggests a small improvement in predictive skill, perhaps attributable in part to a better representation of persistent patterns in the annual cycle. A comparison of the assimilating run against independent nutrient data at Station 'M' was generally inconclusive with regard to the potential of the final parameter estimates to improve predictive skill for the nutrient fields

7.5 Learning from space and time variation in parameter estimates

~~A variety of approaches have been explored for DA with parameters varying in space or time or both.~~ As shown in this section, a variety of approaches have been explored for DA with parameters varying in space or time or both. We conclude the

section by considering what might be learnt from these types of studies. A common finding is that the posterior misfit cost with respect to the assimilated data is reduced by allowing variation, but this provides no evidence in itself to support the case for parameter variation. Allowing parameter variation increases the number of parameter values to be optimised, making it easier to fit a given data set.

Goodness-of-fit statistics that penalise model complexity in terms of number of parameters (e.g. the F-score of Ward et al., 2013, described in Sect. 6.2) could prove more informative, but are not used. Cross-validation can be used to provide a direct demonstration of differences in predictive skill. In the few studies which do use cross-validation to compare uniform and varying parameter solutions (Hemmings et al., 2004; Mattern et al., 2014; El Jarbi et al., 2013), some evidence of predictive skill is seen but the cross-validation schemes are not shown to discriminate reliably between predictive skill associated with model dynamics and that due to interpolation of patterns in space or persistence of an annual cycle. Better cross-validation schemes will be needed before we can convincingly demonstrate real improvements in the models as a result of introducing spatial and/or temporal variation in parameters.

Allowing parameters to vary reduces the extent to which their values can be constrained by a given set of observations, making an already under-determined problem worse. It could therefore be argued that parameter variation is justified only when there is good evidence to infer that a given model cannot adequately represent the observed variability under the uniform parameter vector constraint. The evidence should be statistically robust, taking into account all relevant sources of uncertainty. The consideration of these additional uncertainties, motivated by its potential for improving parameter estimates (Hemmings and Challenor, 2012), may tend to weaken data constraints further and make the introduction of parameter variation less practical, as well as affecting the strength of the evidence in support of it.

Heterogeneity in the parameter vector is most likely to be useful for structurally simple models. Those models may lack the required flexibility to capture some distinct spatial features observed within large domains or they may fail to resolve specific events during a complete annual cycle. Its introduction may be a sensible alternative to increasing structural complexity as it does not increase the computational demands of 3D simulations. From an ecological point of view, the need to introduce space and time variations in parameter values reflects limitations in resolving physical environmental changes, or deficiencies in physiological or ecological processes, or all of these factors together. For example, variations in plankton elemental stoichiometry, e.g. variable Chl*a*:C and C:N ratios, induce variations in photosynthetic rates that may not be well described by a model's parameterisation of Chl*a* synthesis and assimilation of nutrients (as discussed in Sect. 3.2). It is helpful to consider biological or environmental reasons why space or time variations of parameter values are expected to improve model performance.

If good reasons are found to support the use of parameter variation for model improvement, then the issue of how to benefit from this spatio-temporal information must be addressed. Spatially varying parameters can be applied directly in 3D models (e.g., Losa et al., 2006). This should work well for hindcasts and short-term forecasts where the application is not compromised by large scale ecological changes. For forecasting, climatological trajectories such as those estimated by El Jarbi et al. (2013) are likely to be of advantage, although their direct application to long-term prediction in the context of global change would be difficult to justify. Application of spatially varying parameters to long-term predictions of global change is possible but will be more complicated than their use in short-term forecasting and it may be necessary to find ways of allowing spatial patterns in

biogeochemical parameters to evolve with predicted changes in the physical regimes.

8 Emulator approaches

Systematic approaches for parameter optimisation that were successfully applied in 0D or 1D set ups, may become too costly as resolution in space is increased and if the time period for integration is prolonged. This is the case when spatially three-dimensional models with high resolution or steady annual cycles (i.e. periodic solutions) are considered. For the computation of a steady annual cycle (or fixed-point) typically thousands years of model time are necessary, which may result in a number of time steps in the order of $o(10^7)$. Since DA usually involves an iterative optimisation process, typically hundreds or more model evaluations are necessary to obtain a satisfactory parameter set. Thus the necessary time steps during procedures of parameter identification can even reach $o(10^{10})$. Recent attempts aim at replacing computationally costly models with approximations that are less expensive; i.e. emulators have the goal to provide an approximation of the model output trajectory $\boldsymbol{x} := (\boldsymbol{x}_i)_{i=0}^{N_t}$, recalling Eq. (1) of Sect. (2.1):

$$\boldsymbol{x}_{i+1} = M[\boldsymbol{x}_i, \Theta, \boldsymbol{f}_i, \boldsymbol{\eta}_i], \quad i = 0, \ldots, N_t - 1, \tag{26}$$

by substituting the original model M by a simpler one, the emulator ($\widetilde{M}$). Here we disregard a stochastic model approach and consider $\boldsymbol{\eta}_i = 0$ for simplicity.

The application of emulators has emerged in many different fields of science and thus the theoretical background is relatively well developed (e.g. Kennedy and O'Hagan, 2000, 2001; Phillips, 2003; Lucia et al., 2004; van der Merwe et al., 2007; Bliznyuk et al., 2008; Conti et al., 2009; Liu and West, 2009; Castelletti et al., 2012). Two distinct approaches to emulation exist, which we refer to as dynamic emulation and statistical emulation. Both approaches are outlined in the following. Note that the terminology in literature may vary somewhat depending on the respective research field.

8.1 Dynamic Emulators

A dynamic emulator (or reduced order or surrogate model) is a substitute for the original model M. It makes use of the original model equations but is a simpler representation in terms of resolution or details resolved in the dynamics. The term "simple model" refers here to the computational effort needed to evaluate a solution that is a useful approximation of the solution obtained with the full model. A typical number of model evaluations needed for an automised optimisation process can easily reach the order of 10^{10}. In this case an emulator becomes particularly valuable, because its application should be much faster than the original model, while as much as possible main properties of the original model are retained. Only then an emulator-based DA approach will give satisfactory results.

Dynamical or physical emulators are based on a simplified model version ($\widetilde{M}$), which might be additionally aligned with interim evaluations of the original model. The term "dynamic" refers to the fact that the emulator is still based on dynamical physical or biogeochemical equations. These ~~that~~ can be similar to to the ones in the original model but might have some reduced complexity, either by neglecting some processes or by simplifying e.g. the forcing $\widetilde{\boldsymbol{f}}$. Another option is the reduction

of accuracy in model output by coarsening the spatial or temporal discretization. For instance, the Transport Matrix (TM) method (Khatiwala, 2007) can be interpreted as an emulator approach with a kind of coarse model. The TM is an emulator that simplifies the original model M by using an approximated and averaged forcing $\widetilde{\boldsymbol{f}}$ in Eq. (26) and a linear approximation of the spatial discretisation, compared to nonlinear advection schemes typically used in ocean models. For the case of a spin-up, as mentioned above, a reduction of accuracy can be achieved by introducing a different criterion that specifies when a tolerable steady periodic solution as been approached.

When using dynamic emulators, it is often insufficient to take the output of the faster but less accurate coarse model during optimisation, because the accuracy of the coarse model $\widetilde{M}$ might be too low to effectively support parameter search process. It can be worthwhile or even necessary to gradually enhance (or update) the emulator's accuracy during the optimisation procedure by introducing special alignment or correction operators. To explain their definition, let us assume we have computed state vectors of the original and of the coarse model with a current set of values for the parameter vector Θ_ℓ in the ℓ-th step of the optimisation run, i.e.

$$\begin{aligned} \boldsymbol{x}_{i+1} &= M[\boldsymbol{x}_i, \Theta_\ell, \boldsymbol{f}_i], \\ \widetilde{\boldsymbol{x}}_{i+1} &= \widetilde{M}[\widetilde{\boldsymbol{x}}_i, \Theta_\ell, \widetilde{\boldsymbol{f}}_i], \quad i = 0, \ldots, N_t - 1. \end{aligned}$$

We recall that the model state vector $\boldsymbol{x}_i$ consists of the values of the N_x state variables. Thus, in a spatially distributed model, $\boldsymbol{x}_i$ is a vector where every element represents the values at a certain spatial grid point. We here assume that the same numbering is used for the coarse model state $\widetilde{\boldsymbol{x}}_i$.

The alignment operator in optimisation step ℓ is then defined element-wise for $\boldsymbol{x}_i$ and point-wise in time by

$$\boldsymbol{A}_{\ell i} \widetilde{M}[\widetilde{\boldsymbol{x}}_i, \Theta_\ell, \widetilde{\boldsymbol{f}}_i] = M[\boldsymbol{x}_i, \Theta_\ell, \boldsymbol{f}_i]. \tag{27}$$

Thus, every $\boldsymbol{A}_{\ell i}$ is a diagonal matrix. At the current iterate Θ_ℓ, the emulator's output equals the output of the original model. For a parameter vector Θ close to Θ_ℓ, the emulator uses the correction of Eq. (27) – being exact at Θ_ℓ – for the coarse model evaluated at Θ, thus giving only an approximation of the original model. The ~~and the~~ idea of this *response correction method* is that the deviation between both model outputs remains uncritically similar in a vicinity of Θ_ℓ. The emulator is thus not just the coarse model $\widetilde{M}$, but an *aligned* one, $\boldsymbol{A}_{\ell i}\widetilde{M}$, that is now locally optimised. The local optimisation process does not require any additional evaluations of the original model, but only of the cheaper, coarse one. When this inner optimisation gives some new parameter vector $\Theta_{\ell+1}$, the original model is evaluated once again, and the procedure in Eq. (27) is repeated, defining the new emulator for the $(\ell+1)$-th outer optimisation step. In the inner optimisation loop no runs of the original model are needed, and the total number of outer iterations is expected to be lower than in an classical direct optimisation using M. This type of optimisation procedure fits in the framework of trust region methods, a class of state-of-the-art algorithms for which a mathematical convergence analysis is shown in Conn et al. (2000).

The method was successfully applied for parameter identification of a transient 1D configuration with a NPZD ecosystem model and for periodic states with climatological forcing in a three-dimensional setting in a N-based model with dissolved organic phosphorus (DOP) (Prieß et al., 2013a, b). Therein, a coarser time-stepping and a less accurate computation of the

fixed-point (i.e. a shorter spin-up), respectively, was used to construct the simple model $\widetilde{M}$. For this computationally very costly 3D model, it turns out that the most efficient way is to start the optimisation using the emulator- or surrogate-based optimisation procedure (with a very coarse model), and then increase its accuracy during the outer optimisation (Slawig et al., 2014).

8.2 Statistical Emulators

In contrast to a dynamical emulator, statistical emulators relate the input parameters statistically to the model output and thus to $H(\boldsymbol{x})$ regardless of the dynamical model structure. Generally, statistical emulators interpolate the results of a numerical model from a set of training runs with differing parameters. The aim is to approximate the unknown model output for other input parameters, not included in the training parameter set. Common approaches are based on a polynomial fit (of varying degree). Typically, such interpolations are extended by Bayesian techniques to also obtain uncertainty estimates. For this purpose it is commonly assumed that the model outcome can be represented by a Gaussian process and also that the model output changes smoothly as parameter values are varied. Priori assumptions about reliable parameter ranges and their distribution are required. Another ~~necessary a~~ prior choice needed is to determine the respective model output of interest, e.g. results ~~needed~~ required for $H(\boldsymbol{x})$ to determine $p(\Theta \mid \boldsymbol{y})$ or $L(\boldsymbol{y} \mid \Theta)$, Sect. (2.2). Although there are methods available to reduce the dimensionality for multi-dimensional model output (e.g., Higdon et al., 2008; Leeds et al., 2014), it remains practically infeasible to capture the complete output of a 3D-coupled ocean ecosystem model. While the theory for statistical emulation is relatively well described (e.g., Kennedy and O'Hagan, 2000; O'Hagan, 2006; Liu and West, 2009; Conti and O'Hagan, 2010), statistical emulators are so far rarely applied in biogeochemical ocean modelling.

In Fig. (7) an example of a statistical emulator is provided based on a simple NPZD-type box-model. The model setup is adopted from Löptien and Dietze (2015), thereby resolving seasonal variations in photosynthetically available radiation. Since computational costs are low, the chosen example setup would not necessarily require emulation. However, the model is well suited for testing an emulator approach, because it allows us to evaluate a wide range of model solutions. Figure (7) depicts simulated and emulated root mean square (RMS) errors relative to a set of synthetic observations (i.e. with noise added to model results that are obtained for a prescribed set of parameter values). For our example we use the maximum growth rate of phytoplankton and the maximum grazing rate as free model parameters, while all other model parameters remain fixed. The emulation is based on a second order polynomial, following the approach of Kennedy and O'Hagan (2000). The training runs comprise 25 model simulations in a Latin hypercube design, according to (Urban and Fricker, 2010).

Figure (7) shows very similar results for the emulator and for the full model. In particular, the location of the minimum can be well reproduced by the emulator. Thus, the agreement between emulated and simulated model-data misfit is satisfactory and the emulator could be applied for parameter optimisation. The precision might be further enhanced by considering higher order polynomials and/or more trainings data sets. Note, however, that the complexity of the problem increases with the number of free parameters. In particular, the numerous parameter collinearities in biogeochemical models (e.g. Matear, 1995; Kreus and Schartau, 2015; Löptien and Dietze, 2015) can complicate emulation. Increasing the dimension of the model introduces additional difficulties. One suggestion on how to reduce the dimension of a complex model output is given by Hooten et al.

(2011). The authors decomposed modelled surface Chl*a* concentrations of a suite of training runs into singular vectors and predicted the leading modes in dependence of a suite of biological and physical model parameters. During a subsequent parameter optimisation with respect to satellite chlorophyll, they identified zooplankton grazing rate and the light response of phytoplankton to be the most influential parameters. In contrast to most other approaches, where variances are estimated based on Bayesian techniques, Hooten et al. (2011) used a Bayesian approach to estimate the mean values. The study of Leeds et al. (2014) applied a similar technique for DA.

Another example for statistical emulation in biogeochemical modelling is presented by Mattern et al. (2012). Their emulator approach was based on polynomial chaos expansion (e.g. Askey and Wilson, 1985; Wan and Karniadakis, 2006). Mattern et al. (2012) emulated simulation results of Chl*a* concentrations as a function of "maximum zooplankton grazing rate" and the Chl*a*:C-ratio in the Middle Atlantic Bight in year 2006. The authors used an emulator instead of the model to minimise the model-data misfit with respect to daily Chl*a* concentrations observed from remote sensing. They optimised time-constant as well as time-varying parameter estimates. Both approaches improved the overall model performance with respect to Chl*a*. While the original time-varying estimates disregard the actual state of the system, the use of the polynomial chaos method formed the basis of an updated, more reliable method in the study of Mattern et al. (2014) previously discussed in Sect. (7).

Another study of Mattern et al. (2013b) analysed the uncertainty of modelled hypoxia for the Texas-Louisiana shelf based on statistical emulators. The authors investigated the uncertainty due to initial and boundary conditions of biological variables as well as river nutrient loads and phytoplankton growth rate. Additionally, physical factors like river runoff, wind forcing and ocean mixing coefficients were taken into account. The authors revealed considerable uncertainties as their estimates for the hypoxic area varied by more than 40%, when considering reasonable uncertainties in freshwater runoff. Such an extensive analysis would not have been possible without taking advantage of emulators. Furthermore, the use of emulators opens up the possibility of new approaches to exploring the parameter space. One emulator-based technique referred to as "history matching" (Craig et al., 1996), now well-established in other fields and recently applied to the constraint of coupled ocean-atmosphere model parameters (Williamson et al., 2013), seems a particularly promising approach for parameter identification in marine ecosystem modelling. This relatively simple method uses Bayesian inference to rule out areas of parameter space as implausible, given some set of observations. Estimated uncertainties in both the observations (with respect to the truth) and the emulator (with respect to the model) can be taken into account. The method can be applied iteratively with different observation sets to reduce the size of the plausible region at each stage, either as a precursor to more formal model calibration or as a parameter identification method in its own right.

8.3 Combining dynamical and statistical approaches

While emulations based on statistical approaches are comparatively fast, such methods rely on sufficiently large sets of training data (i.e. full model simulations). To generate such training data can be costly, especially for 3D models with high spatial resolutions. To overcome this problem one might consider a combination of statistical and dynamical emulators. ~~The principal advantage of statistical emulators is speed. Dynamical emulators, on the other hand, have the advantage of being inherently related to the models they are emulating. This makes them much less reliant on model output than statistical emulators that rely~~

totally on that output to provide their training data. Suitable training data sets are easy to generate for 3D models with limited domains or fairly coarse resolution. However, the increased integration times required for high resolution 3D simulations or for long spin-up periods inevitably reduce the size of ensemble simulations that can be used for generating training data. It then becomes difficult or impossible to create the relatively large training sets that are needed to explore large high-dimensional parameter spaces.

A two stage emulation process is suggested by Hemmings et al. (2015). Their idea is to use a set of 1D models as a dynamical emulator that describes the evolution of the 3D model at representative sites. This Stage 1 emulator allows large ensemble simulations to be run, providing output that could be used as training data for construction of a statistical emulator (Stage 2). The dynamical emulator of Hemmings et al. (2015) is not used in an inner optimisation loop but is used instead to predict 3D model output for arbitrary parameter vectors. It is thus used more like a statistical emulator. In fact, a particular innovation in their study was to quantify uncertainty in the emulator outputs for inference purposes. Another innovation was the inclusion of biogeochemical perturbations associated with lateral advection that are typically ignored in 1D calibration studies. These were derived by averaging 3D model diagnostics over a 10-member ensemble simulation based on a sample of parameter vectors from the search space. Accounting for the lateral flux information was helpful, contributing strongly to the emulator accuracy. The emulator with uncertainty estimates gave robust results for the surface Chl*a* concentration of an ecosystem model of intermediate complexity, considering variation in 8 parameters.

The ultimate aim of the two stage procedure would be to use a sufficiently large number of state estimates of the model based on a (sufficiently precise) dynamical emulator, which can then be used for the construction of a statistical emulator for a cost function or similar metric. estimates of model output that the dynamic emulator can provide rapidly for 100s of parameter vectors to construct a statistical emulator for a cost function or similar metric that is then used in parameter identification. The dynamical emulator would effectively bridge the gap between a small reference ensemble that is practical to generate with the full 3D model and the statistical emulator that requires a relatively large training set. The respective metric must incorporate an error model that takes into account all sources of uncertainty in the statistical emulation of the full model. Thus, the uncertainty estimates obtained when training the statistical emulator must be inflated by combining them with the dynamical emulator's own uncertainty estimates. Stage 1 emulation results suggest that it may be important to first extend the latter to include temporal covariance estimates for the parametric uncertainty associated with the averaged 3D model output used. Another important consideration is that longer spin-up times for creating the 3D model reference ensemble would be required in a practical application to truly represent the effects of varying parameters in a global circulation model. Another important consideration is that global 3D models require long spin-up times to overcome an initial model drift (see Sect. 9.1). The application of dynamical emulation techniques for accelerated spin-up, such as the TM method (Khatiwala, 2007) mentioned in Sect. (8.1), could help to provide a better representation of the parametric variation by increasing the practical length of the spin-up period.

9 Parameter estimation of large-scale and global biogeochemical ocean circulation models

Global biogeochemical ocean models are commonly used to investigate the mutual interactions between ocean biota and climate change, a famous example being coupled Earth system models (ESMs) applied in the fifth assessment of the Intergovernmental Panel on Climate Change (IPCC, 2014) and those models that are evaluated as part of the Coupled Model Intercomparison Project (CMIP5; Taylor et al., 2012). Besides individual evaluations of biogeochemical ocean model components (e.g., Ilyina et al., 2013; Tjiputra et al., 2013), global ocean biogeochemical simulation results are often specifically evaluated in terms of their representations of the carbon cycle (e.g., Schwinger et al., 2016). More recent studies also focus on analysing the spread of oxygen minimum zones (e.g., Cocco et al., 2013; Cabre et al., 2015).

9.1 Consistency between tracer distribution and ocean circulation field

A major challenge in calibrating biogeochemical models on global scale is that the simulations require many millennia until tracer distributions are in equilibrium with the given circulation field and the biogeochemical processes (Wunsch and Heimbach, 2008). Equilibrium solutions are usually achieved by integrating tracer fields for several thousand years in a so-called model spin-up, based on some seasonally cycling climatological circulation fields. Convergence to steady state conditions depends on the region, tracer type, and form of boundary condition (Wunsch and Heimbach, 2008; Primeau and Deleersnijder, 2009; Siberlin and Wunsch, 2011). It also depends on the values assigned to the parameters of the biogeochemical model, and it is not necessarily a monotonic function of time, but can exhibit inflection points that reflect the interaction of diverse processes happening on different time scales (Kriest and Oschlies, 2015). For parameter optimisation it is meaningful to exclude from a cost function those transient model solutions that involve continuing trends in the redistribution of tracers (see also Séférian et al., 2016).

