Atmos. Chem. Phys. Discuss.,
https://doi.org/10.5194/acp-2018-1081-RC1, 2018

Atmospheric Chemistry and Physics
Discussions

Interactive comment on "Fundamentals of Data Assimilation applied to biogeochemistry" *by* Peter J. Rayner et al.

Anonymous Referee #1

Received and published: 1 December 2018

General comments

This manuscript gives an introduction to the theory of data assimilation and how it is applied in the field of biogeochemistry. The topic is very important and timely given the many diverse data assimilation systems that have been and are being developed. The authors have done a commendable job introducing the topic and bringing together many different methods, uniting them under the common umbrella of Bayesian formalism. However, I have some concerns about the content of the manuscript and how it is presented, and think the manuscript would benefit from some major revisions to improve its clarity and increase its value to the scientific community.

1. My first point of confusion is who this manuscript is aimed at. The authors state that "[the diversification of methods] can be confusing for a novice", and "If we succeed, a

reader should be well-placed to understand the relationships among the methods and applications presented later." These sentences seem to suggest that the manuscript is aimed at people who are not necessarily familiar with data assimilation and related techniques. However, later the authors make statements that seem to assume that the reader knows the meaning of some key terms. For example, the term "target variable" is used on page 5, line 13 without any prior definition. Furthermore, the manuscript makes references to "the inversion (or inverse) problem" many times without explaining what the inverse problem is (it is only mentioned briefly in Introduction on page 1, line 16).

If the aim of the manuscript is to introduce data assimilation for novices, I think the manuscript would benefit from some restructuring and further explanations. For example, a schematic of the data assimilation process and the different components would be helpful and could be referred to when describing the different parts of the system. I also think some concrete examples when describing the different implementations would help readers to follow the whole process better.

On the other hand, if the manuscript is aimed at readers with some familiarity with at least one of the methods (data assimilation, parameter estimation, inverse modelling, etc.), the manuscript should make it clear from the outset. In this case I think it would be valuable if the authors could add a section that describes different data assimilation implementations in biogeochemistry and discuss how they differ using the Bayesian framework they have established in the manuscript. This is done somewhat in section 8, "Historical Overview", but I think a more thorough description of state-of-the-art data assimilation systems and their main differences would add a lot of value for practitioners of biogeochemical data assimilation. Finally, the authors could conclude with current challengers in biogeochemical data assimilation and discuss where they see the most room for improvement.

2. I think the manuscript would be clearer if the authors choose one perspective from the outset (e.g. data assimilation) and present all material from this perspective in a

consistent manner. Currently the manuscript seems to borrow many terms from data assimilation, but frequently, especially in later parts, the language switches to that of inverse modelling (e.g. page 10, line 13: "We need therefore to incorporate these extra variables into the inversion process"; there are many other parts where "inverse problem" and "inversion" are mentioned). Section 6 is even named "Solving the Inverse Problem". I recommend to replace all occurrences of "inversion", "inverse problem" etc. with terms that are more common in data assimilation.

3. Data assimilation is probably best known for its applications in numerical weather prediction, where the technique is used mainly to improve meteorological initial conditions to produce better weather forecasts. In biogeochemistry, on the other hand, data assimilation (and related techniques) are more commonly used to constrain parameters. This difference is alluded to in the manuscript (e.g. page 15, lines 3-10). However, I think it would be better if this distinction is explicitly stated in the beginning of the manuscript. This difference explains e.g. why the manuscript does not focus on the dynamical model (the dynamical model for the target variables in biogeochemical applications is often unknown or assumed to be persistence, while the forecast model is an essential component in atmospheric data assimilation). A broader discussion about the choice of assimilation time window would also be helpful.

4. The authors write that "we will not be using mathematically precise language" (page 2, line 17). I can see where the authors are coming from, but I think this does a disservice to the readers. I recommend the authors to remove this sentence and to be mathematically rigorous to the extent feasible. I understand a mathematical precise language will take away some of the simplicity, but I think the benefits outweigh the added complexity; currently the reader may be left wondering where the language is imprecise and be less likely to refer back to the text when e.g. implementing a data assimilation system.

5. The manuscript does not talk much about the issues of ill-posed problems, which are common in inverse problems, and the need for regularization (except for under

"Historical Overview", page 19). I consider the use of prior error covariances, e.g. errors with a specified correlation length scale, to be a form of regularization. Even if the "true" error correlation length scales are smaller, a larger correlation length scale may be necessary for the data assimilation system to converge to a solution. This constraint will on the other hand lead to larger aggregation errors. It may also be worth to add a discussion about how the number of observations influences the design of the data assimilation system, e.g. the choice of regularization.