To attain some equilibrated biogeochemical cycling requires considerable computational time, which makes it particularly difficult to employ methods that exploit the parameter-cost function manifold with a large ensemble of model runs like the MCMC method. ~~used in our mesocosm example in Sect.).~~ The derivation and application of emulators, as described in Sect. (8), is therefore of great value for parameter optimisation of global biogeochemical ocean models. An alternative approach to accelerate the spin-up time is to apply Newton-Krylov methods, by iteratively solving the dynamical system for steady state (e.g., Khatiwala, 2008; Li and Primeau, 2008; Piwonski and Slawig, 2016).

Some speed up of long-term model simulations can also be achieved with an appropriate balance between a model's spatial resolution and the complexity of biogeochemical tracer dynamics, as approached by Ridgwell et al. (2007). ~~The advantage of their model is the coarse spatial resolution as well as the "abstract" design of the biogeochemical tracer equations, allowing time steps as large as 0.05 yr (≈ 18 days) for integration. This way~~ Using a coarse grid and a time step of 0.05 yr (≈ 18 days), they could apply an Ensemble Kalman Filter for estimating parameters of their relatively "abstract" biogeochemical component of an ESM of intermediate complexity, building on a DA setup of ~~a preceding calibration study of~~ Annan et al. (2005). Another option is to ~~considerably~~ decrease the number of model runs ~~needed for parameter optimisation~~ by applying the variational adjoint method for parameter optimisation (Sect. 2.2.3). ~~can be efficient in this respect. An adjoint global bio-~~

~~geochemical model was devised and utilised in the study of Tjiputra et al. (2007).~~ Results of an adjoint global biogeochemical model were used by Tjiputra et al. (2007) to determine first derivatives of a cost function with respect to the parameters ~~values~~, see ~~Sect. (5.3.3) and~~ also Appendix (C). ~~With this derivative information they could apply a Quasi-Newton descent algorithm for a more target-oriented cost function minimisation, which effectively reduced the number of model runs needed for parameter optimisation.~~ However, because of local minima or flat regions in the cost function ~~one associated problem is that~~ optimal estimates may then depend on the initial guess of parameter values, as discussed in Sect. (2.2.3).

Some DA applications may not require equilibrated tracer dynamics to maintain steady seasonal cycles, e.g. when applying sequential DA approaches with recurrent analyses steps and corrections of the simulated state variables. An example is the study of Simon et al. (2015) who introduced an ensemble-based DA method for a large-scale biogeochemical model of the North Atlantic and Arctic Ocean. The focus of their study was to estimate spatial and temporal variations of phytoplankton and zooplankton loss rate parameters as well as model states, in order to establish an operational system for hind- and forecasts of Chl*a* concentrations. Their model was initialised with climatological data of nutrients and oxygen and initial values of the other biogeochemical state variables were set to low constant values. Prior to the DA period (2007-2010) their model was integrated for a six year period, starting in year 2000. This simulation period is much shorter than the few hundreds of years typically needed to equilibrate tracer distribution and ocean circulation in the North Atlantic and Arctic Ocean (e.g., Wunsch and Heimbach, 2008) and the optimised hindcast simulations may therefore not be expected to represent detrended seasonal cycles of biogeochemical tracer distributions and mass flux.

In summary, various procedures for calibrating large-scale and global biogeochemical ocean circulation models ~~are extensive and~~ exist, but are presently challenged by overcoming limitations in computational time to approach equilibrated steady cycles in biogeochemical tracer distributions. ~~and the ocean circulation. Apart from aspects of computational costs,~~ Data availability on global scale introduces additional limitations to act as constraints for parameter identification of global biogeochemical models.

9.2 Data for parameter estimation and calibration in global ocean biogeochemical models

In regard to the ocean's key role in global carbon cycling and hence for the climate system, four different types of data are typically considered for assessing and calibrating global biogeochemical ocean models: i) ~~climatological~~ data of dissolved inorganic tracers, e.g. distributions of nutrients, oxygen, alkalinity and dissolved inorganic carbon, ii) data products derived from remote sensing measurements, e.g. of chlorophyll *a*, or plankton primary- or net community production, iii) in situ measurements or composite data of organic and inorganic matter concentrations, fluxes and rates e.g. at different time-series stations, and iv) observations of the gravitational flux of organic particles to the ocean interior, transporting particulate organic matter through the water column.

For the calibration and assessment of large-scale or global biogeochemical models many studies resort to using climatological data sets, e.g. of nutrients and oxygen, components of the carbonate system (e.g., Watanabe et al., 2011; Tjiputra et al., 2013). Also common is the additional or exclusive use of observational estimates that were derived from remote sensing measurements, like primary production rates and surface concentrations of Chl*a* (e.g., Carr et al., 2006; Tjiputra et al., 2013;

Nevison et al., 2015; Simon et al., 2015). Given the often high level of structural complexity of ocean biogeochemical models we find only few studies that involved more elaborate data such as organism groups or fluxes of organic matter. Examples can be found in Gehlen et al. (2006), who compared simulated and observed particle fluxes, or Aumont et al. (2015), who compared simulated and observed dissolved iron concentrations and nitrogen fixation rates. Likewise, Ward et al. (2012) considered satellite based estimates of surface Chl*a* concentrations of different taxonomic groups as specified in Hirata et al. (2011).

One reason for the ~~usual~~ fallback to rather basic data types such as climatological nutrient concentrations for global model evaluation is the sparse distribution of open ocean, in situ observations. ~~A good~~ One example is the scarcity of global microzooplankton biomass observations in the ocean, as depicted in Buitenhuis et al. (2010). Direct, in situ, open ocean ship-based observations are sparse in space and time mainly for logistic reasons (and costs) and we therefore find available sets of situ data to be noticeably biased towards certain areas and periods (e.g. towards coastal areas, summer season in the high latitudes, and the northern hemisphere, Kriest et al., 2010).

Ocean measurements of rates are particularly valuable, but ~~rate measurements~~ these may not be straightforward to ~~be~~ accomplish ~~on a research vessel~~, e.g. isotopic measurements on a research vessel. Some rate measurements may also suffer from large methodological uncertainties, e.g. measurements of nitrogen fixation ~~rate~~. Of similar value, comparable to rate measurements, are observations of oceanic particle flux, as obtained from sediment traps or from optical methods (e.g., Gardner, 2000; Buesseler, 1991). These data provide only patchy information about the particle flux in the world ocean. Their analysis and interpretation are also difficult, since particles produced at the surface are subject to horizontal transport by advection, hampering the establishment of correlations between surface and deep fluxes, particularly for slowly sinking particles (e.g. a meter per day) in energetic current fields (e.g. a meter per second) (e.g., Siegel et al., 2008; Frigstad et al., 2015). ~~Possibly for these uncertainties,~~ Attempts to calibrate global models against individual observations of particle flux have not yet revealed any unique "best" model solution (Gehlen et al., 2006; Kriest and Oschlies, 2013). ~~, although they provide some insight into the general ability of models to reproduce the observed magnitude of the deep particle flux.~~ To establish a consistent linkage between surface primary production rates, e.g. as derived from remote sensing, and observed in-situ measurements of particle flux remains a major challenge. ~~affect plankton growth within the upper ocean's layers in combination with parameters that describe vertical flux of organic matter and those that~~ ultimately determine carbon storage (Kwon et al., 2009). This requires a close look at parameters that link production the euphotic zone to deep carbon export. Parameters that specify vertical flux and remineralisation of organic matter ultimately determine carbon storage (Kwon et al., 2009).

9.3 Parameters relevant for global ocean biogeochemical modelling

~~Parameters of phytoplankton growth and of organic matter remineralisation determine simulated primary production rates. But model calibration against primary production data can be ambiguous, because simulated rates can be tuned to some desired magnitude, by adjusting those parameters that regulate nutrient turnover within the upper ocean layers (e.g. algal exudation rate or assimilation efficiency of zooplankton grazing) without affecting organic matter flux to the ocean interior (Oschlies, 2001). Thus, simulated tracer concentrations at great depth (> 1000 m) may not critically depend on parameters that specify seasonal variations in primary production. The large scale distributions of dissolved nutrients and of oxygen are sensitive to changes in~~

~~ocean circulation dynamics. However, deep global tracer concentrations are also sensitive to biogeochemical parameters that describe global particle export at depth (e.g. between 500 and 2000 m). This sensitivity suggests that global nutrient and oxygen data may help to identify credible values for parameters of particle flux and remineralisation (Kwon and Primeau, 2006; Kriest and Oschlies, 2013). In this manner the model s ability to adequately simulate organic matter flux and carbon storage in the ocean can be improved, e.g. for refining future projections.~~

The joint effect of particle flux and remineralisation is often described by one or two parameters in global models. Early models referred to an exponential function of remineralisation with depth (Bacastow and Maier-Reimer, 1991), which - in equilibrium - would correspond to a constant particle sinking velocity and constant remineralisation. Another, common description of particle flux (and hence of subsequent remineralisation) is the consideration of a power law of depth: $F(z)\ z^{-b}$, where b is usually set to $b = 0.858$, representing the open-ocean composite value derived by Martin et al. (1987) from sediment traps (e.g., Maier-Reimer, 1993). Empirical fits to various observations of particle flux suggest that b may vary between 0.3-1.4 (Martin et al., 1987; Berelson, 2001; Van Mooy et al., 2002; Buesseler et al., 2007). This typical range of variation of b has been used and tested in global biogeochemical models e.g. analysing how its value affects dissolved tracer concentrations in the ocean (Kwon and Primeau, 2006, 2008; Kriest and Oschlies, 2013). Kwon et al. (2009) coupled a simple global biogeochemical model with a one-box atmosphere and found a large effect of this parameter on atmospheric pCO_2, highlighting the relevance of this parameterisation in ESM simulations. Since this parameterisation is widely used (e.g., Kwon and Primeau, 2006, 2008; Najjar et al., 2007; Parekh et al., 2005) we will have a closer look at its implicit assumptions in the following and discuss potential constraints for the estimation of respective parameters.

Under steady state conditions b can be interpreted as being equal to a constant remineralisation rate r divided by a particle sinking speed a that increases with depth: $b = r/a$ (Kriest and Oschlies, 2008). The associated potential mechanisms that may lead to a vertical increase in sinking speed are selective export of large and fast particles to deeper layers, or repackaging of small particles into larger ones by zooplankton egestion. An alternative interpretation is to assume the sinking speed to be constant while the remineralisation rate decreases with depth. This implies that particles may become more refractory and less susceptible to bacterial degradation, or that bacterial activity is reduced by the decrease in temperature at depth. Other parameterisations of particle flux profiles have been applied in global models, e.g., constant sinking and remineralisation (leading to an exponential flux curve; e.g., Bacastow and Maier-Reimer, 1991), or models that explicitly simulate different groups of particles with different size and properties (e.g., Gehlen et al., 2006; Schwinger et al., 2016). Cabre et al. (2015) provide an excellent overview about different parameterisations for models applied in CMIP5.

So far few attempts have been made to systematically calibrate parameterisations of particle export and remineralisation in global biogeochemical models. Kwon and Primeau (2006) assimilated annual mean phosphate data into a simple global ocean biogeochemical circulation model to optimise globally uniform b. Their study shows that the value of of $b \approx 1$ can be well identified for their model when using global climatological data. According to their approach, the tracer distributions are dynamically consistent with their solution of ocean circulation. ~~Note that their findings do not imply that steady state profiles of export and remineralistion, and thus of b, can be derived by applying any simulated circulation field to climatological data, e.g. of phosphate. Wilson et al. (2015) have addressed this aspect and in their study they show this shortcoming to be mainly due to~~

~~uncertainties in the circulation-driven transport and in observations. Furthermore, it may also be associated with uncertainties in the transport and remineralisation of dissolved organic matter.~~ Such consistency is relevant and b may not be derivable by applying any simulated circulation field to climatological data, e.g. of phosphate (Wilson et al., 2015). Furthermore, Wilson et al. (2015) also discussed how the identification of b is affected by uncertainties in the transport and remineralisation of dissolved organic matter.

In a recent study of Kriest et al. (2017) the export parameter b turned out to be well identifiable, with an optimal value of ≈ 1.3, based on annual mean climatologies of dissolved nutrients and oxygen. As in Kwon and Primeau (2006) their biogeochemical model explicitly resolves seasonal cycles. Plankton parameters that act on seasonal scale within the upper, near surface layers are more difficult to identify, if annual mean climatological data are used. Figure (8) exemplifies this difficulty, based on results from Kriest et al. (2017) who optimised six biogeochemical parameters in total. The example reveals differences in the sensitivity of the cost function with respect to variations of two contrasting parameters, the zooplankton mortality (κ_{zoo}) and b respectively. These differences can be visualised from projections of the parameter-cost function manifold $(\mathbf{\Theta}, J(\mathbf{\Theta}))$, as obtained during parameter optimisation (Schartau and Oschlies, 2003; Ward et al., 2010). To better illustrate the discrepancy between the two parameters in Fig. (8) we defined two arbitrary cost function threshold limits $\Delta_J = J(\Theta)/J(\widehat{\Theta}) - 1$ and $\Delta_J = 0.01$ and $\Delta_J = 0.001$ (see Eq. 16 in Sect. 5). The projected pattern of the zooplankton mortality reveals a much smaller sensitivity of the cost function (larger uncertainty), compared to the robust (nearly quadratic) pattern of the export parameter b. Furthermore, for κ_{zoo} some bimodal structure exists within $\Delta_J \leq 0.01$, which impedes parameter identification. Clearly, annual mean climatologies of dissolved inorganic tracers provide only little information on plankton dynamics in the upper layers, while particle export dynamics (which integrates over large spatial and temporal scales) are well constrained by the large-scale distribution of dissolved inorganic tracers. Thus, simulated tracer concentrations at great depth do not critically depend on every parameter that specifies growth and mortality of the plankton.

In the presence of very diverse time and space scales, which is typical in global biogeochemical ocean modelling, the selection of data sets and the definition of the error model strongly affect parameter identification. We also stress that parameter estimates of global biogeochemical modelling studies are conditioned by the applied circulation, which can have a large impact on simulated tracer fields (Najjar et al., 2007), and by the boundary conditions of e.g. of organic matter burial at the sea floor (Kriest and Oschlies, 2013). To date, it remains unclear whether parameters optimized for a given circulation field will improve model simulations in a different setting, e.g. with a different circulation or forcing, as induced by climate change scenarios.

9.4 Impact of parameter uncertainties on climate model projections into the future

A typical large-scale application of marine biogeochemical models is their use in ESMs from which projections of future climate change can be derived for different emission and land-use scenarios. Output of such models helps to inform scientists, but also society and policymakers about possible consequences of human action on the climate system. A key example is the most recent assessment report of the IPCC that featured ESMs with fully interactive carbon cycles (IPCC, 2014). An appropriate treatment of the uncertainties contained in the applied scenarios and employed models is crucial for correctly interpreting model projections, informing the societal debate about climate policies and thus strengthening the base for developing relevant

measures. A full treatment of uncertainties in the projections of ESM is beyond the scope of our review and we can only address this topic here briefly.

A comprehensive attempt to account for uncertainties in the models when determining likelihoods of reaching certain climate goals, like the politically widely accepted 2°C warming goal, was presented by Steinacher et al. (2013) and Steinacher and Joos (2016). Employing a somewhat simplified ESM of intermediate complexity, they ran perturbed parameter ensembles with some ad hoc assumptions about prior probability distributions of the model parameters. The skill of individual ensemble members was then measured by comparison of model hindcasts with available observations of the current state of the Earth system. A single, pragmatic skill score was used in the assessment and led to an improved posterior estimate of parameter probability distributions. The model dynamics then mapped the parametric uncertainty onto the model projections. From the large ensemble of model solutions that were, in hindcast mode, not inconsistent with the observational constraints, the authors could then successfully derive likelihoods of reaching various climate goals.

Note that reproducing the current climate state is merely a necessary condition for model skill, but may not constrain the model's ability to correctly simulate the sensitivity to natural or anthropogenic environmental change. Observational information on past climate change, such as glacial-interglacial changes may help to better constrain the models' sensitivity to changing environmental conditions, even though no historical analog of the current anthropogenic perturbation is known in terms of the rapid rate of change. Still, any information about model sensitivities to applied perturbations is extremely valuable, be it derived from lab or mesocosm experiments or from historical information. DA is a promising tool to combine such information on very different space and time scales and to develop an improved understanding of how the earth system works and may respond to ongoing environmental change.

10 Summary and perspectives

The survey of Arhonditsis and Brett (2004) revealed that relatively few aquatic biogeochemical modelling studies a) considered parameter optimisation (8.5%), b) provided values of data-model misfit (30 %), or c) performed quantitative parameter sensitivity analyses (28%). Since then there has been a vast increase in the number of those studies where the assimilation of biological and chemical data into planktonic ecosystem models is described. Likewise, we now find a wide field of different studies that address problems of parameter identification. Although positive, this development has also brought up diverse approaches whose contexts and connections are sometimes difficult to understand. Furthermore, we face a variety in terminology and notation, which makes it even more arduous to comprehend the various studies and the significance of their findings. With this review we aim to provide support to readers.

The theoretical backbone for studies of parameter estimation and uncertainty builds first of all on how model errors and observational errors are treated. Specifying the error model is an essential first step in the workflow of parameter identification, enabling the subsequent derivation of conditional probabilities and cost functions. Our review shows that there is no ultimate standard error model or procedure but a meaningful practice is to become explicit about these errors and to reconsider the underlying assumptions for discussions of parameter estimates and model results. Whether the DA approach conserves mass

and/or energy is relevant in this respect, depending on the scientific problem addressed. Some ecosystem model applications may not critically depend on mass conservation, e.g. when simulating plankton growth to act as food source in regional simulations of fish stock size and recruitment. In biogeochemical ~~models~~ applications the conservation of mass can be essential, in particular for large-scale or global ocean ~~simulations~~ applications.

As in many other fields of science, the basic estimation methods considered in plankton ecosystem DA studies are Bayesian estimation and Maximum Likelihood. Their major differences are how prior information enters the DA approach and how estimates and uncertainties are evaluated. The consideration of prior parameter values from preceding studies is meaningful and likely alleviates parameter identification problems. A drawback then is that asymptotic (point-wise) approximations of posterior uncertainty covariance matrices, as described herein, may not apply. But when the model parameters in question have been estimated before in a number of comparable settings, it may seem a tragic waste of effort and information to pursue an ML approach without prior information. A similar issue arises in specifying an "ignorance" prior, and the choice of using BEs when no prior information is available can also be questioned.

We included a section on typical basic parameterisations of plankton models, mainly to stress that the treatment of light- and nutrient limitation may differ between modelling studies. Furthermore, we touched on the problem of resolving phytoplankton losses specified by e.g. grazing and aggregation parameters. Latest plankton growth models account for physiological acclimation effects, responsible for variations between carbon fixation, cellular allocation of nitrogen and phosphorus, and Chl*a* synthesis. Those variations are relevant for DA, in particular if flux estimates of carbon (e.g. CO_2 utilisation and respiration) are of primary concern. It is thus worthwhile to discuss some of the underlying dynamics that can be resolved with the plankton ecosystem model rather than treating it as a "black box" for simulating Chl*a* concentrations.

Many acclimation or optimality-based models have been qualitatively calibrated with data from laboratory experiments. DA approaches for parameter estimation were only done in a few of these studies. Going from laboratory data to the assimilation of data from mesocosm experiments can be a useful intermediate step for testing e.g. acclimation or adaptive models and for assessing uncertainty ranges of parameter values. In this respect, parameter estimates of one experiment can be used for cross-validation with data of another independent mesocosm experiment. On the one hand, simulations of the physical environment of mesocosms are easier to implement, compared e.g. to setting up a 1D model for an ocean site. On the other hand, parameter estimates obtained from the assimilation of mesocosm data might not be representative for ocean simulations. Although more difficult, model cross-validations between different ocean sites or regions provide valuable insight, eventually specifying a model's predictive skill under oceanic conditions.

Some studies have shown that an increase in model complexity may not automatically improve predictive skill. This can be partially attributed to over-fitting, which can yield parameter estimates that improve model-data misfits at one site but induce unreasonable model results at other ocean sites. Such results illustrate the vital role played by well-designed cross-validation experiments. A critical element of cross-validation is whether the assimilated data are truly independent from the data used for testing model skill. This is, for instance, not typically the case if observations from different years but of the same characteristic region are used unless inter-annual variability dominates over the repeating seasonal dynamics. Regional

differences between parameter estimates are informative and have the potential to reveal a model's limitations in a way that can suggest improvements.

Parameter identification becomes more difficult as we go from local and regional scale to large-scale and global model simulations. Algorithms for parameter optimisation require multiple model evaluations, which can be computationally expensive for global biogeochemical models. The procedure for optimising parameter values can be accelerated with the application of an emulator. We discussed the use of dynamical and statistical emulators. The dynamical emulator is a simpler representation of a full model operator that is computionally expensive, thereby approximating the underlying model dynamics. A statistical emulator interpolates model output from a set of training runs with different values assigned to the parameter vector. Based on the derived statistics it can be applied to approximate unknown model output for other input parameters. Both emulator approaches have been shown to efficiently support the search for optimal parameter values. The development and use of emulators of biogeochemical models will likely gain in importance along with improved computer performance. A promising approach is to apply models with coarser resolution or a series of 1D models (distributed over ocean regions) as dynamical emulators for 3D global biogeochemical model simulations. Studies have shown that sufficient accuracy of the emulator can be achieved with repeated intermediate alignments of the dynamical emulator. Alternatively, differences between 1D- and 3D results can be statistically quantified as emulator uncertainty, impacting on the parameter search process and used to modify the emulator-based cost function.

Parameter identification in global marine biogeochemical circulation models is still in its infancy, due to the high computational requirements, the huge range of spatial and temporal scales to be covered, and the comparatively sparse spatial-temporal distribution of data in the ocean. In contrast to local optimisations, the consideration of all relevant spatial and temporal scales has one major advantage in that it provides the opportunity to rigorously test and benchmark biogeochemical models. In addition to tasks and complications mentioned in our review, care must be taken in the selection of appropriate data sets, assuring their relevance (or potential) for answering the questions posed. Moreover a critical evaluation of the respective roles of physics, biogeochemistry, exchanges across the model's boundaries and, possibly, ecology is an as yet unresolved task.

A recurring problem associated with parameter optimisation is that marine biogeochemical models are often unrealistically simplified, while at the same time remaining unconstrained by data. Ideally, models should be developed to minimise the number of uncertain parameters yet maintain a level of complexity that is suited to their intended use in answering specific questions (e.g., Denman, 2003). To accomplish this we may not only think of new model approaches, but also of collecting respective data that can help to constrain solutions of these models.

10.1 Modelling prospects

A commonality of new model formulations is to focus on principles, e.g. by considering the adaptation of traits towards optimal trade-offs (e.g., Wirtz and Pahlow, 2010; Dutkiewicz et al., 2009; Smith et al., 2015), or by accounting for allometric relationships in growth and plankton interaction (e.g., Banas, 2011; Acevedo-Trejos et al., 2015), or by using microbial traits in a functional gene approach (Reed et al., 2014). Recent studies have begun to simulate ecosystem complexity and allow the model to "self-organise" according to a relatively simple set of ecological and physiological rules or "trade-offs" (Bruggeman

and Kooijman, 2007; Follows et al., 2007). A major advantage of this approach is that the models are able to resolve greater ecological diversity with fewer specified parameters whose values can be assumed to be spatially invariant. This diversity allows the simulated plankton community to reorganise across broad environmental (e.g. spatial) gradients. But the identification of the most important trade-offs governing competition between organisms remains a major challenge (Tilman, 1990; Litchman et al., 2007, 2012).

Perhaps one of the most remarkable developments is the revival of thermodynamically inspired ecosystem theories for modelling biogeochemical cycling in the oceans (e.g., Vallino, 2011). In the review of Vallino and Algar (2016) the concept and potential of the maximum entropy production principle are addressed. In this modelling approach life in the ocean is perceived as units of e.g. covalent bonded chains of carbon atoms that create disequilibria of energy and mass between organisms. These disequilibria lead to different functional pathways in biogeochemical cycling, accompanied by a flexible evolution of structural dependencies between nutrient or substrate availability, plankton and other organisms. Such novel or revised approaches are expedient and help to create new ideas in terms of how to design models and measurement strategies that may alleviate the problems of parameter identification.

10.2 Examples of recent advances in data availability

The use of previously underexploited data sets (for example those linking organism size to key ecophysiological rates; Baird and Suthers, 2007; Banas, 2011; Ward et al., 2012) have the potential to bring new constraints on model behaviour, and may go some way to alleviating the degree of underdetermination that is typically associated with parameter estimation. New data sources, such as the Bio-Argo profiling floats, should also advance our understanding, e.g. by documenting seasonal variations of deep Chl*a* maxima in remote oligotrophic regions (Mignot et al., 2014). These Bio-Argo profile data have the advantage that they resolve biogeochemical properties with a relatively high frequency of five to ten days over a sampling period of up to two years.

A substantial fraction of recent fluourescence measurements from Bio-Argo platforms has already been included in a new global Chl*a* database described and provided by Sauzède et al. (2015b). Their quality-controlled data comprise profiles of total Chl*a* concentration together with some additional estimates of the relative contributions from pico-, nano-, and micro phytoplankton. The employed relationship between the relative size distributions and total Chl*a* concentration was derived from an extensive analysis of High- Performance Liquid Chromatography pigment data in combination with Chl*a* fluorescence measurements (Sauzède et al., 2015a). The consideration of these profile data will possibly facilitate the estimation of photoacclimation parameters in particular, and of phytoplankton growth parameters in general.

Data products from remote sensing measurements are continueously improved and new empirical relationships between photosynthesis and respiration are derived to estimate net community production (NCP) on the global scale (e.g., Westberry et al., 2012; Tilstone et al., 2015). These spatially resolved estimates may help to constrain parameters of plankton respiration and remineralisation rates. In spite of large uncertainties, the assimilation of NCP estimates from remote sensing into biogeochemical models may impose additional constraints on parameters that affect solutions of air-sea exchange of CO_2 and of organic matter export. In this respect we also stress that upgrades and analyses of time-series data are more then ever essential

to make inference about organic matter flux and ecosystem functioning (e.g., Emerson, 2014), which may introduce additional constraints for identifying values of a larger number of parameters of plankton ecosystem models. Finally, we point to latest products from compilations and syntheses of oceanic and atmospheric CO_2 data collected by a large international community (Rödenbeck et al., 2015; Bakker et al., 2016). Data products like air-sea CO_2 flux of specified ocean regions (biomes), as derived in (Rödenbeck et al., 2015), in combination with data of nutrient concentrations and O_2 will likely put new light on those parameters that determine variations of the elemental stoichiometry (C:N:P:O_2) in model results of inorganic and organic matter cycling.

10.3 Harmonising research foci in marine ecosystem modelling and data assimilation

The application of DA methods has become standard for calibrating marine ecosystem- and biogeochemical models. But scientific insight can differ between DA studies considerably. In the literature we find that there is often an imbalance between level of sophistication of the ecosystem model used and the DA method employed. This is likely due to the fact that marine ecosystem-/biogeochemical modelling studies integrate knowledge from different scientific fields, of which each has its own foci, objectives, and expertise i.e. plankton ecology, physical oceanography, marine geochemistry, and mathematics and statistics. It is difficult to track major advancements in marine ecosystem modelling when considering the different views from each of these research fields. Furthermore, the design of experimental studies and the collection of field data are often achieved without harmonising the needs of biologists with the modelers' exigencies (Flynn, 2010).

Facets of parameter identification in biological modelling disclose major commonalities and disparities between the objectives expressed in the different research fields. Discussions on parameter identification are therefore helpful to achieve a common understanding and to promote communication between observers, modelers, and statisticians. Problems of parameter identification may thus be well addressed by pooling expertise across multiple disciplines, without losing sight of scientific objectives. Such joint efforts should help planktonic ecosystem models to fulfil their potential as quantitative tools for aquatic sciences.

Author contributions. Individual sections of our review were written by one or more lead author(s), with contributions from the other authors (Phil Wallhead, PW; John Hemmings, JH; Ben Ward, BW; Ulrike Löptien, UL; Thomas Slawig, TS; Iris Kriest, IK; Andreas Oschlies, OA, and Markus Schartau, MS). All authors were involved in mutual revisions of the individual sections. The sections' lead authors are: 1. Introduction (MS), 2. Theoretical background (PW, MS, and JH), 3. Typical parameterisations of plankton models (MS), 4. Error models (PW), 5. Parameter uncertainties (MS), 6. Cross-validation and model complexity (BW), 7. Space-time variations in model parameters (JH), 8. Emulator approaches (UL and TS), 9. Parameter estimation of large-scale biogeochemical ocean circulation models (IK, AO, and MS), 10. Summary and perspectives (MS), Appendix A (PW), Appendix B (MS), and Appendix C (MS and TS). Shubham Krishna performed parameter optimisations, MCMC computations of the mesocosm modelling example, as well as calculations of the 2D parameter arrays.

Acknowledgements. We gratefully acknowledge the support from the International Space Science Institute (ISSI). This publication is an outcome of the ISSI's Working Group on "Carbon Cycle Data Assimilation: How to consistently assimilate multiple data streams". We like

to thank four anonymous referees who provided constructive and helpful comments. The time and effort they spend on our manuscript is much appreciated. The examples of mesocosm data assimilation are based on the mesocosm modelling environment designed for the large integrated projects Surface Ocean Processes in the Anthropocene (SOPRAN, 03F0662A) and BIOACID (03F0728A), both funded by the German Federal Ministry of Education and Research (BMBF). Contributions from Iris Kriest, Ulrike Löptien, and Thomas Slawig were supported by the BMBF funded PalMod - Paleo Modelling: A national paleo climate modelling initiative.

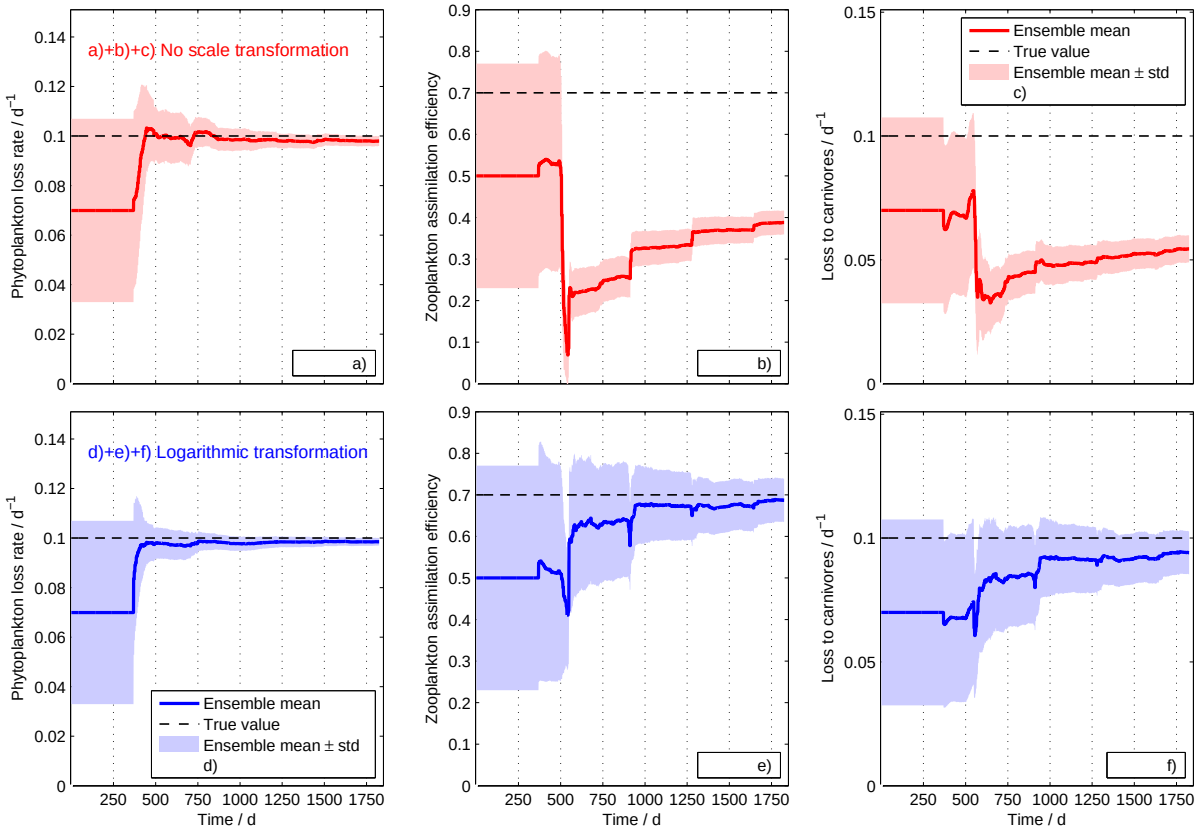

Figure 1. Time evolution of parameter estimates in a simulation test of an Ensemble Kalman Filter using untransformed data (a-c, top row) and using logarithmic transformed data (d-f, bottom row) (Simon and Bertino, 2012, Fig. 3). Solid lines and shading show ensemble means and standard deviations averaged over 20 simulation experiments, while dashed lines show the true parameter values. The data were generated using Gamma-distributed observational errors with standard deviation 30% (see Simon and Bertino, 2012). A transformation can significantly reduce the bias of parameter estimates by the end of the assimilation period. Figure was redrawn from results provided by Ehouarn Simon, with permission from Elsevier. Copyright of figure content by Elsevier.

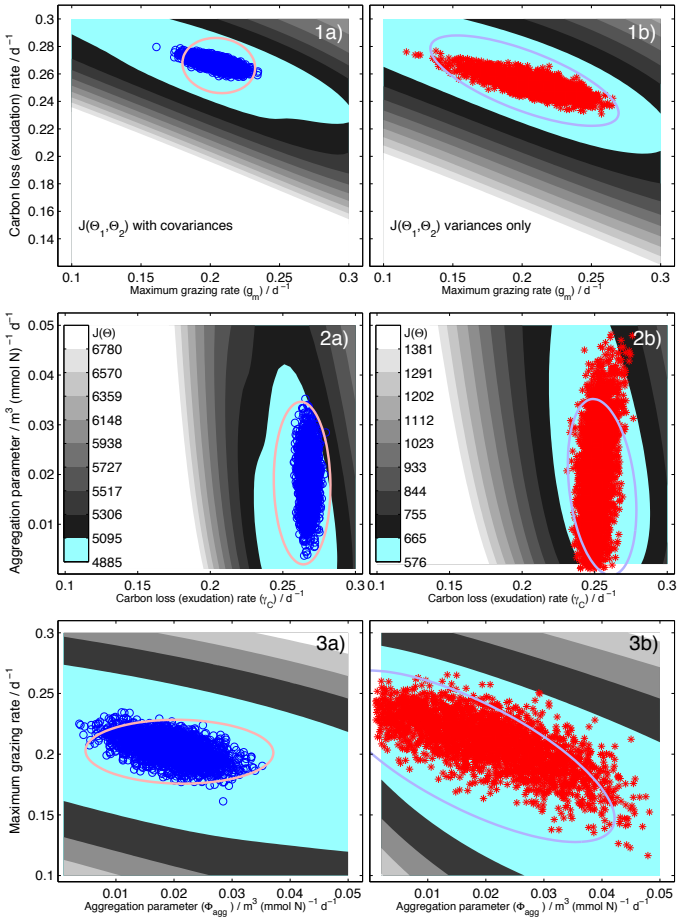

Figure 2. Cost function contours when varying values of a combination of two parameters $J(\widehat{\Theta}_m \pm \Delta_m, \widehat{\Theta}_n \pm \Delta_n)$ around the optimum estimate at $(\widehat{\Theta}_m, \widehat{\Theta}_n, \min(J))$, while values of all other parameters remain fixed. Each plot resolves a pairwise combination out of three parameters that all specify phytoplankton biomass losses. The two columns reveal differences in error margins due to different cost functions with same data for the same model: a) with covariances explicitly regarded and b) all data are assumed to be independent. First row (1a and 1b): combination of maximum grazing rate ($\Theta_1 = g_m$) and carbon exudation rate ($\Theta_2 = \gamma_C$). Second row (2a and 2b): combination of the aggregation parameter ($\Theta_3 = \Phi_{\text{agg}}$) and γ_C. Third row (3a and 3b): combination of g_m and Φ_{agg}. Markers show credible regions of parameter estimates obtained with Markov Chain Monte-Carlo (MCMC) method (dots for J with covariances, asterisks for J with variances only). Error ellipses (lines) depict point-wise 95% confidence regions derived from an approximated and inverted Hessian matrix, according Eq. (22). The cyan colored region embeds all cost function values that are lower than an upper threshold $\triangle J^*(\alpha = 0.05)$, derived from a distribution of $J(\widehat{\Theta}) - J^*(\widehat{\Theta})$, where $J^*(\widehat{\Theta})$ are cost function values at $\widehat{\Theta}$ using resampled data (Fig. B1 in Appendix).

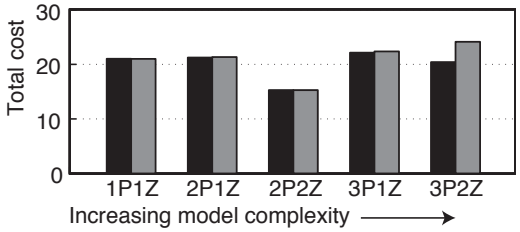

Figure 3. Predictive skill for five ecosystem models of different complexity, after assimilation of satellite data (black) and after assimilation of satellite data with 20% added noise (grey) (Xiao and Friedrichs, 2014a). The most complex model appear to be the most sensitive to errors in the data, in terms of its cross-validated predictive skill.

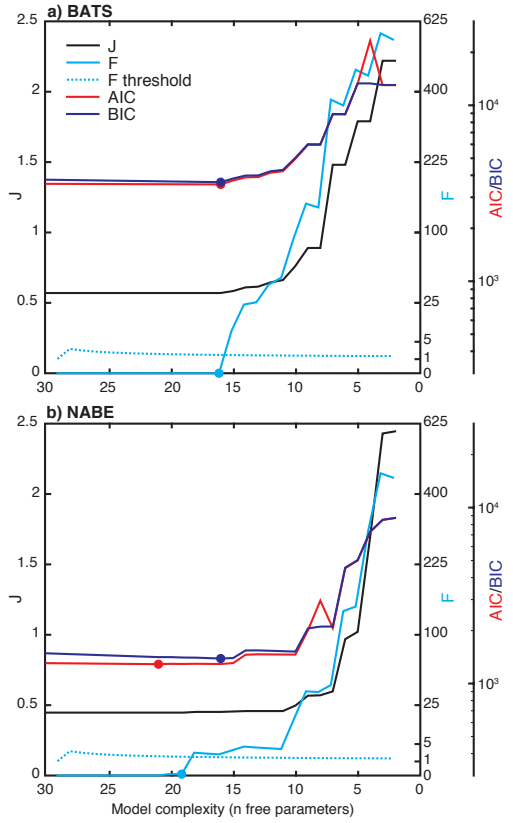

Figure 4. Model selection metrics at the Bermuda Atlantic Time-series Study (BATS) and the North Atlantic Bloom Experiment (NABE), as a function of complexity across a suite of nested ocean biogeochemical models (Ward et al., 2013). The least-squares misfit, J (left-hand axis), increases monotonically with decreasing complexity, as it does not penalise model complexity. The likelihood ratio test, F (first right-hand axis), compares each reduced model to the full model, and selects the simplest that is not significantly worse than the full model (F<F threshold). The AIC and BIC (second right-hand axis) both contain terms that account for model data misfit and complexity, and the optimal model is the one with the lowest score. In each case, the optimal model is indicated by a dot.

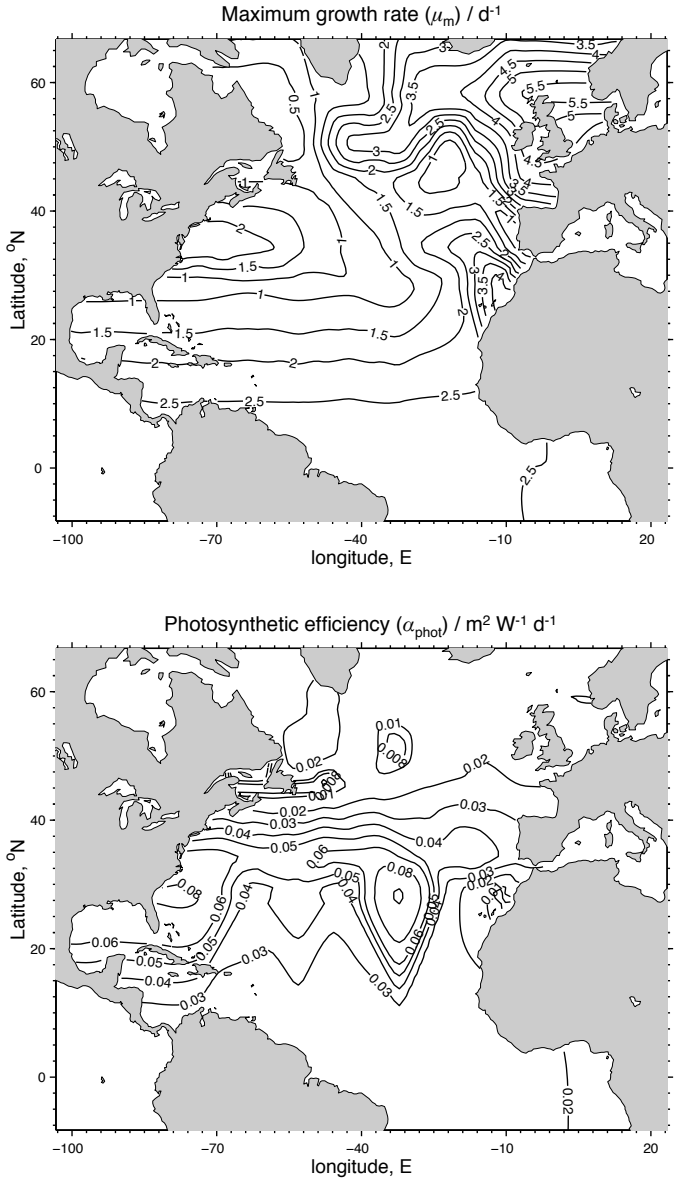

Figure 5. Spatially varying estimates for the phytoplankton maximum growth rate (μ_m in unit d^{-1}) and photosynthetic efficiency (α_{phot}, in $m^2\ W^{-1}\ d^{-1}$) used in a 3D modelling study of the North Atlantic (Losa et al., 2006). The parameter estimates are based on those obtained in a previous assimilation of satellite chlorophyll data (Losa et al., 2004). Permission to include Fig. (2) from Losa et al., (2006) was granted by the authors. Figure is used with permission from Elsevier. Copyright of original figure by Elsevier.