Specific comments

1. Page 3, lines 23-25: "As a practical example a frequentist may estimate a mean by averaging his sample while a Bayesian may calculate an integral over her probability density." I do not think this is a good example of the difference between Bayesian and frequentist statisticians. A better example could illustrate how Bayesian and frequentists interpret probabilities (e.g., a frequentist may only consider the long-term frequency of occurrence of a random event, while a Bayesian may draw from other prior information to assign probabilities, even for non-repeatable events).

2. Page 4, line 2: "we have followed it [the notion of Ide et al., 1997] here". It would be helpful to the reader to highlight exactly what is new with the notation introduced in this manuscript. Is it simply an extension of the Ide et al. (1997) notation for biogeochemistry data assimilation? Is it a generalization? More about this in the next point.

3. Table 1. Multiple points:

3.1. Consider adding a "Remarks" column and note when e.g. a notation differs from the notation in Ide et al. (1997).

3.2. For the definition of G, should mu and U be bold to show that they are a vector and a matrix, respectively? In that case also change "mean mu and covariance U" to "means mu and covariances U".

3.3. Descriptions of superscripts "a" and "b": "Posterior or analysis" and "Background or prior". Change to "Analysis or posterior" (to be consistent with "Background or prior").

3.4 Symbol Q: From my understanding Q is often used to denote model uncertainty (for the dynamical model). "Forecast uncertainty" here seems to include uncertainties due to initial conditions and boundary conditions.

3.5. Description of R: Add "(Observation uncertainty)" or something similar.

3.6. Symbol A: I do not think I have seen "A" used for "posterior uncertainty covariance" before. Maybe use Pˆb for background uncertainty covariance and Pˆa for analysis uncertainty covariance.

4. Page 6, Figure 1: It took some time for me to interpret this schematic. It may be helpful to mention that the observation operator in this case is a simple 1:1 mapping to the system state.

5. Page 10, lines 6-7: "then the dynamical model forms part of the mapping between the unknowns and the observations so is properly considered part of the observation operator". I know that this notation is common in e.g. atmospheric inversion, but I personally think it is unfortunate and easily leads to confusion (as the authors also note on page 9, lines 18-19, data uncertainty or data error may refer to the uncertainty due to errors in both observations and the observation operator, which is misleading). In e.g. 4D-Var for atmospheric data assimilation, I believe the observation operator and dynamical model are usually kept separate in the formulation of the cost function. Given that this manuscript focuses on clarity and fundamentals of data assimilation, I think it would wise to adopt the notation of atmospheric data assimilation and not conflate the observation operator and dynamical model.

6. Page 10, lines 12-13: "Frequently we regard these [parameters] as fixed, which is likely to underestimate the uncertainty of the estimates. We need therefore to incorporate these extra variables into the inversion process." I think "need to" is a bit too strong; maybe say instead that some methods incorporate these extra variables.

7. Page 20, lines 16-19: The authors mention that the time delay in EnKF may be problematic for tracers that live longer than the assimilation window. Is this not a common problem for all implementations?

Technical corrections

1. Page 1, footnote 1: I find the footnote unnecessary and suggest to put this information directly in the text.

2. Page 5, line 11: "by an observation operator". May be worth mentioning that this operator is sometimes also referred to as the forward operator.

3. Page 5, line 13: "target variable". Define the term.

4. Page 7, line 18: "to find variables to which it is sensitive". Replace "it" with "the quantity of interest" or something similar.

5. Page 8, line 10: Capitalize "This".

6. Page 8, line 10: Fix citation format (no parentheses).

7. Page 13, line 14-15: "These estimates will generally yield larger variability than that from our most likely flux". I do not believe "flux" is defined in this context. Maybe change to "realisation".

8. Page 14, line 23: "superscript f indicates the application of the forward model". I believe the authors are referring to the dynamical or forecast model here. The forward model is, from my understanding, often synonymous with the observation operator.

9. Page 15, line 4: "For data assimilation our motivation is to hindcast the state of the system". Consider changing to "For data assimilation applied to biogeochemistry our motivation is often to hindcast the state of the system", or something similar.

10. Page 16, line 30: "Another advantage is that, ...". It is not clear from the previous sentences that the previous statement (need to run a dynamical model for each realisation of the ensembe) is an advantage. Maybe change to "An advantage of EnKF is that, ..."

11. Page 19, line 7: Capitalize "possibilities".

12. Page 19, line 26: "(references (e.g. Gloor et al., 2000;". Remove "(references" or "(e.g.".

13. Page 20, line 16: Change format of citations (no parentheses).

Interactive comment on Atmos. Chem. Phys. Discuss., https://doi.org/10.5194/acp-2018-1081, 2018.