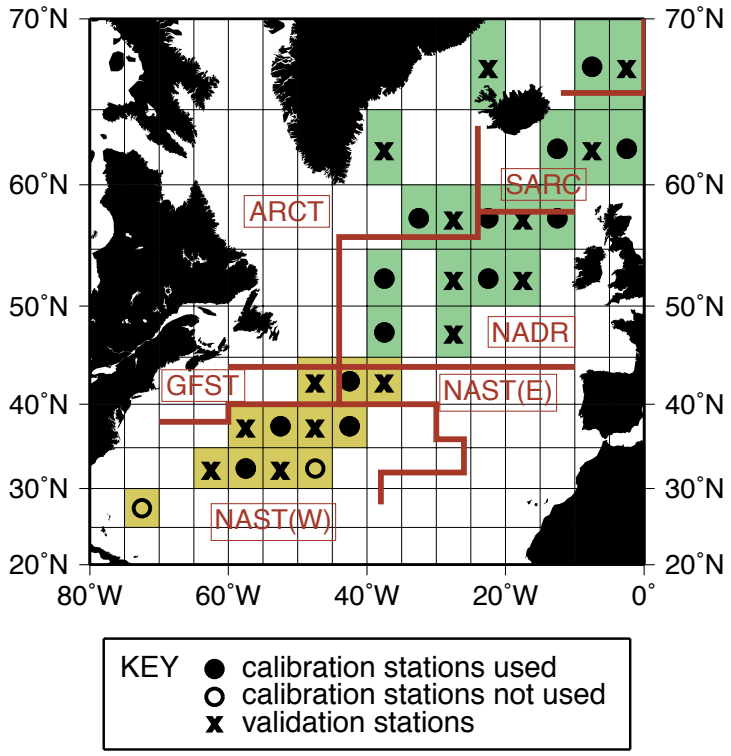

Figure 6. Geographic extent of the two sub-domains giving the optimal calibration in the split-domain calibration study of Hemmings et al. (2004), shown here in yellow and green. Also shown are the distributions of the sites used from the calibration set to obtain the parameter vectors for each sub-domain and the sites used for cross-validation. Biogeochemical provinces defined by Longhurst (1998) are shown for reference. ARCT: Atlantic Arctic Province; SARC: Atlantic Subarctic Province; NADR: North Atlantic Drift Province; GFST: Gulf Stream Province; NAST: North Atlantic Subtropical Gyral Province. Figure (6a) of Hemmings et al., (2004) is shown with permission from Elsevier. Copyright of original figure by Elsevier.

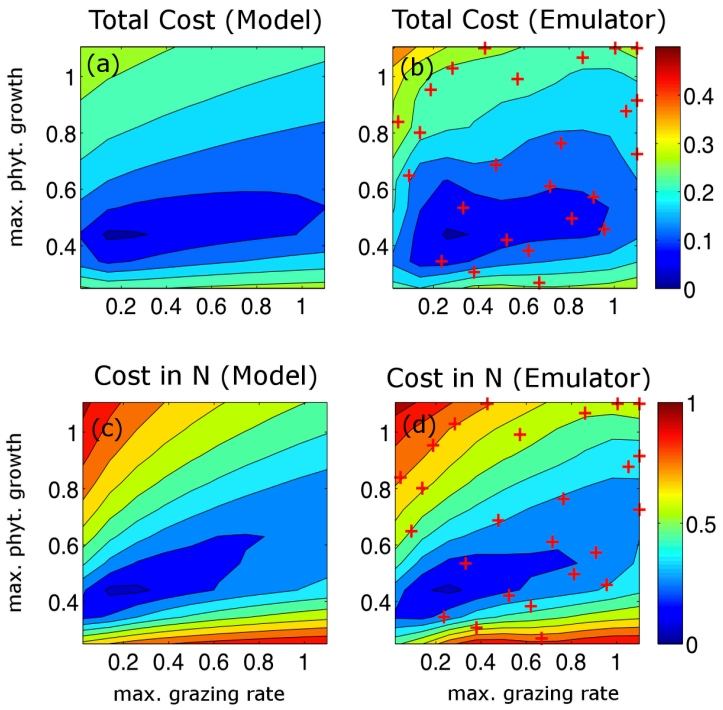

Figure 7. Simulated (a,c) and emulated (b,d) RMS (root mean square) error depending on the maximum growth rate of phytoplankton and the maximum grazing rate. Simulated and emulated RMS-errors are provided relative to "synthetic observations", based on a simulation for a given parameter set (HI=15 W m^{-2}; m=0.06 d^{-1}; μ_{max}=0.51 d^{-1}; H_n=0.8 mmol N m^{-3}; m_{PD}=0.1 d^{-1}; m_{DN}=0.1 d^{-1}; H_Z=0.9 mmol N m^{-6}; m_{ZN}=0.01 d^{-1}; m_{ZD}=0.01 d^{-1}; g_{max}=0.21 d^{-1}), which is disrupted by reddish noise (AR(3)-process) with a standard deviation of 0.09 mmol N m^{-3}. (Notation after Löptien and Dietze, 2015). Sub panels (a,b) are based on all prognostic variables, while the RMS error in (c,d) is based on nitrate (NO_3^-) only (c,d). Red crosses mark the training data.

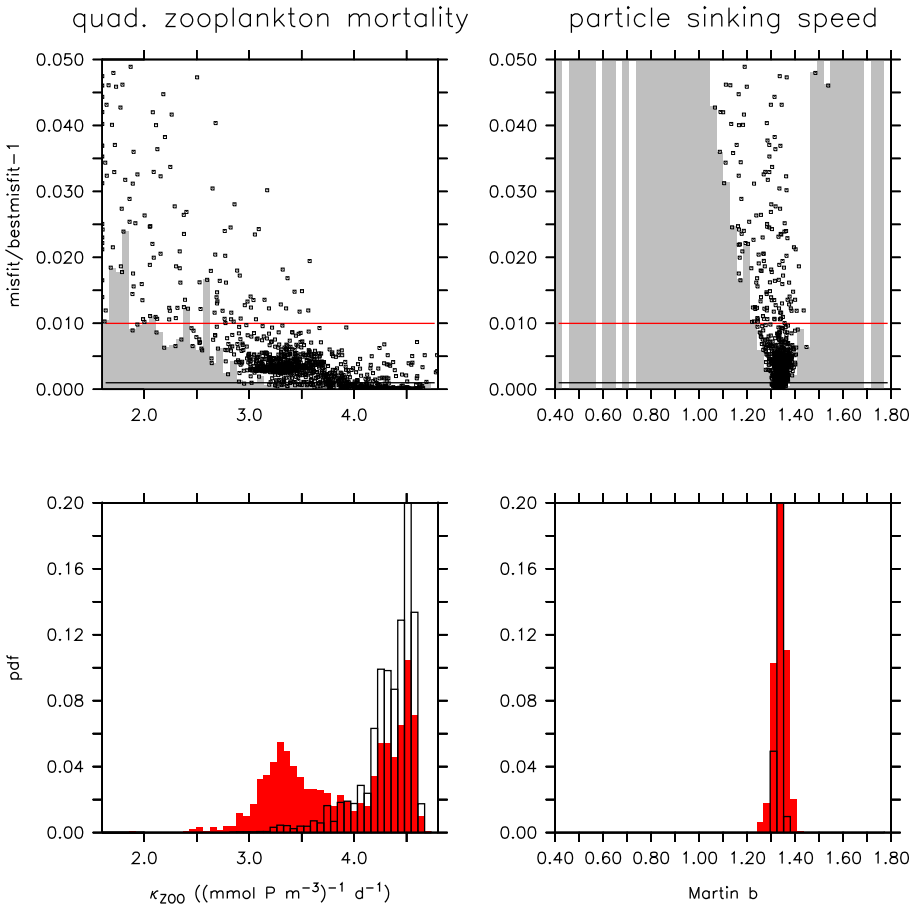

Figure 8. Projections from parameter-cost function manifold $(\widehat{\Theta}_l, J(\Theta))$ as obtained during the optimisation of six biogeochemical parameters. Parameters shown are quadratic zooplankton mortality κ_{zoo} (left panels) and rate of vertical increase of particle sinking speed, a, expressed as quotient $b = r/a$, where r is particle remineralization rate (right panels). Upper panels: cost function (volume-weighted root-mean square error, divided by global mean concentration of each tracer) expressed as its deviation from the minimum. Parameters of all model simulations in the optimisation trajectory were grouped into 50 classes. Grey bars show minimum cost within each class. Red and black horizontal lines indicate deviation from minimum cost of 1% and 0.1%, respectively. Squares show the cost of each individual. Note that the y-axis only extends to 5% above minimum cost at $(\widehat{\Theta}, J(\widehat{\Theta})$. Lower panels: parameter distribution (PDF) of all model simulations, whose cost do not exceed a threshold limit of $\triangle_J = 1.01 \cdot J(\widehat{\Theta})$ (1%, red bars) or $\triangle_J = 1.001 \cdot J(\widehat{\Theta})$ (0.1%, open bars) 0.1% (open bars) of the minimum cost, see Eq. (16 and text).

Appendix A: The Variable Lag Fit with unknown error variances (Sect. 4.4)

In a Variable Lag Fit (VLF), we assume that the truth at time t_i is related to the model output by a kinematic model error (ζ) in phase or time lag τ_i. Equation (2) becomes:

$$\mathbf{x}^t(t_i) = \mathbf{x}(t_i + \tau_i) \tag{A1}$$

A notable feature of this model error representation is that it introduces unknowns τ that can be conditionally optimised by searching forwards and backwards in time within saved model output, i.e. *without rerunning the dynamical model*. For the demonstration in Fig. (A1) we assumed that the time lag errors are normal and independent: $\tau_i \sim N(0, \sigma_\tau^2)$. This independence assumption may seem restrictive; for example, a misplaced eddy might be expected to impose some correlation between the τ_i for a set of cruise data. Nevertheless, we find that the method is somewhat robust to neglected lag correlation. Moreover, this formal neglect enables a large computational simplification since the lags can then be optimised one by one, see Wallhead et al. (2006).

For the observational error in Fig. (A1) we assumed lognormal errors with no interpolation or conversion factors, and that all measured variables were sampled simultaneously. Equation (3) becomes:

$$y_{ij} = x_{ij}^t \cdot \exp\left(\epsilon_{ij} - \frac{\sigma_j^2}{2}\right) \tag{A2}$$

at each measurement time t_i and for each measured variable j (Nutrient, Phytoplankton and Zooplankton). For simplicitly we further assumed that the observational errors were independent between measurements and data types, hence $\epsilon_{ij} \sim N(0, \sigma_j^2)$. Note that the ϵ may be considered to include a component of kinematic model error (ζ) without affecting the parameter estimation, hence we refer to them as *residual* errors below. Assuming that the ecosystem parameters θ_e, time lags $\boldsymbol{\tau}$, time lag variance σ_τ and observational error variances σ are all unknown, a joint posterior mode estimate of $\Theta = (\theta_e, \boldsymbol{\tau}, \sigma_\tau, \boldsymbol{\sigma})$ is obtained by maximising the posterior density $p(\Theta \mid \boldsymbol{y})$, equivalent to minimising the following cost function:

$$\begin{aligned} J(\Theta) =& n \log \sigma_\tau^2 + \sum_i \frac{\tau_i^2}{\sigma_\tau^2} + n \sum_j \log \sigma_j^2 \\ &+ \sum_{ij} \frac{(\log y_{ij} - \log x_j(t_i + \tau_i) + 0.5\sigma_j^2)^2}{\sigma_j^2} \end{aligned} \tag{A3}$$

To test this cost function, we simulated data from the NPZD model of Oschlies and Garçon (1999) in a 0D setting using the parameters values and sine-squared forcing function from Wallhead et al. (2013). Three years of simultaneous weekly samples of N, P, and Z were simulated assuming independent normal time lag errors with standard deviation σ_τ = 10 days and independent normal residual errors $\sigma_{\log N} = 0.1$, $\sigma_{\log P} = 0.2$, $\sigma_{\log Z} = 0.3$. The data were assimilated into the same NPZD model by one of two methods. In the 'standard fit', no time lag error was assumed and search parameters $\Theta = \{\theta_e, \sigma_{\log N}, \sigma_{\log P}, \sigma_{\log Z}\}$ were estimated by minimising only the final two terms in A3 with $\tau_i = 0$ for all i. In the VLF, $\Theta = \{\theta_e, \boldsymbol{\tau}, \sigma_\tau, \sigma_{\log N}, \sigma_{\log P}, \sigma_{\log Z}\}$ was estimated by minimising Eq. (A3). In both cases, we assume uncertainty in only two of the 15 biological parameters, namely the phytoplankton maximum uptake rate V_m and the zooplankton maximum grazing rate g (hence $\theta_e = (V_m, g)$). For all search parameters, allowed ranges were $\pm 50\%$ about the true values, equivalent to unbiased uniform priors with 29% prior uncertainty. Initial values of the search parameters were chosen at random from this prior, and optimisations were repeated over 10 random restarts to avoid local minima. The experiment was repeated over 20 simulated data sets to obtain the statistics in Table (A).

Caution must be exercised here regarding the estimation of σ_τ. If the prior for σ_τ permits very low or zero values then the MAP estimation will push the estimate of σ_τ towards zero irrespective of its true value. This is because, unlike the fourth term in Eq. (A3), the second term

can be made exactly zero with $\boldsymbol{\tau} = \mathbf{0}$ as long as $\sigma_\tau^2 > 0$, in which case the negative contribution of $n \log \sigma_\tau^2$ may produce a spurious, deeper minimum of J near to $\sigma_\tau = 0$. We have found that this spurious minimum need not influence estimation as long as the sample size and the lower limit of the allowed range or rectangular prior for σ_τ are sufficiently large, Fig. (A2). An alternative solution may be to assume a prior that drops smoothly to zero as $\sigma_\tau^2 \to 0$, such as an inverse gamma distribution (cf., Kavetski et al., 2006).

To investigate estimation of the time lag variance parameter σ_τ we obtained cost function profiles by fitting the same data set using a range of fixed values of σ_τ, Fig. (A2). We see that with three years of weekly NPZ sampling the cost function function has a strong minimum close to the true value of 10 days, and this minimum should be approached even if the allowed range (prior uncertainty) for σ_τ reaches as low as 1 day. However, if we decrease the number of sampled years, or especially the number of sampled variables, the minimum becomes weaker and a spurious minimum close to σ_τ = 0 starts to encroach on the profile. A sufficiently low minimum allowed value $\sigma_\tau^{(\min)}$ may then lead to estimates converging to this spurious minimum.

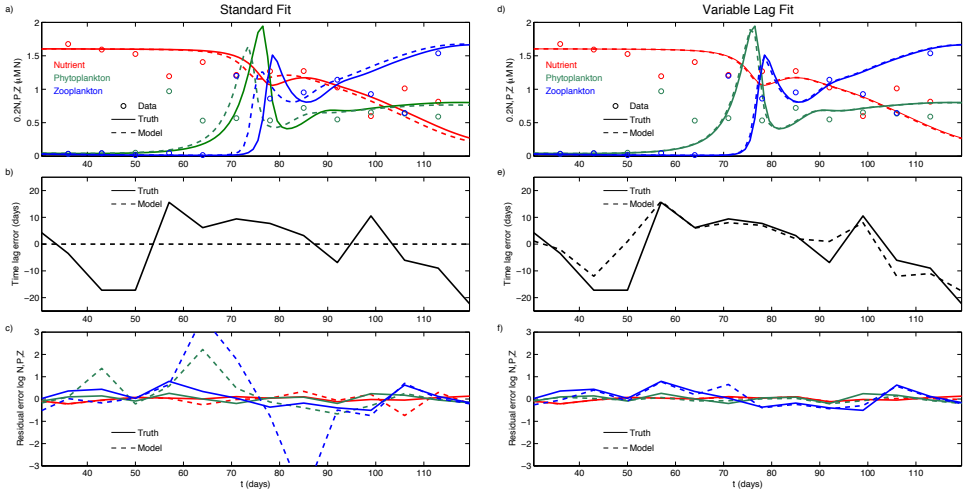

Figure A1. Demonstration of the Variable Lag Fit (VLF) applied to a simulated data set. a) shows the system trajectory with the true parameter values (solid lines), the data (dots) simulated assuming normal and independent time lag errors (σ_τ = 10 days) and residual errors ($\sigma_{\log N,P,Z}$ = 0.1, 0.2, 0.3, see Table A), and the system trajectory with the VLF parameter estimates (dashed lines, overlapping with solid). b) compares the true time lags (solid) with those estimated from the VLF (dashed). c) compares the true residual errors with those estimated by the VLF (dashed, same colour code as in a)). Three years of data were assimilated but only the initial and post-bloom period of the first year is shown for clarity.

Table A. True parameter values and means $\pm$ 1 SD of estimates over 20 simulated data sets, using a standard fit method and a variable lag fit method (see Eq. A3). Three years of weekly NPZ data were simulated using the true values (first row) for the maximum nutrient uptake rate V_m, zooplankton grazing rate g, residual standard deviations $\sigma_{\log N,P,Z}$, and time lag standard deviation σ_τ (for experiments with lags imposed). With no time lags, the standard fit accurately recovers the true parameter values (third row), but with time lags (fourth row) the standard grazing rate estimates are biased and imprecise, while the residual variances have strong positive bias as they are forced to account for the time lag errors. The variable lag fit avoids these biases and accurately partitions the variance between residual error and time lag error (fifth row).

	Lags?	V_m (day^{-1})	g (day^{-1})	$\sigma_{\log N}$	$\sigma_{\log P}$	$\sigma_{\log Z}$	σ_τ (days)
True values	—	0.66	2.00	0.10	0.20	0.30	10.0
First guesses	—	0.66 ± 0.19	2.00 ± 0.58	0.10 ± 0.03	0.20 ± 0.06	0.30 ± 0.09	10.0 ± 2.9
Standard fit	No	0.66 ± 0.00	2.03 ± 0.07	0.10 ± 0.01	0.20 ± 0.01	0.31 ± 0.01	—
Standard fit	Yes	0.68 ± 0.03	2.61 ± 0.44	0.27 ± 0.02	0.46 ± 0.07	0.75 ± 0.14	—
Variable Lag Fit	Yes	0.67 ± 0.01	2.03 ± 0.19	0.07 ± 0.01	0.18 ± 0.02	0.29 ± 0.02	9.2 ± 0.7

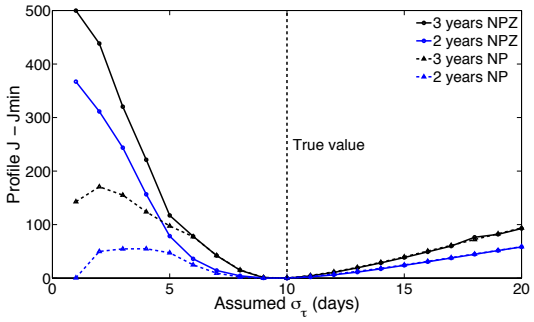

Figure A2. Profiles of the Variable Lag Fit cost function (-2 $\times$ posterior density) relative to the minimum value for a range of assumed values of the time lag error standard deviation σ_τ. For each σ_τ, Eq. (A3) was minimised over (θ, τ, $\sigma_{\log N,P,Z}$) for the same data set. Different curves correspond to different scenarios for the number of sampled years (at weekly sampling frequency) and number of simultaneously sampled variables (black = 3 years, blue = 2 years, solid lines with circles = Nutrient-Phytoplankton-Zooplankton sampling, dashed lines with triangles = Nutrient-Phytoplankton sampling). The extent to which each curve has a deep minimum close to the true value σ_τ(true) = 10 days indicates the feasibility of estimating σ_τ for the corresponding sampling plan.

Appendix B: Mesocosm example (Sect. 5.4)

For our example we account for six different types of measurements from mesocosms of the Pelagic Ecosystem CO_2 Enrichment Study (PeECE I, Engel et al., 2005; Delille et al., 2005): 1) dissolved inorganic carbon (DIC, mmol m^{-3}), 2) nitrate (NO_3^-, mmol m^{-3}), 3) nitrite (NO_2^-, mmol m^{-3}), 4) Chl*a* (mg m^{-3}), 5) PON (mmol m^{-3}), 6) POC (mmol m^{-3}). Concentrations of NO_3^- and NO_2^- are not explicitly resolved by the model and therefore these measurements are combined. We refer to their sum as dissolved inorganic nitrogen (DIN). Thus, the number of components of the observation vector is $\boldsymbol{y}$ is N_y = 5. Observations are available on a daily basis over a period of 23 days (N_t = 23). The vector includes daily means of nine mesocosms at t_i, i = 1, ... , N_t. The dynamical model equations determine twelve state variables (N_x = 12). The corresponding vector of model counterparts to observations is $H_i(\boldsymbol{x})$, with carbon and nitrogen biomass concentrations of phytoplankton (PhyN & PhyC), of zooplankton (ZooN & ZooC), of detritus (DetN & DetC), and carbon concentration of (particulate) macrogels (GelC). The data-model residual vector is:

$$
\begin{aligned}
\boldsymbol{d}_i &= \boldsymbol{y}_i - H_i(\boldsymbol{x}) \\
&= \underbrace{\begin{pmatrix} \mathrm{DIC}_i \\ \mathrm{DIN}_i \\ \mathrm{Chl}a_i \\ \mathrm{PON}_i \\ \mathrm{POC}_i \end{pmatrix}}_{\text{obs}} - \underbrace{\begin{pmatrix} \mathrm{DIC}_i \\ \mathrm{DIN}_i \\ \theta_i^{\mathrm{Chl:C}} \cdot \mathrm{PhyC}_i \\ (\mathrm{PhyN} + \mathrm{ZooN} + \mathrm{DetN})_i \\ (\mathrm{PhyC} + \mathrm{ZooC} + \mathrm{DetC} + \mathrm{GelC})_i \end{pmatrix}}_{\text{model}}
\end{aligned}
\tag{B1}
$$

As an error model we assume additive Gaussian errors applying Eq. (4) in Sect. (2.1.3). The standard errors ($\boldsymbol{\sigma}_i$) represent the observed variability between the nine mesocosms, based on daily measurements. Residual error covariance matrices can thus be derived for every sampling day: $\mathbf{R}_i = \boldsymbol{S}_i \, \boldsymbol{C}_{(\boldsymbol{y})} \, \boldsymbol{S}_i$. The matrices $\boldsymbol{S}_i$ include diagonal elements with $\boldsymbol{\sigma}_i$ at date t_i, while off-diagonal elements are zero. The elements of matrix $\boldsymbol{C}_{(\boldsymbol{y})}$ represent correlations between the different types of observations, which were determined for two time intervals: exponential growth and post-bloom period. The distinction between periods of bloom buildup and post-bloom can be particularly meaningful when C and N (or P) data are assimilated. Correlations can switch sign and thus the sign of the data-model residual $\boldsymbol{d}_i = \boldsymbol{y}_i - H_i(\boldsymbol{x})$ matters. For example, PON and dissolved inorganic carbon (DIC) are strongly negatively correlated during the exponential growth phase. During the post-bloom period DIC may still decrease at times when PON concentration declines as well, which yields a weak but positive correlation.

The standard errors (σ_i) can be written in matrix notation with off-diagonal elements being zero:

$$
\boldsymbol{S}_i = \begin{pmatrix} \sigma_i^{(\mathrm{DIC})} & 0 & \cdots & 0 \\ 0 & \sigma_i^{(\mathrm{DIN})} & \cdots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 0 & \cdots & 0 & \sigma_i^{(\mathrm{POC})} \end{pmatrix}
\tag{B2}
$$

Correlations during exponential gowth (t_i; i = 1, ... , 13) / and during post bloom period (t_i; i = 14, ... , 22):

$$\boldsymbol{C}_{(\mathbf{y})} = \begin{pmatrix} \text{DIC} & \text{DIN} & \text{Chl}a & \text{PON} & \text{POC} \\ 1 & 0.96/0.2 & -0.95/-0.22 & -0.97/\mathbf{0.20} & -0.97/-0.64 \\ . & 1 & -0.96/-0.37 & -0.95/-0.26 & -0.95/\mathbf{0.16} \\ . & . & 1 & 0.96/0.63 & 0.92/\mathbf{-0.26} \\ . & . & . & 1 & 0.94/\mathbf{-0.55} \\ . & . & . & . & 1 \end{pmatrix} \tag{B3}$$

For days with some missing observations (e.g. no PON measurements), the dimension of the vectors $H_i(\boldsymbol{x})$ and $\boldsymbol{y}_i$ and matrices $\boldsymbol{S}_{(\mathbf{y}_i)}$ and $\boldsymbol{C}_{(\mathbf{y})}$ have to be adjusted for that date accordingly. We disregard any prior information and the cost function (Eq. (13) in Sect. 2.3) reduces to:

$$J(\Theta) = \sum_{i=1}^{N_t} (\boldsymbol{y}_i - H_i(\boldsymbol{x}))^T \mathbf{R}_i^{-1} (\boldsymbol{y}_i - H_i(\boldsymbol{x})) \tag{B4}$$

For our second cost function we assume all data to be independent (i.e. all off-diagonals of $\boldsymbol{C}_{(\boldsymbol{y})}$ are zero) and Eq. (B4) can be further simplified to a sum over all individual vector components (indexed with j):

$$J(\Theta) = \sum_{i=1}^{N_t} \sum_{j=1}^{N_y} \frac{(y_{ij} - H_{ij}(\boldsymbol{x}))^2}{\sigma_{ij}^2} \tag{B5}$$

The mesocosm model environment was coded in FORTRAN and compiled as shared library so that we could use R as free software environment for statistical computations. For parameter optimisation (simulated anneadling) and for the analysis of the posterior (Markov chain Monte Carlo method) we applied the R package FME of Soetaert and Petzoldt (2010).

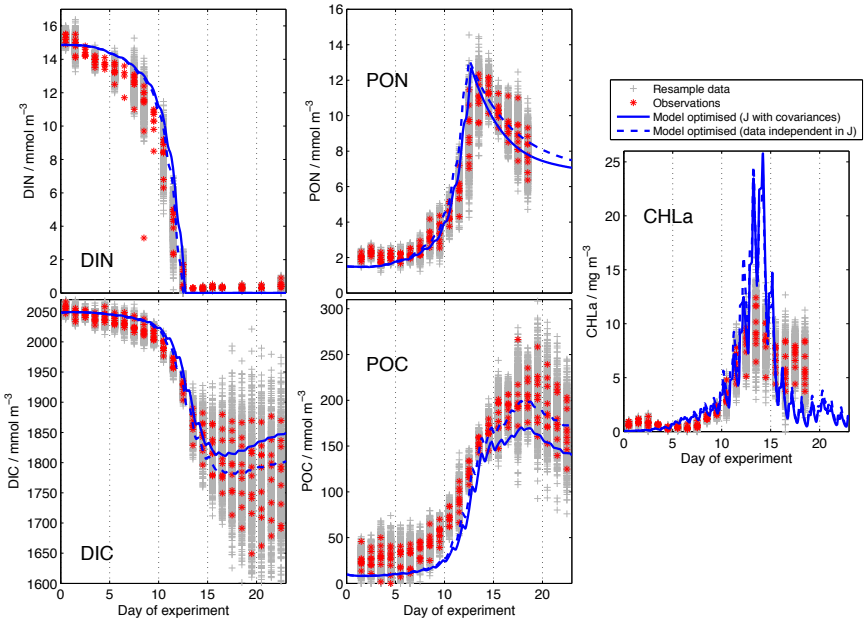

Figure B1. Observations of nine mesocosms (red asterisks), resampled data (gray markers) and optimised simulation results (blue lines): Dissolved inorganic nitrogen and carbon (DIN and DIC), particulate nitrogen and carbon (PON and POC), and chlorophyll *a* concentration (CHLa).

Appendix C: Development of an adjoint model (Sect. 5.3.3)

Adjoint models can be used to efficiently compute the derivative (or gradient) of the cost function J. In a parameter identification problem, J depends on Θ both indirectly via the state variable $\boldsymbol{x}$ and also directly if prior information is incorporated. The optimisation problem can thus be written as

$$\min_{\Theta} J(\boldsymbol{x}(\Theta), \Theta), \tag{C1}$$

where $\boldsymbol{x} = (\boldsymbol{x}_i)_{i=0}^{N_t}$ summarize all time instances of the model variables. To evaluate the derivative of the cost w.r.t. the parameters Θ, we may apply the chain rule and obtain

$$\frac{dJ}{d\Theta} = \sum_{i=0}^{N_t} \frac{\partial J}{\partial \boldsymbol{x}_i} \frac{d\boldsymbol{x}_i}{d\Theta} + \frac{\partial J}{\partial \Theta}, \tag{C2}$$

where we omitted the arguments $\boldsymbol{x}(\Theta)$ and Θ for brevity.

The needed derivatives of the model variables $\boldsymbol{x}_i$ w.r.t. the parameters Θ can be obtained by taking the total derivative w.r.t. Θ of the equations of the dynamical model, Eq. (1):

$$\frac{d\boldsymbol{x}_{i+1}}{d\Theta} = \frac{\partial M}{\partial \boldsymbol{x}_i} \frac{d\boldsymbol{x}_i}{d\Theta} + \frac{\partial M}{\partial \Theta}, \quad i = 0, \ldots, N_t - 1. \tag{C3}$$

This time propagation scheme for the derivatives is often called the tangent linear model.

The idea behind adjoint models is to avoid this direct computation, whose effort grows linear with the number of parameters Θ. For this purpose, we re-formulate Eq. (C1), treat both arguments of J independently and use the model equation as a constraint in the optimisation process. This can be expressed as

$$\min_{(\boldsymbol{x}, \Theta)} J(\boldsymbol{x}, \Theta) \text{ s. t. } \boldsymbol{x}_{i+1} = M\left[\boldsymbol{x}_i, \theta_e, \boldsymbol{f}\right], i = 0, \ldots, N_t - 1. \tag{C4}$$

A useful overview of adjoint model construction and applications is given in Kasibhatla (2000). An established approach to construct an adjoint model is to generate adjoint code directly from the numerical code of a model, based on algorithms that implement the chain rule for automatic differentiation (Griewank, 1989, 2003). According to the description of Giering and Kaminski (1998), a numerical model can be treated as a composition of differentiable functions, where each function represents a statement in the numerical code. The differentiation of such composition can be automated by highly sophisticated tools that yield tangent linear and adjoint FORTRAN code (e.g., Faure and Papegay, 1997; Giering and Kaminski, 1998). The application of adjoint construction tools (e.g., Tangent linear and Adjoint Model compiler, TAMC, of Giering and Kaminski, 1998) have been shown to perform well for studies with large-scale ocean general circulation models that include even complicated boundary conditions (e.g., Stammer et al., 1997; Marotzke et al., 1999; Wunsch and Heimbach, 2007; Heimbach et al., 2011).

Another approach is based on a discretised extended Lagrange equation. Under certain mathematical assumptions, a solution of Eq. (C4) corresponds to a saddle-point $(\boldsymbol{x}, \Theta, \boldsymbol{\lambda})$ of the Lagrangian

$$\mathcal{L}(\boldsymbol{x}, \Theta, \boldsymbol{\lambda}) = J(\boldsymbol{x}, \Theta) + \sum_{i=0}^{N_t - 1} \boldsymbol{\lambda}_i^{\top} \left(M\left[\boldsymbol{x}_i, \theta_e, \boldsymbol{f}\right] - \boldsymbol{x}_{i+1}\right). \tag{C5}$$

The vector $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_i)_{i=0}^{N_t-1}$ contains the Lagrange multipliers $\boldsymbol{\lambda}_i$, each of which corresponds to one time step in the model. A saddle-point of $\mathcal{L}$ satisfies the conditions

$$0 = \frac{\partial \mathcal{L}}{\partial \boldsymbol{x}_i} = \frac{\partial J}{\partial \boldsymbol{x}_i} + \boldsymbol{\lambda}_i^\top \frac{\partial M}{\partial \boldsymbol{x}_i} - \boldsymbol{\lambda}_{i-1}^\top, \qquad i = 1, \ldots, N_t \tag{C6}$$

$$0 = \frac{\partial \mathcal{L}}{\partial \Theta} = \frac{\partial J}{\partial \Theta} + \sum_{i=0}^{N_t-1} \boldsymbol{\lambda}_i^\top \frac{\partial M}{\partial \Theta} \tag{C7}$$

$$0 = \frac{\partial \mathcal{L}}{\partial \boldsymbol{\lambda}} \tag{C8}$$

Here, we again omitted the arguments, and set $\lambda_{N_t} = 0$ in the first equation to keep the compact notation. Note that all derivatives are partial ones since the idea is to decouple $\boldsymbol{x}$ and Θ and realize their dependency by implying the constraint in Eq. (C4). For simplicity we neglect additional parameter bounds which otherwise would affect Eq. (C7). Taking the derivative in Eq. (C8) for each $\boldsymbol{\lambda}_i$ separately results in the model equations (Eq. 1) again. From (C6) we deduce

$$\frac{\partial J}{\partial \boldsymbol{x}_i}\frac{d\boldsymbol{x}_i}{d\Theta} = \boldsymbol{\lambda}_{i-1}^\top \frac{d\boldsymbol{x}_i}{d\Theta} - \boldsymbol{\lambda}_i^\top \frac{\partial M}{\partial \boldsymbol{x}_i}\frac{d\boldsymbol{x}_i}{d\Theta}, \quad i = 1, \ldots, N_t$$

and apply Eq. (C3) to obtain

$$\frac{\partial J}{\partial \boldsymbol{x}_i}\frac{d\boldsymbol{x}_i}{d\Theta} = \boldsymbol{\lambda}_{i-1}^\top \frac{d\boldsymbol{x}_i}{d\Theta} - \boldsymbol{\lambda}_i^\top \left(\frac{d\boldsymbol{x}_{i+1}}{d\Theta} - \frac{\partial M}{\partial \Theta} \right), \; i = 1, \ldots, N_t$$

where $\lambda_{N_t} = 0$ as above. Summing up gives

$$\begin{aligned} \sum_{i=1}^{N_t} \frac{\partial J}{\partial \boldsymbol{x}_i}\frac{d\boldsymbol{x}_i}{d\Theta} &= \boldsymbol{\lambda}_0^\top \frac{d\boldsymbol{x}_1}{d\Theta} + \sum_{i=1}^{N_t} \boldsymbol{\lambda}_i^\top \frac{\partial M}{\partial \Theta} \\ &= \boldsymbol{\lambda}_0^\top \frac{\partial M}{\partial \boldsymbol{x}_0}\frac{d\boldsymbol{x}_0}{d\Theta} + \sum_{i=0}^{N_t} \boldsymbol{\lambda}_i^\top \frac{\partial M}{\partial \Theta} \end{aligned}$$

where we used again Eq. (C3) for $i = 1$. The first term includes the derivative of the initial values $\boldsymbol{x}_0$ w.r.t. the paraemters and in many cases will be zero. As result, the derivative of the cost can be computed from Eq. (C2) using the multiplier vector $\boldsymbol{\lambda}$, but without the tangent linear model. Note that the derivative of the model w.r.t. Θ in the sum is a *partial* derivative only, thus it does not include the derivaitive of the model variables, but only those of the model equations w.r.t. Θ.

The multipliers $\boldsymbol{\lambda}_i$ satsify a time-stepping scheme themselves, but in reverse direction. Using the transposed form of (C6), we obtain

$$\boldsymbol{\lambda}_{i-1} = \left(\frac{\partial M}{\partial \boldsymbol{x}_i}\right)^\top \boldsymbol{\lambda}_i + \left(\frac{\partial J}{\partial \boldsymbol{x}_i}\right)^\top, \quad i = N_t, \ldots, 1, \tag{C9}$$

with $\lambda_{N_t} = 0$ (see above) as starting point of the computation. Since here the transposed (or adjoint) of the linearisation of the model operator M occurs, these equations are referred to as the adjoint equations or the adjoint model. Accordingly, the multipliers $\boldsymbol{\lambda}$ are also referred to as adjoint variables or adjoints. Given a model trajectory $\boldsymbol{x}$ and using Eq. (C9), the trajectory of the adjoints $\boldsymbol{\lambda}$ can be computed. It is crucial to note that both time-stepping schemes, for the variables $\boldsymbol{x}$ and the adjoints $\boldsymbol{\lambda}$, have opposite directions. This requires – except for the case of a linear model M – the complete model trajectory to be stored or recomputed in order to compute $\boldsymbol{\lambda}$.

The adjoint model construction starting from a discretised extended Lagrange equation, Eq. (C5) can easily become extensive, in particular when discretisations of advection and mixing are included in the model dynamics. Furthermore, even small changes in the equations can entail considerable additional efforts in updating the adjoint model equations. The application of automatic differentiation tools may therefore be better suited for cases where the ecosystem dynamical model is subject to regular modifications.

References

Acevedo-Trejos, E., Brandt, G., Bruggeman, J., and Merico, A.: Mechanisms shaping size structure and functional diversity of phytoplankton communities in the ocean, Scientific Reports, 5, 2015.

Akaike, H.: Information theory and an extension of the maximum likelihood principle, in: Proceeding of the Second International Symposium on Information Theory, edited by Petrov, B. N. and Caski, F., pp. 267–281, Akademiai Kiado, 1973.

Aksnes, D. L. and Egge, J. K.: A theoretical model for nutrient uptake in phytoplankton., Marine Ecology Progress Series, 70, 65–72, 1991.

Allen, J., Eknes, M., and Evensen, G.: An Ensemble Kalman Filter with a complex marine ecosystem model: hindcasting phytoplankton in the Cretan Sea, in: Annales Geophysicae, vol. 21, pp. 399–411, 2003.

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

Annan, J., Hargreaves, J., Edwards, N., and Marsh, R.: Parameter estimation in an intermediate complexity earth system model using an ensemble Kalman filter, Ocean Modelling, 8, 135–154, 2005.

Anning, T., MacIntyre, H. L., Pratt, S. M., Sammes, P. J., Gibb, S., and Geider, R. J.: Photoacclimation in the marine diatom *Skeletonema costatum*, Limnology and Oceanography, 45, 1807–1817, 2000.

Arhonditsis, G. and Brett, M.: Evaluation of the current state of mechanistic aquatic biogeochemical modeling, Marine Ecology Progress Series, 271, 13–26, doi:10.3354/meps271013, 2004.

Arhonditsis, G. B., Papantou, D., Zhang, W., Perhar, G., Massos, E., and Shi, M.: Bayesian calibration of mechanistic aquatic biogeochemical models and benefits for environmental management, Journal of Marine Systems, 73, 8–30, 2008.

Armi, L. and Flament, P.: Cautionary remarks on the spectral interpretation of turbulent flows, Journal of Geophysical Research: Oceans (1978–2012), 90, 11 779–11 782, 1985.

Armstrong, R. A.: Optimality-based modeling of nitrogen allocation and photoacclimation in photosynthesis, Deep Sea Research Part II: Topical Studies in Oceanography, 53, 513–531, 2006.

Armstrong, R. A.: Nutrient uptake rate as a function of cell size and surface transporter density: A Michaelis-like approximation to the model of Pasciak and Gavis, Deep Sea Research Part I: Oceanographic Research Papers, 55, 1311–1317, 2008.

Arrhenius, S.: Über die Dissociationswärme und den Einfluss der Temperatur auf den Dissociationsgrad der Elektrolyte, Wilhelm Engelmann, 1889.

Askey, R. and Wilson, J. A.: Some basic hypergeometric orthogonal polynomials that generalize Jacobi polynomials, vol. 319, American Mathematical Society, 1985.

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

Bacastow, R. and Maier-Reimer, E.: Dissolved organic carbon in modeling oceanic new production, Global Biogeochemical Cycles, 5, 71–85, 1991.

Baird, M. E. and Suthers, I. M.: A size-resolved pelagic ecosystem model, Ecological Modelling, 203, 185–203, 2007.

Bakker, D. C. E., Pfeil, B., Landa, C. S., Metzl, N., O'Brien, K. M., Olsen, A., Smith, K., Cosca, C., Harasawa, S., Jones, S. D., Nakaoka, S.-I., Nojiri, Y., Schuster, U., Steinhoff, T., Sweeney, C., Takahashi, T., Tilbrook, B., Wada, C., Wanninkhof, R., Alin, S. R., Balestrini, C. F., Barbero, L., Bates, N. R., Bianchi, A. A., Bonou, F., Boutin, J., Bozec, Y., Burger, E. F., Cai, W.-J., Castle, R. D., Chen, L., Chierici, M.,

Currie, K., Evans, W., Featherstone, C., Feely, R. A., Fransson, A., Goyet, C., Greenwood, N., Gregor, L., Hankin, S., Hardman-Mountford, N. J., Harlay, J., Hauck, J., Hoppema, M., Humphreys, M. P., Hunt, C. W., Huss, B., Ibánhez, J. S. P., Johannessen, T., Keeling, R., Kitidis, V., Körtzinger, A., Kozyr, A., Krasakopoulou, E., Kuwata, A., Landschützer, P., Lauvset, S. K., Lefèvre, N., Lo Monaco, C., Manke, A., Mathis, J. T., Merlivat, L., Millero, F. J., Monteiro, P. M. S., Munro, D. R., Murata, A., Newberger, T., Omar, A. M., Ono, T., Paterson, K., Pearce, D., Pierrot, D., Robbins, L. L., Saito, S., Salisbury, J., Schlitzer, R., Schneider, B., Schweitzer, R., Sieger, R., Skjelvan, I., Sullivan, K. F., Sutherland, S. C., Sutton, A. J., Tadokoro, K., Telszewski, M., Tuma, M., van Heuven, S. M. A. C., Vandemark, D., Ward, B., Watson, A. J., and Xu, S.: A multi-decade record of high-quality fCO_2 data in version 3 of the Surface Ocean CO_2 Atlas (SOCAT), Earth System Science Data, 8, 383–413, doi:10.5194/essd-8-383-2016, http://www.earth-syst-sci-data.net/8/383/2016/, 2016.

Banas, N. S.: Adding complex trophic interactions to a size-spectral plankton model: Emergent diversity patterns and limits on predictability, Ecological Modelling, 222, 2663–2675, 2011.

Baretta-Bekker, J. G., Riemann, B., Baretta, J. W., and Rasmussen, E. K.: Testing the microbial loop concept by comparing mesocosm data with results from a dynamical simulation-model, Marine Ecology Progress Series, 106, 187–198, 1994.

Baretta-Bekker, J. G., Baretta, J. W., Hansen, A. S., and Riemann, B.: An improved model of carbon and nutrient dynamics in the microbial food web in marine enclosures, Aquatic Microbial Ecology, 1998.

Baumert, H.: On the theory of photosynthesis and growth in phytoplankton. Part I: Light limitation and constant temperature, Internationale Revue der gesamten Hydrobiologie und Hydrographie, 81, 109–139, 1996.

Bayes, T.: A letter from the late Reverend Mr. Thomas Bayes, F. R. S. to John Canton, A.M.F.R.S., Philosophical Transactions (1683-1775), pp. 269–271, 1763.

Bayes, T. and Price, R.: An Essay towards solving a Problem in the Doctrine of Chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S., Philosophical Transactions (1683-1775), pp. 370–418, 1763.

Béal, D., Brasseur, P., Brankart, J.-M., Ourmieres, Y., and Verron, J.: Characterization of mixing errors in a coupled physical biogeochemical model of the North Atlantic: implications for nonlinear estimation using Gaussian anamorphosis, Ocean Science, 6, 247–262, 2010.

Bennett, A. F.: Inverse modeling of the ocean and atmosphere, Cambridge University Press, 2002.

Berelson, W.: The flux of particulate organic carbon into the ocean interior: A comparison of four US JGOFS regional studies, Oceanography, 14, 59–67, 2001.

Bertino, L., Evensen, G., and Wackernagel, H.: Sequential data assimilation techniques in oceanography, International Statistical Review, 71, 223–241, 2003.

Blackman, F. F.: Optima and limiting factors, Annals of Botany, 19, 281–295, 1905.

Bliznyuk, N., Ruppert, D., Shoemaker, C., Regis, R., Wild, S., and Mugunthan, P.: Bayesian calibration and uncertainty analysis for computationally expensive models using optimization and radial basis function approximation, Journal of Computational and Graphical Statistics, 17, 2008.

Bocquet, M.: An introduction to inverse modelling and parameter estimation for atmosphere and ocean sciences, in: International Summer School - Advanced Data Assimilation for Geosciences, edited by Blayo, E., Bocquet, M., Cosme, E., and Cugliandolo, L. F., p. 608, Oxford University Press, 2014.

Box, G. E. and Cox, D. R.: An analysis of transformations, Journal of the Royal Statistical Society. Series B (Methodological), pp. 211–252, 1964.

Brasseur, P., Bahurel, P., Bertino, L., Birol, F., Brankart, J.-M., Ferry, N., Losa, S., Rémy, E., Schröter, J., Skachko, S., et al.: Data assimilation for marine monitoring and prediction: the MERCATOR operational assimilation systems and the MERSEA developments, Quarterly Journal of the Royal Meteorological Society, 131, 3561–3582, 2005.

Brown, J. H., Gillooly, J. F., Allen, A. P., Savage, V. M., and West, G. B.: Toward a metabolic theory of ecology, Ecology, 85, 1771–1789, 2004.

Bruggeman, J.: A phylogenetic approach to the estimation of phytoplankton traits, Journal of Phycology, 47, 52–65, 2011.

Bruggeman, J. and Kooijman, S. A. L. M.: A biodiversity-inspired approach to aquatic ecosystem modeling, Limnology and Oceanography, 52, 1533, 2007.

Bruggeman, J., Heringa, J., and Brandt, B. W.: PhyloPars: estimation of missing parameter values using phylogeny, Nucleic Acids Research, 37, W179–W184, 2009.

Brun, R., Reichert, P., and Künsch, H. R.: Practical identifiability analysis of large environmental simulation models, Water Resources Research, 37, 1015–1030, 2001.

Buesseler, K.: Do upper-ocean sediment traps provide an accurate record of particle flux?, Nature, 353, 420–423, 1991.

Buesseler, K., Lamborg, C., Boyd, P., Lam, P., Trull, T., Bidigare, R., Bishop, J., Casciotti, K., Dehairs, F., Elskens, M., Honda, M., Karl, D., Siegel, D., Silver, M., Steinberg, D., Valdes, J., Mooy, B. V., and Wilson, S.: Revisiting carbon flux through the ocean's twilight zone, Science, 316, 567–570, doi:10.1126/science.1137959, 2007.

Buitenhuis, E., Le Quéré, C., Aumont, O., Beaugrand, G., Bunker, A., Hirst, A., Ikeda, T., O'Brien, T., Piontkovski, S., and Straile, D.: Biogeochemical fluxes through mesozooplankton, Global Biogeochemical Cycles, 20, 2006.

Buitenhuis, E. T., Rivkin, R. B., Sailley, S., and Le Quéré, C.: Biogeochemical fluxes through microzooplankton, Global Biogeochemical Cycles, 24, 2010.

Burd, A. B. and Jackson, G. A.: Particle aggregation, Annual Review of Marine Science, 1, 65–90, 2009.

Burmaster, D. E.: The continuous culture of phytoplankton: mathematical equivalence among three steady-state models, American Naturalist, pp. 123–134, 1979.

Burnham, K. P. and Anderson, D. R.: Model selection and multimodel inference: a practical information-theoretic approach, Springer Science & Business Media, 2002.

Burnham, K. P. and Anderson, D. R.: Multimodel inference understanding AIC and BIC in model selection, Sociological Methods & Research, 33, 261–304, 2004.

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

Cao, X. and Spall, J.: Comparison of Expected and Observed Fisher Information in Variance Calculations for Parameter Estimates, Johns Hopkins APL technical digest, 28, 294, 2010.

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

Castelletti, A., Galelli, S., Ratto, M., Soncini-Sessa, R., and Young, P.: A general framework for dynamic emulation modelling in environmental problems, Environmental Modelling & Software, 34, 5–18, 2012.

Ciavatta, S., Torres, R., Saux-Picart, S., and Allen, J. I.: Can ocean color assimilation improve biogeochemical hindcasts in shelf seas?, Journal of Geophysical Research: Oceans, 116, 2011.

Cocco, V., Joos, F., Steinacher, M., Frölicher, T., Bopp, L., Dunne, J., Gehlen, M., Heinze, C., Orr, J., Oschlies, A., Schneider, B., Segschneider, J., and Tjiputra, J.: Oxygen and indicators of stress for marine life in multi-model global warming projections, Biogeosciences, 10, 1849–1868, doi:10.5194/bg-10-1849-2013, 2013.

Conn, A. R., Gould, N. I. M., and Toint, P. L.: Trust-region methods, MPS-SIAM Series on Optimization, Society for Industrial and Applied Mathematics, Philadelphia, 2000.

Conti, S. and O'Hagan, A.: Bayesian emulation of complex multi-output and dynamic computer models, Journal of Statistical Planning and Inference, 140, 640–651, 2010.

Conti, S., Gosling, J. P., Oakley, J. E., and O'hagan, A.: Gaussian process emulation of dynamic computer codes, Biometrika, p. asp028, 2009.

Côté, B. and Platt, T.: Day-to-day variations in the spring-summer photosynthetic parameters of coastal marine phytoplankton, Limnology and Oceanography, 28, 320–344, 1983.

Cox, D.: Frequentist and Bayesian statistics: a critique, Proceedings of the Statistical Problems in Particle Physics, Astrophysics and Cosmology, 2005.

Cox, D. R. and Hinkley, D. V.: Theoretical Statistics, CRC Press, 1974.

Craig, P., Goldstein, M., Seheult, A., and Smith, J.: Bayes linear strategies for matching hydrocarbon reservoir history, in: Bayesian statistics 5: Proceedings of the Fifth Valencia International Meeting, June 5-9, 1994, edited by Bernardo, J., Berger, J., Dawid, A., and Smith, A., pp. 69–95, Oxford University Press, 1996.

Crout, N. M., Tarsitano, D., and Wood, A. T.: Is my model too complex? Evaluating model formulation using model reduction, Environmental Modelling & Software, 24, 1–7, 2009.

Cullen, J. J., Yang, X., and MacIntyre, H. L.: Nutrient limitation of marine photosynthesis, in: Primary Productivity and Biogeochemical Cycles in the Sea, edited by Falkowski, P., Woodhead, A., and Vivirito, K., vol. 43 of *Environmental Science Research*, pp. 69–88, Springer, 1992.

Delille, B., Harlay, J., Zondervan, I., Jacquet, S., Chou, L., Wollast, R., Bellerby, R. G., Frankignoulle, M., Borges, A. V., Riebesell, U., and Gattuso, J.-P.: Response of primary production and calcification to changes of pCO_2 during experimental blooms of the coccolithophorid *Emiliania huxleyi*, Global Biogeochemical Cycles, 19, 2005.

Denman, K.: Modelling planktonic ecosystems: parameterizing complexity, Progress in Oceanography, 57, 429–452, doi:10.1016/S0079-6611(03)00109-5, 2003.

Doney, S. C., Glover, D. M., and Najjar, R. G.: A new coupled, one-dimensional biological-physical model for the upper ocean: Applications to the JGOFS Bermuda Atlantic Time-series Study (BATS) site, Deep Sea Research Part II: Topical Studies in Oceanography, 43, 591–624, 1996.

Doron, M., Brasseur, P., Brankart, J.-M., Losa, S. N., and Melet, A.: Stochastic estimation of biogeochemical parameters from Globcolour ocean colour satellite data in a North Atlantic 3D ocean coupled physical–biogeochemical model, Journal of Marine Systems, 117, 81–95, 2013.

Dowd, M.: A sequential Monte Carlo approach for marine ecological prediction, Environmetrics, 17, 435–455, 2006.

Dowd, M.: Bayesian statistical data assimilation for ecosystem models using Markov Chain Monte Carlo, Journal of Marine Systems, 68, 439–456, 2007.

Dowd, M.: Estimating parameters for a stochastic dynamic marine ecological system, Environmetrics, 22, 501–515, 2011.

Dowd, M. and Meyer, R.: A Bayesian approach to the ecosystem inverse problem, Ecological Modelling, 168, 39–55, doi:10.1016/S0304-3800(03)00186-8, 2003.

Dowd, M., Jones, E., and Parslow, J.: A statistical overview and perspectives on data assimilation for marine biogeochemical models, Environmetrics, 25, 203–213, 2014.

Droop, M.: 25 Years of Algal Growth Kinetics A Personal View, Botanica Marina, 26, 99–112, 1983.

Dutkiewicz, S., Follows, M. J., and Bragg, J. G.: Modeling the coupling of ocean ecology and biogeochemistry, Global Biogeochemical Cycles, 23, 2009.

Edwards, K. F., Thomas, M. K., Klausmeier, C. A., and Litchman, E.: Allometric scaling and taxonomic variation in nutrient utilization traits and maximum growth rate of phytoplankton, Limnology and Oceanography, 57, 554–566, 2012.

Efron, B.: Bootstrap confidence intervals for a class of parametric problems, Biometrika, 72, 45–58, 1985.

Efron, B.: Why isn't everyone a Bayesian?, The American Statistician, 40, 1–5, 1986.

Efron, B. and Hinkley, D. V.: Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information, Biometrika, 65, 457–483, 1978.

Efron, B. and Tibshirani, R.: Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy, Statistical science, pp. 54–75, 1986.

El Jarbi, M., Rückelt, J., Slawig, T., and Oschlies, A.: Reducing the model-data misfit in a marine ecosystem model using periodic parameters and linear quadratic optimal control, Biogeosciences, 10, 1169–1182, 2013.

Emerson, S.: Annual net community production and the biological carbon flux in the ocean, Global Biogeochemical Cycles, 28, 14–28, 2014.

Engel, A., Zondervan, I., Aerts, K., Beaufort, L., Benthien, A., Chou, L., Delille, B., Gattuso, J.-P., Harlay, J., Heemann, C., Hoffmann, L., Jacquet, S., Nejstgaard, J., Pizay, M.-D., Rochelle-Newall, E., Schneider, U., Terbrueggen, A., and Riebesell, U.: Testing the direct effect of CO_2 concentration on a bloom of the coccolithophorid *Emiliania huxleyi* in mesocosm experiments, Limnology and Oceanography, 50, 493–507, 2005.

Eppley, R. W.: Temperature and phytoplankton growth in the sea, Fishery Bulletin, 70, 1063–1085, 1972.

Eppley, R. W., Rogers, J. N., and McCarthy, J. J.: Half-saturation constants for uptake of nitrate and ammonium by marine phytoplankton, Limnology and Oceanography, 14, 912–920, 1969.

Evans, G. T.: A framework for discussing seasonal succession and coexistence of phytoplankton species, Limnology and Oceanography, 33, 1988.

Evans, G. T.: Defining misfit between biogeochemical models and data sets, Journal of Marine Systems, 40, 49–54, 2003.

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

Evensen, G.: The ensemble Kalman filter: Theoretical formulation and practical implementation, Ocean Dynamics, 53, 343–367, 2003.

Evensen, G.: The ensemble Kalman filter for combined state and parameter estimation, IEEE Control Systems, 29, 83–104, 2009.

Falkowski, P. G.: Nitrate uptake in marine phytoplankton: Comparison of half-saturation constants from seven species1, Limnology and Oceanography, 20, 412–417, 1975.

Fan, W. and Lv, X.: Data assimilation in a simple marine ecosystem model based on spatial biological parameterizations, Ecological Modelling, 220, 1997–2008, 2009.

Fasham, M., Ducklow, H., and McKelvie, S.: A nitrogen-based model of plankton dynamics in the oceanic mixed layer, Journal of Marine Research, 48, 591–639, 1990.

Fasham, M. J. R. and Evans, G. T.: The use of optimization techniques to model marine ecosystem dynamics at the JGOFS station 47 Deg N 20 Deg W, Philosophical Transactions of The Royal Society: Biological Sciences, 348, 203–209, 1995.

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

Faure, C. and Papegay, Y.: Odyssée Version 1.6, the language reference manual, Rapport Technique, 211, 1997.

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

Fiechter, J., Herbei, R., Leeds, W., Brown, J., Milliff, R., Wikle, C., Moore, A., and Powell, T.: A Bayesian parameter estimation method applied to a marine ecosystem model for the coastal Gulf of Alaska, Ecological Modelling, 258, 122–133, 2013.

Fiksen, Ø., Follows, M. J., and Aksnes, D. L.: Trait-based models of nutrient uptake in microbes extend the Michaelis-Menten framework, Limnology and Oceanography, 58, 193–202, 2013.

Fisher, R. A.: On the mathematical foundations of theoretical statistics, Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, pp. 309–368, 1922.

Fisher, R. A.: Two new properties of mathematical likelihood, Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character, 144, 285–307, 1934.

Fletcher, S.: Mixed Gaussian-lognormal four-dimensional data assimilation, Tellus A, 62, 266–287, 2010.

Flynn, K., Davidson, K., and Leftley, J.: Carbon-nitrogen relations at whole-cell and free-amino-acid levels during batch growth of *Isochrysis galbana* (Prymnesiophyceae) under conditions of alternating light and dark, Marine Biology, 118, 229–237, 1994.

Flynn, K. J.: Modelling multi-nutrient interactions in phytoplankton; balancing simplicity and realism, Progress in Oceanography, 56, 249–279, 2003.

Flynn, K. J.: Ecological modelling in a sea of variable stoichiometry: dysfunctionality and the legacy of Redfield and Monod, Progress in Oceanography, 84, 52–65, 2010.

Flynn, K. J., Marshall, H., and Geider, R. J.: A comparison of two N-irradiance interaction models of phytoplankton growth, Limnology and Oceanography, 46, 1794–1802, 2001.

Follows, M. J., Dutkiewicz, S., Grant, S., and Chisholm, S. W.: Emergent biogeography of microbial communities in a model ocean, science, 315, 1843–1846, 2007.

Franks, P. J.: NPZ models of plankton dynamics: their construction, coupling to physics, and application, Journal of Oceanography, 58, 379–387, 2002.

Franks, P. J.: Plankton patchiness, turbulent transport and spatial spectra, Marine Ecology Progress Series, 294, 295–309, 2005.

Franks, P. J. S.: Planktonic ecosystem models: perplexing parameterizations and a failure to fail, Journal of Plankton Research, 31, 1299–1306, doi:10.1093/plankt/fbp069, 2009.

Freeman, J. and Modarres, R.: Inverse Box–Cox: the power-normal distribution, Statistics & Probability Letters, 76, 764–772, 2006.

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, 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, 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.

Friedrichs, M. A. M., Dusenberry, J. A., Anderson, L. A., Armstrong, R. A., Chai, F., Christian, J. R., Doney, S. C., Dunne, J., Fujii, M., Hood, R., McGillicuddy, D. J., Moore, J. K., Schartau, M., Spitz, Y. H., and Wiggert, J. D.: Assessment of skill and portability in regional marine biogeochemical models: Role of multiple planktonic groups, Journal of Geophysical Research, 112, C08 001, doi:10.1029/2006JC003852, 2007.

Frigstad, H., Henson, S., Hartman, S., Omar, A., Jeansson, E., Cole, H., Pebody, C., and Lampitt, R.: Links between surface productivity and deep ocean particle flux at the Porcupine Abyssal Plain sustained observatory, Biogeosciences, 12, 5885–5897, 2015.

Fulton, E. A., Smith, A. D. M., and Johnson, C. R.: Effect of complexity on marine ecosystem models, Marine Ecology Progress Series, 253, 1–16, doi:10.3354/meps253001, 2003.

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, 233–252, 2003.

Gardner, W.: Sediment trap sampling in surface waters: issues and recommendations, in: The Changing Ocean Carbon Cycle, edited by *et al.*, R. H., pp. 240–284, Cambridge University Press, 2000.

Gehlen, M., Bopp, L., Emprin, N., Aumont, O., Heinze, C., and Ragueneau, O.: Reconciling surface ocean productivity, export fluxes and sediment composition in a global biogeochemical ocean model, Biogeosciences, 3, 521–537, 2006.

Geider, R. J., Maclntyre, H. L., and Kana, T. M.: A dynamic regulatory model of phytoplanktonic acclimation to light, nutrients, and temperature, Limnology and Oceanography, 43, 679–694, 1998.

Gentleman, W., Leising, A., Frost, B., Strom, S., and Murray, J.: Functional responses for zooplankton feeding on multiple resources: a review of assumptions and biological dynamics, Deep Sea Research Part II: Topical Studies in Oceanography, 50, 2847–2875, doi:10.1016/j.dsr2.2003.07.001, 2003.

Giering, R. and Kaminski, T.: Recipes for adjoint code construction, ACM Transactions on Mathematical Software (TOMS), 24, 437–474, 1998.

Gregg, W. W., a.M. Friedrichs, M., 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.

Griewank, A.: On automatic differentiation, Mathematical Programming: recent developments and applications, 6, 83–107, 1989.

Griewank, A.: A mathematical view of automatic differentiation, Acta Numerica, 12, 321–398, 2003.

Guieu, C., Dulac, F., Ridame, C., and Pondaven, P.: Introduction to project DUNE, a DUst experiment in a low Nutrient, low chlorophyll Ecosystem, Biogeosciences, 11, 425–442, 2014.

Gunson, J., Oschlies, A., and Garçon, V.: Sensitivity of ecosystem parameters to simulated satellite ocean color data using a coupled physical-biological model of the North Atlantic, Journal of Marine Research, 57, 613–639, doi:10.1357/002224099321549611, 1999.

Hald, A.: On the history of maximum likelihood in relation to inverse probability and least squares, Statistical Science, pp. 214–222, 1999.

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

Hastie, T., Tibshirani, R., and Friedman, J.: Unsupervised learning, in: The elements of statistical learning, pp. 485–585, Springer, 2009.

Hastings, W. K.: Monte Carlo sampling methods using Markov chains and their applications, Biometrika, 57, 97–109, 1970.

Healey, F. P.: Interacting effects of light and nutrient limitation on the growth rate of *Synechococcus linearis* (cyanophyceae) 1, Journal of Phycology, 21, 134–146, 1985.

Heath, M. R.: Ecosystem limits to food web fluxes and fisheries yields in the North Sea simulated with an end-to-end food web model, Progress in Oceanography, 102, 42–66, 2012.

Heimbach, P., Wunsch, C., Ponte, R. M., Forget, G., Hill, C., and Utke, J.: Timescales and regions of the sensitivity of Atlantic meridional volume and heat transport: Toward observing system design, Deep Sea Research Part II: Topical Studies in Oceanography, 58, 1858–1879, 2011.

Hemmings, J. C. 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.

Hemmings, J. C., Srokosz, M. A., Challenor, P., and Fasham, M. J.: Assimilating satellite ocean-colour observations into oceanic ecosystem models, Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, 361, 33–39, 2003.

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, 2004.

Hemmings, J. C., Challenor, P. G., and Yool, A.: Mechanistic site-based emulation of a global ocean biogeochemical model (MEDUSA 1.0) for parametric analysis and calibration: an application of the Marine Model Optimization Testbed (MarMOT 1.1), Geoscientific Model Development, 8, 697–731, doi:10.5194/gmd-8-697-2015, 2015.

Higdon, D., Gattiker, J., Williams, B., and Rightley, M.: Computer model calibration using high-dimensional output, Journal of the American Statistical Association, 103, 2008.

Hirata, T., Hardman-Mountford, N., Brewin, R., Aiken, J., Barlow, R., Suzuki, K., Isada, T., Howell, E., Hashioka, T., Noguchi-Aita, M., and Yamanaka, Y.: Synoptic relationships between surface Chlorophyll-*a* and diagnostic pigments specific to phytoplankton functional types, Biogeosciences, 8, 311–327, 2011.

Hoffman, R. N., Liu, Z., Louis, J.-F., and Grassoti, C.: Distortion representation of forecast errors, Monthly Weather Review, 123, 2758–2770, 1995.

Hooten, M. B., Leeds, W. B., Fiechter, J., and Wikle, C. K.: Assessing first-order emulator inference for physical parameters in nonlinear mechanistic models, Journal of Agricultural, Biological, and Environmental Statistics, 16, 475–494, 2011.

Huisman, J. and Weissing, F. J.: Biodiversity of plankton by species oscillations and chaos, Nature, 402, 407–410, 1999.

Huret, M., Gohin, F., Delmas, D., Lunven, M., and Garçon, V.: Use of SeaWiFS data for light availability and parameter estimation of a phytoplankton production model of the Bay of Biscay, Journal of Marine Systems, 65, 509–531, 2007.

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, 653–683, 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 Papers, 46, 27–61, 1999.

Ilyina, T., Six, K. D., Segschneider, J., Maier-Reimer, E., Li, H., and Núñez-Riboni, I.: Global ocean biogeochemistry model HAMOCC: Model architecture and performance as component of the MPI-Earth System Model in different CMIP5 experimental realizations, Journal of Advances in Modeling Earth Systems, 5, 287–315, 2013.

IPCC: Synthesis report, Contributions of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, RK Pachauri and LA Meyer (eds.)]. IPCC, Geneva, Switzerland, 2014.

Jackson, G. A.: A model of the formation of marine algal flocs by physical coagulation processes, Deep Sea Research Part A. Oceanographic Research Papers, 37, 1197–1211, 1990.

Jassby, A. D. and Platt, T.: Mathematical formulation of the relationship between photosynthesis and light for phytoplankton, Limnology and Oceanography, 21, 540–547, 1976.

Jazwinski, A. H.: Stochastic processes and filtering theory, Courier Corporation, 2007.

Joassin, P., Delille, B., Soetaert, K., Harlay, J., Borges, A. V., Chou, L., Riebesell, U., Suykens, K., and Grégoire, M.: Carbon and nitrogen flows during a bloom of the coccolithophore *Emiliania huxleyi*: Modelling a mesocosm experiment, Journal of Marine Systems, 85, 71–85, 2011.

Johnson, J. B. and Omland, K. S.: Model selection in ecology and evolution, Trends in Ecology & Evolution, 19, 101–108, 2004.

Jones, E., Parslow, J., and Murray, L.: A Bayesian approach to state and parameter estimation in a Phytoplankton-Zooplankton model, Australian Meteorological and Oceanographic Journal, 59, 7–16, 2010.

Kane, A., Moulin, C., Thiria, S., Bopp, L., Berrada, 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, 2011.

Kasibhatla, P.: Inverse methods in global biogeochemical cycles, 114, American Geophysical Union, 2000.

Kavetski, D., Kuczera, G., and Franks, S. W.: Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory, Water Resources Research, 42, 2006.

Kennedy, M. C. and O'Hagan, A.: Predicting the output from a complex computer code when fast approximations are available, Biometrika, 87, 1–13, 2000.

Kennedy, M. C. and O'Hagan, A.: Bayesian calibration of computer models, Journal of the Royal Statistical Society. Series B, Statistical Methodology, pp. 425–464, 2001.

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

Khatiwala, S.: Fast spin up of ocean biogeochemical models using matrix-free Newton–Krylov, Ocean Modelling, 23, 121–129, 2008.

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, 2011.

Klausmeier, C. A. and Litchman, E.: Algal games: The vertical distribution of phytoplankton in poorly mixed water columns, Limnology and Oceanography, 46, 1998–2007, 2001.

Kooijman, S.: Population dynamics on basis of budgets, in: The dynamics of physiologically structured populations, vol. 68, pp. 266–297, Springer Berlin, 1986.

Kreus, M. and Schartau, M.: Variations in the elemental ratio of organic matter in the central Baltic Sea: Part II–Sensitivities of annual mass flux estimates to model parameter variations, Continental Shelf Research, 100, 46–63, 2015.

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

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

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

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., Oschlies, A., and Khatiwala, S.: Sensitivity analysis of simple global marine biogeochemical models, Global Biogeochemical Cycles, 26, 2012.

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, in press, 1–28, doi:10.5194/gmd-10-1-2017, http://www.geosci-model-dev.net/10/1/2017/, 2017.

Kuczera, G.: Assessing hydrologic model nonlinearity using response surface plots, Journal of Hydrology, 118, 143–161, 1990.

Kuhn, A. M., Fennel, K., and Mattern, J. P.: Model investigations of the North Atlantic spring bloom initiation, Progress in Oceanography, 138, 176–193, 2015.

Kwon, E. Y. and Primeau, F.: Optimization and sensitivity study of a biogeochemistry ocean model using an implicit solver and in situ phosphate data, Global Biogeochemical Cycles, 20, 2006.

Kwon, E. Y. and Primeau, F.: Optimization and sensitivity of a global biogeochemistry ocean model using combined in situ DIC, alkalinity, and phosphate data, Journal of Geophysical Research: Oceans, 113, 2008.

Kwon, E. Y., Primeau, F., and Sarmiento, J. L.: The impact of remineralization depth on the air–sea carbon balance, Nature Geoscience, 2, 630–635, 2009.

Laws, E. A. and Bannister, T.: Nutrient-and light-limited growth of *Thalassiosira fluviatilis* in continuous culture, with implications for phytoplankton growth in the ocean, Limnology and Oceanography, 25, 457–473, 1980.

Laws, E. A., Redalje, D. G., Karl, D. M., and Chalup, M. S.: A theoretical and experimental examination of the predictions of two recent models of phytoplankton growth, Journal of theoretical biology, 105, 469–491, 1983.

Lawson, L. M., Spitz, Y. H., Hofmann, E. E., and Long, R. B.: A data assimilation technique applied to a predator-prey model, Bulletin of Mathematical Biology, 57, 593–617, 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.

Lawson, W. G. and Hansen, J. A.: Alignment error models and ensemble-based data assimilation, Monthly Weather Review, 133, 1687–1709, 2005.

Le Dimet, F.-X. and Talagrand, O.: Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects, Tellus A, 38, 97–110, 1986.

Le Queré, C.: Reply to Horizons Article 'Plankton functional type modelling: running before we can walk'Anderson (2005): I. Abrupt changes in marine ecosystems?, Journal of Plankton Research, 28, 871–872, 2006.

Leeds, W., Wikle, C., and Fiechter, J.: Emulator-assisted reduced-rank ecological data assimilation for nonlinear multivariate dynamical spatio-temporal processes, Statistical Methodology, 17, 126–138, 2014.

Leeds, W. B., Wikle, C. K., Fiechter, J., Brown, J., and Milliff, R. F.: Modeling 3-D spatio-temporal biogeochemical processes with a forest of 1-D statistical emulators, Environmetrics, 24, 1–12, 2013.

Lele, S. R. and Dennis, B.: Bayesian methods for hierarchical models: are ecologists making a Faustian bargain, Ecological Applications, 19, 581–584, 2009.

Lewis, F., Butler, A., and Gilbert, L.: A unified approach to model selection using the likelihood ratio test, Methods in Ecology and Evolution, 2, 155–162, doi:10.1111/j.2041-210X.2010.00063.x, 2011.

Li, X. and Primeau, F. W.: A fast Newton–Krylov solver for seasonally varying global ocean biogeochemistry models, Ocean Modelling, 23, 13–20, doi:http://dx.doi.org/10.1016/j.ocemod.2008.03.001, http://www.sciencedirect.com/science/article/pii/S1463500308000371, 2008.

Li, X., McGillicuddy, D. J., Durbin, E. G., and Wiebe, P. H.: Biological control of the vernal population increase of *Calanus finmarchicus* on Georges Bank, Deep Sea Research Part II: Topical Studies in Oceanography, 53, 2632–2655, 2006.

Li, X., Wang, C., Fan, W., and Lv, X.: Optimization of the Spatiotemporal Parameters in a Dynamical Marine Ecosystem Model Based on the Adjoint Assimilation, Mathematical Problems in Engineering, 2013, 2013.

Lignell, R., Haario, H., Laine, M., and Thingstad, T. F.: Getting the "right" parameter values for models of the pelagic microbial food web, Limnology and Oceanography, 58, 301–313, 2013.

Link, W. A. and Barker, R. J.: Model weights and the foundations of multimodel inference, Ecology, 87, 2626–2635, 2006.

Litchman, E., Klausmeier, C. A., Schofield, O. M., and Falkowski, P. G.: The role of functional traits and trade-offs in structuring phytoplankton communities: scaling from cellular to ecosystem level, Ecology Letters, 10, 1170–1181, 2007.

Litchman, E., Edwards, K. F., Klausmeier, C. A., and Thomas, M. K.: Phytoplankton niches, traits and eco-evolutionary responses to global environmental change, Marine Ecology Progress Series, 470, 235–248, 2012.

Liu, F. and West, M.: A dynamic modelling strategy for Bayesian computer model emulation, Bayesian Analysis, 4, 393–411, 2009.

Longhurst, A.: Seasonal cycles of pelagic production and consumption, Progress in Oceanography, 36, 77–167, 1995.

Longhurst, A. R.: Ecological Geography of the Sea, San Diego, Academic Press, 1998.

Löptien, U. and Dietze, H.: Constraining parameters in state-of-the-art 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.

Löptien, U. and Meier, H. M.: The influence of increasing water turbidity on the sea surface temperature in the Baltic Sea: A model sensitivity study, Journal of Marine Systems, 88, 323–331, 2011.

Löptien, U., Eden, C., Timmermann, A., and Dietze, H.: Effects of biologically induced differential heating in an eddy-permitting coupled ocean-ecosystem model, Journal of Geophysical Research: Oceans (1978–2012), 114, 2009.

Losa, S. N., Kivman, G. A., Schröter, J., and Wenzel, M.: Sequential weak constraint parameter estimation in an ecosystem model, Journal of Marine Systems, 43, 31–49, 2003.

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

Losa, S. N., Vézina, A., Wright, D., Lu, Y., Thompson, K., and Dowd, M.: 3D ecosystem modelling in the North Atlantic: Relative impacts of physical and biological parameterizations, Journal of marine systems, 61, 230–245, 2006.

Lucia, D. J., Beran, P. S., and Silva, W. A.: Reduced-order modeling: new approaches for computational physics, Progress in Aerospace Sciences, 40, 51–117, 2004.

Maier-Reimer, E.: Geochemical cycles in an ocean general circulation model. Preindustrial tracer distributions, Global Biogeochemical Cycles, 7, 645–677, 1993.

Malve, O., Laine, M., Haario, H., Kirkkala, T., and Sarvala, J.: Bayesian modelling of algal mass occurrences—using adaptive MCMC methods with a lake water quality model, Environmental Modelling & Software, 22, 966–977, 2007.

Marotzke, J., Giering, R., Zhang, K. Q., Stammer, D., Hill, C., and Lee, T.: Construction of the adjoint MIT ocean general circulation model and application to Atlantic heat transport sensitivity, Journal of Geophysical Research, 104, 529–548, 1999.

Marsili-Libelli, S., Guerrizio, S., and Checchi, N.: Confidence regions of estimated parameters for ecological systems, Ecological Modelling, 165, 127–146, 2003.

Martin, A.: Phytoplankton patchiness: the role of lateral stirring and mixing, Progress in Oceanography, 57, 125–174, 2003.

Martin, J. H., Knauer, G. A., Karl, D. M., and Broenkow, W. W.: VERTEX: carbon cycling in the northeast Pacific, Deep Sea Research Part A. Oceanographic Research Papers, 34, 267–285, 1987.

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

Matear, R. J. and Jones, E.: Marine biogeochemical modelling and data assimilation, in: Operational Oceanography in the 21st Century, pp. 295–317, Springer, 2011.

Mattern, J. P. and Edwards, C. A.: Simple parameter estimation for complex models—Testing evolutionary techniques on 3-dimensional biogeochemical ocean models, Journal of Marine Systems, 165, 139–152, 2017.

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, 32–47, 2012.

Mattern, J. P., Dowd, M., and Fennel, K.: Particle filter-based data assimilation for a three-dimensional biological ocean model and satellite observations, Journal of Geophysical Research: Oceans, 118, 2746–2760, 2013a.

Mattern, J. P., Fennel, K., and Dowd, M.: Sensitivity and uncertainty analysis of model hypoxia estimates for the Texas-Louisiana shelf, Journal of Geophysical Research: Oceans, 118, 1316–1332, 2013b.

Mattern, J. P., Fennel, K., and Dowd, M.: Periodic time-dependent parameters improving forecasting abilities of biological ocean models, Geophysical Research Letters, 41, 6848–6854, 2014.

McDonald, C. P. and Urban, N. R.: Using a model selection criterion to identify appropriate complexity in aquatic biogeochemical models, Ecological Modelling, 221, 428–432, 2010.

Meeker, W. Q. and Escobar, L. A.: Teaching about approximate confidence regions based on maximum likelihood estimation, The American Statistician, 49, 48–53, 1995.

Melbourne-Thomas, J., Wotherspoon, S., Corney, S., Molina-Balari, E., Marini, O., and Constable, A.: Optimal control and system limitation in a Southern Ocean ecosystem model, Deep Sea Research Part II: Topical Studies in Oceanography, 2013.

Merico, A., Bruggeman, J., and Wirtz, K.: A trait-based approach for downscaling complexity in plankton ecosystem models, Ecological Modelling, 220, 3001–3010, 2009.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E.: Equation of state calculations by fast computing machines, The Journal of Chemical Physics, 21, 1087–1092, 1953.

Mignot, A., Claustre, H., Uitz, J., Poteau, A., D'Ortenzio, F., and Xing, X.: Understanding the seasonal dynamics of phytoplankton biomass and the deep chlorophyll maximum in oligotrophic environments: A Bio-Argo float investigation, Global Biogeochemical Cycles, 28, 856–876, 2014.

Mittermaier, M. P.: Improving short-range high-resolution model precipitation forecast skill using time-lagged ensembles, Quarterly Journal of the Royal Meteorological Society, 133, 1487–1500, 2007.

Monod, J.: Recherches sur la croissance des cultures bacteriennes, Ph. D Thesis, 1942.

Monod, J.: The growth of bacterial cultures, Selected Papers in Molecular Biology by Jacques Monod, p. 139, 2012.

Murtugudde, R., Beauchamp, J., McClain, C. R., Lewis, M., and Busalacchi, A. J.: Effects of penetrative radiation on the upper tropical ocean circulation, Journal of Climate, 15, 470–486, 2002.

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

Natvik, L.-J. and Evensen, G.: Assimilation of ocean colour data into a biochemical model of the North Atlantic: Part 1. Data assimilation experiments, Journal of Marine Systems, 40, 127–153, 2003.

Nerger, L. and Gregg, W. W.: Improving assimilation of SeaWiFS data by the application of bias correction with a local SEIK filter, Journal of marine systems, 73, 87–102, 2008.

Nevison, C., Manizza, M., Keeling, R., Kahru, M., Bopp, L., Dunne, J., Tiputra, J., Ilyina, T., and Mitchell, B.: Evaluating the ocean biogeochemical components of Earth system models using atmospheric potential oxygen and ocean color data, Biogeosciences, 12, 193–208, doi:10.5194/bg-12-193-2015, 2015.

Omlin, M. and Reichert, P.: A comparison of techniques for the estimation of model prediction uncertainty, Ecological Modelling, 115, 45–59, 1999.

Omlin, M., Reichert, P., and Forster, R.: Biogeochemical model of Lake Zürich: model equations and results, Ecological Modelling, 141, 77–103, 2001.

Oschlies, A.: Feedbacks of biotically induced radiative heating on upper-ocean heat budget, circulation, and biological production in a coupled ecosystem-circulation model, Journal of Geophysical Research: Oceans, 109, 2004.

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

Oschlies, A. and Schartau, M.: Basin-scale performance of a locally optimized marine ecosystem model, Journal of Marine Research, 63, 335–358, 2005.

O'Hagan, A.: Bayesian analysis of computer code outputs: a tutorial, Reliability Engineering & System Safety, 91, 1290–1300, 2006.

Pahlow, M.: Linking chlorophyll-nutrient dynamics to the Redfield N:C ratio with a model of optimal phytoplankton growth, Marine Ecology Progress Series, 287, 33–43, 2005.

Pahlow, M. and Oschlies, A.: Chain model of phytoplankton P, N and light colimitation, Marine Ecology Progress Series, 376, 2009.

Pahlow, M. and Oschlies, A.: Optimal allocation backs Droop's cell-quota model, Marine Ecology Progress Series, 473, 1–5, 2013.

Pahlow, M., Vézina, A. F., Casault, B., Maass, H., Malloch, L., Wright, D. G., and Lu, Y.: Adaptive model of plankton dynamics for the North Atlantic, Progress in Oceanography, 76, 151–191, 2008.

Parekh, P., Follows, M. J., and Boyle, E. A.: Decoupling of iron and phosphate in the global ocean, Global Biogeochemical Cycles, 19, 2005.

Parslow, J., Cressie, N., Campbell, E. P., Jones, E., and Murray, L.: Bayesian learning and predictability in a stochastic nonlinear dynamical model, Ecological Applications, 23, 679–698, 2013.

Pelc, J. S., Simon, E., Bertino, L., El Serafy, G., and Heemink, A. W.: Application of model reduced 4D-Var to a 1D ecosystem model, Ocean Modelling, 57, 43–58, 2012.

Peterson, D., Perry, M., Bencala, K., and Talbot, M.: Phytoplankton productivity in relation to light intensity: a simple equation, Estuarine, Coastal and Shelf Science, 24, 813–832, 1987.

Phillips, J. R.: Projection-based approaches for model reduction of weakly nonlinear, time-varying systems, Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on, 22, 171–187, 2003.

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, http://www.geosci-model-dev.net/9/3729/2016/, 2016.

Platt, T. and Jassby, A. D.: The relationship between photosynthesis and light for natural assemblages of coastal marine phytoplankton, Journal of Phycology, 12, 421–430, 1976.

Platt, T., Caverhill, C., and Sathyendranath, S.: Basin-scale estimates of oceanic primary production by remote sensing: The North Atlantic, Journal of Geophysical Research: Oceans, 96, 15 147–15 159, 1991.

Platt, T., Sathyendranath, S., Ulloa, O., and Harrison, W.: Nutrient control of phytoplankton photosynthesis in the Western North Atlantic, Nature, 356, 229–231, 1992.

Powell, T. M., Lewis, C. V., Curchitser, E. N., Haidvogel, D. B., Hermann, A. J., and Dobbins, E. L.: Results from a three-dimensional, nested biological-physical model of the California Current System and comparisons with statistics from satellite imagery, Journal of Geophysical Research: Oceans, 111, 356–362, doi:10.1029/2004JC002506, 2006.

Prieß, M., Koziel, S., and Slawig, T.: Marine ecosystem model calibration with real data using enhanced surrogate-based optimization, Journal of Computational Science, 4, 423–437, 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, 2013b.

Primeau, F. and Deleersnijder, E.: On the time to tracer equilibrium in the global ocean, Ocean Science, 5, 13–28, 2009.

Prunet, P., Minster, J.-F., Echevin, V., and Dadou, I.: Assimilation of surface data in a one-dimensional physical-biogeochemical model of the surface ocean: 2. Adjusting a simple trophic model to chlorophyll, temperature, nitrate, and pCO2 data, Global Biogeochemical Cycles, 10, 139–158, 1996.

Raue, A., Kreutz, C., Maiwald, T., Bachmann, J., Schilling, M., Klingmüller, U., and Timmer, J.: Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood, Bioinformatics, 25, 1923–1929, 2009.

Raue, A., Kreutz, C., Maiwald, T., Klingmuller, U., and Timmer, J.: Addressing parameter identifiability by model-based experimentation, Systems Biology, IET, 5, 120–130, 2011.

Ravela, S., Emanuel, K., and McLaughlin, D.: Data assimilation by field alignment, Physica D: Nonlinear Phenomena, 230, 127–145, 2007.

Rayner, P., Michalak, A. M., and Chevallier, F.: Fundamentals of Data Assimilation, Geoscientific Model Development Discussions, doi:doi:10.5194/gmd-2016-148, 2016.

Reed, D. C., Algar, C. K., Huber, J. A., and Dick, G. J.: Gene-centric approach to integrating environmental genomics and biogeochemical models, Proceedings of the National Academy of Sciences, 111, 1879–1884, 2014.

Ridgwell, A., Hargreaves, J., Edwards, N. R., Annan, J., Lenton, T. M., Marsh, R., Yool, A., and Watson, A.: Marine geochemical data assimilation in an efficient Earth System Model of global biogeochemical cycling, Biogeosciences, 4, 87–104, 2007.

Riebesell, U., Bellerby, R., Grossart, H.-P., and Thingstad, F.: Mesocosm CO_2 perturbation studies: from organism to community level, Biogeosciences, 5, 1157–1164, 2008.

Robinson, A. R. and Lermusiaux, P. F. J.: Data assimilation for modeling and predicting coupled physical–biological interactions in the sea, in: The Sea, edited by Robinson, A. R., J., M. J., and J., R. B., vol. 12, pp. 475–536, John Wiley & Sons, Inc New York, 2002.

Rödenbeck, C., Bakker, D. C. E., Gruber, N., Iida, Y., Jacobson, A. R., Jones, S., Landschützer, P., Metzl, N., Nakaoka, S., Olsen, A., Park, G.-H., Peylin, P., Rodgers, K. B., Sasse, T. P., Schuster, U., Shutler, J. D., Valsala, V., Wanninkhof, R., and Zeng, J.: Data-based estimates of the ocean carbon sink variability – first results of the Surface Ocean pCO_2 Mapping intercomparison (SOCOM), Biogeosciences, 12, 7251–7278, doi:10.5194/bg-12-7251-2015, http://www.biogeosciences.net/12/7251/2015/, 2015.

Roy, S., Broomhead, D. S., Platt, T., Sathyendranath, S., and Ciavatta, S.: Sequential variations of phytoplankton growth and mortality in an NPZ model: A remote-sensing-based assessment, Journal of Marine Systems, 92, 16–29, 2012.

Rückelt, J., Sauerland, V., Slawig, T., Srivastav, A., Ward, B., and Patvardhan, C.: Parameter optimization and uncertainty analysis in a model of oceanic CO_2 uptake using a hybrid algorithm and algorithmic differentiation, Nonlinear Analysis: Real World Applications, 11, 3993–4009, 2010.

Ruiz, J., Prieto, L., and Ortegón, F.: Diatom aggregate formation and fluxes: a modeling analysis under different size-resolution schemes and with empirically determined aggregation kernels, Deep Sea Research Part I: Oceanographic Research Papers, 49, 495–515, 2002.

Sarmiento, J. L., Slater, R. D., Fasham, M. J. R., Ducklow, H. W., Toggweiler, J. R., and Evans, G. T.: A seasonal three-dimensional ecosystem model of nitrogen cycling in the North Atlantic Euphotic Zone, Global Biogeochemical Cycles, 7, 417–450, 1993.

Sasaki, Y.: Some basic formalisms in numerical variational analysis, Monthly Weather Review, 98, 875–883, 1970.

Sauzède, R., Claustre, H., Jamet, C., Uitz, J., Ras, J., Mignot, A., and D'Ortenzio, F.: Retrieving the vertical distribution of chlorophyll a concentration and phytoplankton community composition from in situ fluorescence profiles: A method based on a neural network with potential for global-scale applications, Journal of Geophysical Research: Oceans, 120, 451–470, 2015a.

Sauzède, R., Lavigne, H., Claustre, H., Uitz, J., Schmechtig, C., D'Ortenzio, F., Guinet, C., and Pesant, S.: Vertical distribution of chlorophyll *a* concentration and phytoplankton community composition from in situ fluorescence profiles: a first database for the global ocean, Earth System Science Data, 7, 261–273, 2015b.

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, 2003.

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

Schartau, M., Engel, A., Schröter, J., Thoms, S., Völker, C., and Wolf-Gladrow, D.: Modelling carbon overconsumption and the formation of extracellular particulate organic carbon, Biogeosciences, 4, 433–454, 2007.

Schwinger, J., Goris, N., Tjiputra, J., Kriest, I., Bentsen, M., Bethke, I., Ilicak, M., Assmann, K., and Heinze, C.: Evaluation of NorESM-OC (versions 1 and 1.2), the ocean carbon-cycle stand-alone configuration of the Norwegian Earth System Model (NorESM1), Geoscientific Model Development Discussions, pp. 1–73, 2016.

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

Shuter, B.: A model of physiological adaptation in unicellular algae, Journal of Theoretical Biology, 78, 519–552, 1979.

Siberlin, C. and Wunsch, C.: Oceanic tracer and proxy time scales revisited, Climate of the Past, 7, 27–39, 2011.

Siegel, D. A., Fields, E., and Buesseler, K. O.: A bottom-up view of the biological pump: Modeling source funnels above ocean sediment traps, Deep Sea Research Part I: Oceanographic Research Papers, 55, 108–127, 2008.

Simon, E. and Bertino, L.: Application of the Gaussian anamorphosis to assimilation in a 3-D coupled physical-ecosystem model of the North Atlantic with the EnKF: a twin experiment, Ocean Science, 5, 495–510, 2009.

Simon, E. and Bertino, L.: Gaussian anamorphosis extension of the DEnKF for combined state parameter estimation: application to a 1D ocean ecosystem model, Journal of Marine Systems, 89, 1–18, 2012.

Simon, E., Samuelsen, A., Bertino, L., and Mouysset, S.: Experiences in multiyear combined state–parameter estimation with an ecosystem model of the North Atlantic and Arctic Oceans using the Ensemble Kalman Filter, Journal of Marine Systems, 152, 1–17, 2015.

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

Slawig, T., Prieß, M., and Kratzenstein, C.: Surrogate-Based and One-Shot Optimization Methods for PDE-Constrained Problems with an Application in Climate Models, in: Solving Computationally Expensive Engineering Problems, edited by Koziel, S., Leifsson, L., and Yang, X.-S., vol. 97 of *Springer Proceedings in Mathematics & Statistics*, pp. 1–24, Springer International Publishing, 2014.

Smith, E. L.: Photosynthesis in relation to light and carbon dioxide, Proceedings of the National Academy of Sciences of the United States of America, 22, 504, 1936.

Smith, R. A.: The theoretical basis for estimating phytoplankton production and specific growth rate from chlorophyll, light and temperature data, Ecological Modelling, 10, 243–264, 1980.

Smith, S. L. and Yamanaka, Y.: Quantitative comparison of photoacclimation models for marine phytoplankton, Ecological Modelling, 201, 547–552, 2007a.

Smith, S. L., Yamanaka, Y., Pahlow, M., and Oschlies, A.: Optimal uptake kinetics: physiological acclimation explains the pattern of nitrate uptake by phytoplankton in the ocean, Marine Ecology Progress Series, 384, 1–12, 2009.

Smith, S. L., Merico, A., Wirtz, K. W., and Pahlow, M.: Leaving misleading legacies behind in plankton ecosystem modelling, Journal of Plankton Research, 36, 613–620, 2014.

Smith, S. L., Pahlow, M., Merico, A., Acevedo-Trejos, E., Sasai, Y., Yoshikawa, C., Sasaoka, K., Fujiki, T., Matsumoto, K., and Honda, M. C.: Flexible phytoplankton functional type (FlexPFT) model: size-scaling of traits and optimal growth, Journal of Plankton Research, p. fbv038, 2015.

Soetaert, K. and Petzoldt, T.: Inverse Modelling, Sensitivity and Monte Carlo Analysis in R Using Package FME, Journal of Statistical Software, 33, 2010.

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

Stammer, D., Wunsch, C., Giering, R., Zhang, Q., Marotzke, J., Marshall, J., and Hill, C.: The global ocean circulation estimated from TOPEX/POSEIDON altimetry and the MIT general circulation model, MIT Center for Global Change Science Report, No. 49, 1997.

Steinacher, M. and Joos, F.: Transient Earth system responses to cumulative carbon dioxide emissions: linearities, uncertainties, and probabilities in an observation-constrained model ensemble, Biogeosciences, 13, 1071–1103, 2016.

Steinacher, M., Joos, F., and Stocker, T. F.: Allowable carbon emissions lowered by multiple climate targets, Nature, 499, 197–201, 2013.

Stock, C. A., McGillicuddy, D. J., Solow, A. R., and Anderson, D. M.: Evaluating hypotheses for the initiation and development of *Alexandrium fundyense* blooms in the western Gulf of Maine using a coupled physical–biological model, Deep Sea Research Part II: Topical Studies in Oceanography, 52, 2715–2744, 2005.

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

Tarantola, A.: Inverse problems theory, Methods for Data Fitting and Model Parameter Estimation. Elsevier, Southampton, 1987.

Tarantola, A.: Inverse problem theory and methods for model parameter estimation, siam, 2005.

Taylor, K. E., Stouffer, R. J., and Meehl, G. A.: An overview of CMIP5 and the experiment design, Bulletin of the American Meteorological Society, 93, 485–498, 2012.

Terry, K. L., Hirata, J., and Laws, E. A.: Light-limited growth of two strains of the marine diatom *Phaeodactylum tricornutum Bohlin*: chemical composition, carbon partitioning and the diel periodicity of physiological processes, Journal of Experimental Marine Biology and Ecology, 68, 209–227, 1983.

Terry, K. L., Hirata, J., and Laws, E. A.: Light-, nitrogen-, and phosphorus-limited growth of *Phaeodactylum tricornutum Bohlin* strain TFX-1: Chemical composition, carbon partitioning, and the diel periodicity of physiological processes, Journal of Experimental Marine Biology and Ecology, 86, 85–100, 1985.

Thacker, W. C.: The role of the Hessian matrix in fitting models to measurements, Journal of Geophysical Research: Oceans, 94, 6177–6196, 1989.

Tilman, D.: Constraints and tradeoffs: toward a predictive theory of competition and succession, OIKOS, pp. 3–15, 1990.

Tilstone, G. H., Xie, Y.-y., Robinson, C., Serret, P., Raitsos, D. E., Powell, T., Aranguren-Gassis, M., Garcia-Martin, E. E., and Kitidis, V.: Satellite estimates of net community production indicate predominance of net autotrophy in the Atlantic Ocean, Remote Sensing of Environment, 164, 254–269, 2015.

Tjiputra, J., Roelandt, C., Bentsen, M., Lawrence, D., Lorentzen, T., Schwinger, J., Seland, Ø., and Heinze, C.: Evaluation of the carbon cycle components in the Norwegian Earth System Model (NorESM), Geoscientific Model Development, 6, 301–325, 2013.

Tjiputra, J. F., Polzin, D., and Winguth, A. M.: 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, 2007.

Torres, R., Allen, J. I., and Figueiras, F. G.: Sequential data assimilation in an upwelling influenced estuary, Journal of Marine Systems, 60, 317–329, 2006.

Tziperman, E. and Thacker, W. C.: An Optimal-Control/Adjoint-Equations Approach to Studying the Oceanic General Circulation, Journal of Physical Oceanography, 19, 1471–1485, doi:10.1175/1520-0485(1989)019<1471:AOCEAT>2.0.CO;2, 1989.

Urban, N. M. and Fricker, T. E.: A comparison of Latin hypercube and grid ensemble designs for the multivariate emulation of an Earth System Model, Computers & Geosciences, 36, 746–755, 2010.

Vallino, J.: Differences and implications in biogeochemistry from maximizing entropy production locally versus globally, Earth System Dynamics, 2, 69–85, 2011.

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.

Vallino, J. J. and Algar, C. K.: The Thermodynamics of Marine Biogeochemical Cycles: Lotka Revisited, Annual Review of Marine Science, 8, 333–356, 2016.

Van den Meersche, K., Middelburg, J. J., Soetaert, K., Van Rijswijk, P., Boschker, H. T., and Heip, C. H.: Carbon-nitrogen coupling and algal-bacterial interactions during an experimental bloom: Modeling a 13C tracer experiment, Limnology and Oceanography, 49, 862–878, 2004.

van der Meer, J.: Metabolic theories in ecology, Trends in Ecology & Evolution, 21, 136–140, 2006.

van der Merwe, R., Leen, T. K., Lu, Z., Frolov, S., and Baptista, A. M.: Fast neural network surrogates for very high dimensional physics-based models in computational oceanography, Neural Networks, 20, 462–478, 2007.

van Leeuwen, P. J.: Particle Filtering in Geophysical Systems, Monthly Weather Review, 137, 4089–4114, 2009.

van Leeuwen, P. J.: Nonlinear data assimilation in geosciences: an extremely efficient particle filter, Quarterly Journal of the Royal Meteorological Society, 136, 1991–1999, 2010.

Van Mooy, B. A., Keil, R. G., and Devol, A. H.: Impact of suboxia on sinking particulate organic carbon: Enhanced carbon flux and preferential degradation of amino acids via denitrification, Geochimica et Cosmochimica Acta, 66, 457–465, 2002.

Venzon, D. and Moolgavkar, S.: A method for computing profile-likelihood-based confidence intervals, Applied Statistics, pp. 87–94, 1988.

Wallhead, P. J., Martin, A. P., Srokosz, M. a., and Fasham, M. J. R.: Accounting for unresolved spatial variability in marine ecosystems using time lags, Journal of Marine Research, 64, 881–914, doi:10.1357/002224006779698387, 2006.

Wallhead, P. J., Garçon, V. C., and Martin, A. P.: Efficient upscaling of ocean biogeochemistry, Ocean Modelling, doi:10.1016/j.ocemod.2012.12.002, 2013.

Wallhead, P. J., Garçon, V. C., Casey, J. R., and Lomas, M. W.: Long-term variability of phytoplankton carbon biomass in the Sargasso Sea, Global Biogeochemical Cycles, 28, 825–841, 2014.

Wan, X. and Karniadakis, G. E.: Beyond Wiener–Askey expansions: handling arbitrary pdfs, Journal of Scientific Computing, 27, 455–464, 2006.

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

Ward, B. A., Dutkiewicz, S., Jahn, O., and Follows, M.: A size-structured food-web model for the global ocean, Limnology and Oceanography, 57, 1877–1891, 2012.

Ward, B. A., Schartau, M., Oschlies, A., Martin, A. P., Follows, M. J., and Anderson, T. R.: When is a biogeochemical model too complex? Objective model reduction and selection for North Atlantic time-series sites, Progress in Oceanography, 116, 49–65, doi:10.1016/j.pocean.2013.06.002, 2013.

Watanabe, S., Hajima, T., Sudo, K., Nagashima, T., Takemura, T., Okajima, H., Nozawa, T., Kawase, H., Abe, M., Yokohata, T., Ise, T., Sato, H., Kato, E., Takata, K., Emori, S., and Kawamiya, M.: MIROC-ESM 2010: model description and basic results of CMIP5-20c3m experiments, Geoscientific Model Development, 4, 845–872, doi:10.5194/gmd-4-845-2011, http://www.geosci-model-dev.net/4/845/2011/, 2011.

Watts, M. C. and Bigg, G. R.: Modelling and the monitoring of mesocosm experiments: two case studies, Journal of Plankton Research, 23, 1081–1093, 2001.

Weir, B., Miller, R. N., and Spitz, Y. H.: Implicit estimation of ecological model parameters, Bulletin of Mathematical Biology, 75, 223–257, 2013.

Westberry, T. K., Williams, P. J. l. B., and Behrenfeld, M. J.: Global net community production and the putative net heterotrophy of the oligotrophic oceans, Global Biogeochemical Cycles, 26, 2012.

Wikle, C. K. and Berliner, L. M.: A Bayesian tutorial for data assimilation, Physica D: Nonlinear Phenomena, 230, 1–16, 2007.

Wikle, C. K., Milliff, R. F., Herbei, R., Leeds, W. B., et al.: Modern statistical methods in oceanography: A hierarchical perspective, Statistical Science, 28, 466–486, 2013.

Williams, P. l. B. and Egge, J.: The management and behaviour of the mesocosms, Estuarine, Coastal and Shelf Science, 46, 3–14, 1998.

Williamson, D., Goldstein, M., Allison, L., Blaker, A., Challenor, P., Jackson, L., and Yamazaki, K.: History matching for exploring and reducing climate model parameter space using observations and a large perturbed physics ensemble, Climate Dynamics, 41, 1703–1729, doi:10.1007/s00382-013-1896-4, 2013.

Wilson, J., Ridgwell, A., and Barker, S.: Can organic matter flux profiles be diagnosed using remineralisation rates derived from observed tracers and modelled ocean transport rates?, Biogeosciences, 12, 5547–5562, doi:10.5194/bg-12-5547-2015, 2015.

Wirtz, K.-W. and Eckhardt, B.: Effective variables in ecosystem models with an application to phytoplankton succession, Ecological Modelling, 92, 33–53, 1996.

Wirtz, K. W. and Pahlow, M.: Dynamic chlorophyll and nitrogen: carbon regulation in algae optimizes instantaneous growth rate, Marine Ecology Progress Series, 402, 81–96, 2010.

Wood, S. N.: Statistical inference for noisy nonlinear ecological dynamic systems, Nature, 466, 1102–1104, 2010.

Wunsch, C. and Heimbach, P.: Practical global oceanic state estimation, Physica D: Nonlinear Phenomena, 230, 197–208, 2007.

Wunsch, C. and Heimbach, P.: How long to oceanic tracer and proxy equilibrium?, Quaternary Science Reviews, 27, 637–651, doi:10.1016/j.quascirev.2008.01.006, 2008.

Wunsch, C., Heimbach, P., and Ponte, R. M.: The Global General Circulation of the Ocean estimated by the ECCO-consortium, Oceanography, 22, 88, 2009.

Xiao, Y. and Friedrichs, M. A. M.: Using biogeochemical data assimilation to assess the relative skill of multiple ecosystem models in the Mid-Atlantic Bight: effects of increasing the complexity of the planktonic food web, Biogeosciences, 11, 3015–3030, doi:10.5194/bg-11-3015-2014, 2014a.

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, 2014b.

Young, G. A. and Smith, R. L.: Essentials of statistical inference, vol. 16, Cambridge University Press, 2005.

Zhang, W. and Arhonditsis, G. B.: A Bayesian hierarchical framework for calibrating aquatic biogeochemical models, Ecological Modelling, 220, 2142–2161, 2009.

Zhao, L., Wei, H., Xu, Y., and Feng, S.: An adjoint data assimilation approach for estimating parameters in a three-dimensional ecosystem model, Ecological Modelling, 186, 235–250, 2005.

Ziegeler, S. B., Dykes, J. D., and Shriver, J. F.: Spatial error metrics for oceanographic model verification, Journal of Atmospheric and Oceanic Technology, 29, 260–266, 2012.