Geosci. Model Dev. Discuss.,
doi:10.5194/gmd-2016-25-AC2, 2016

Geoscientific Model Development Discussions
Open Access

Interactive comment on "Consistent assimilation of multiple data streams in a carbon cycle data assimilation system" *by* Natasha MacBean et al.

Natasha MacBean et al.

natasha.macbean@lsce.ipsl.fr

Received and published: 29 June 2016

Response to Interactive comment on "Consistent assimilation of multiple data streams in a carbon cycle data assimilation system" by Natasha MacBean et al. by Anonymous Referee #2

Received and published: 11 April 2016

This manuscript examines aspects of assimilating multiple data streams into carbon cycle models, includes discussion of the preceding literature and makes recommendations for the carbon cycle data assimilation (DA) community as to best practice when performing DA experiments. A real strength of this paper lies in the clarity of the description of the Data Assimilation problem.

Overall the work presented is well written, appears technically sound and should be easily reproducible. However the value of the individual parts of the manuscript feel somewhat limited, and as a whole I am not convinced they combine to make a complete piece of work. Although I don't doubt that setting up the DA system itself was technically complex, the experiments performed with it are rather limited in scope. My feeling is that it would have been easy to explore some further aspects of the carbon cycle DA problem and make the resulting manuscript much stronger with relatively little extra work.

» RESPONSE

We thank the reviewer for their clear and constructuve review of our manuscript. We understand all his/her concerns about the different parts combining to make a complete piece of work, and we have tried to address these concerns by following all of their suggestions, as detailed below.

»

The "advice for land surface modellers" in section 4 is a good concept but could be better organised. For example the points "conduct preliminary..." and "set up experiments..." are very related. I think the list should be tidied up - perhaps broken into different sections, for example "understanding errors", "preliminary analyses" and so on. Each of these sections can then contain the smaller points.

» RESPONSE

We agree with the reviewer on this. We have re-ordered the advice section accordingly taking into the suggestions above. However, we have also deleted many of the points related to general DA issues (e.g. conduct preliminary sensitivity analyses) as we felt, particularly after reading reviewer 1's comments, that this was confusing the focus of the paper on multiple data stream assimilation. Although many of the issues we raise are indeed general issues related to the assimilation of only one data stream, we tried hard in the manuscript to show how they affected an assimilation with more than one data stream. However we can now see that some of the points made in advice section were counteractive to this goal and could confuse the reader. We hope that in the process we have also tidied up the list, but given the list is now shorter and (hopefully) more focused, we have not broken the points up into sections. However we would be happy to have sub-sections instead of bullets for each of the points. We have put the new advice section at the bottom of this response.

»

The literature review section is reasonable but does not go into some of the preceding work in sufficient depth. In particular there are two studies I can think of that also look at carbon cycle DA problems with simple models that should have been dealt with in more detail. The Optic paper by Trudinger et al. (2007) is referenced, but a discussion of what experiments were performed and what they authors found is lacking. I think this is an important oversight given that this manuscript uses the same model. The Reflex paper by Fox et al. (2009) which looks at parameter estimation using a variety of DA techniques using a simple model and synthetic data isn't referenced. Furthermore the ordering of the manuscript feels a bit backward. One would normally expect the literature review to come prior to the experimental component and to set up the rationale for the experiments that follow.

» RESPONSE

We agree with the reviewer about the structure of the paper and so have put the literature review before the experimental component and slightly re-ordered it to better fit as an introductory section (see point (2) below). We have addressed the suggestions for additional papers below.

»

I have the following major recommendations to make the manuscript publishable: 1)

The experiments performed with the model need to be broader. There are several issues brought up later in the manuscript which could be easily examined. For example some simple experiments looking at populating the off-diagonal elements of the R matrix to set correlation between observations of S1 and S2 would seem to be an easy thing to do. I would be happy to see any sensible additional experiment though.

» RESPONSE

We agree that we could, or should, have added more experiments. Indeed we thought of such experiments from the outset of this work but ended up not including such experiments for fear the paper was too long or the message too complex. We agree that the most obvious, and hopefully most informative, experiment would be one investigating the impact of having correlated observations and populating (or not) the off-diagonal elements of R. We considered examining temporal autocorrelation, but as we want to focus on the multiple data stream aspects we have just considered the correlation between the two data streams. We have implemented this test, but the results we obtained were not what we expected (little impact). As we think this is the most useful extra experiment to include, we have asked the editors for more time (from the 24th June) to investigate this issue and we will provide a further update to this response within the next month. However we will upload the response to the rest of the comments now so the reviewer has more time to look and reply should they wish. The results from these experiments will be presented in a separate section at the end of the experimental section (now Section 3).

»

2) The literature review should be moved before the experimental section and modified so that it builds the rationale for performing the specific experiments undertaken. It should include greater discussion of the papers mentioned above. There are also classic problems in data assimilation which have not been well investigated in the carbon cycle to date such as localisation and errors or representativity and these have not

been mentioned. They should be added into the discussion.

» RESPONSE

We have moved the literature review before the experimental section and have removed some sections, either those that described the experimental results (P21 lines 22 to 27 and P22 lines 1 to 8 – which have now been included in the "advice and perspectives section 4 – see the end of the response), and we have deleted sections that we felt were superfluous, in order to shorten the length as requested by reviewer 1 (e.g. P21 lines 14-15 and lines 27 to 32, P22 lines 9 to 14). We have added in more refererence to the Trudinger and Fox et al. papers but we have not discussed these in too much detail because we want the emphasis of the literature review to be on multiple data assimilation. In this context the Fox et al paper is perhaps more relevant, so we were wrong not to include it in the original text. It is now included in Section 2.1 – Extra constraint from multiple data streams (P21 line 9 before "Thum et al."). The focus of the Trudinger paper is on testing the assimilation set up more than testing issues related to multiple data stream assimilation, and therefore we have not discussed the paper in too much depth. However, given we do want to emphasise the focus on multiple data stream assimilation (please see the response to reviewer 1 for further comments and changes to the manuscript in this regard), we have expanded the last paragraph in the introduction so it starts with the following: "This tutorial-style paper highlights some of the challenges of multiple data stream optimisation of carbon cycle models discussed above. Note that we do not aim to explore all possible issues related to a DA system, for example the choice of the cost function, minimization algorithm, or the characterization of the prior error distributions; indeed previous studies have investigated such aspects at length (e.g. Fox et al., 2009; Trudinger et al., 2007), therefore we refer the reader to these papers for more information. Section 2 reviews recent carbon cycle multiple data stream assimilation studies with reference to some of the aforementioned challenges. Section 3. . ." We hope that these additions are sufficient? Finally we have moved the following section from the advice to the literature review (end of Section 2.2

– impact of bias) because we felt it was better placed there and gave more context to the discussion on bias in FAPAR data seen in previous studies: "Aside from simple corrections, Quaife et al. (2008) and Zobitz et al. (2014) suggested that LSMs should be coupled to radiative transfer models to provide a more realistic and mechanistic observation operator between the quantities simulated by the model and the raw radiance measured by satellite instruments. This proposition followed the experience gained in the case of atmospheric models for several decades (Morcrette, 1991)."

We have tried to further build the rationale for performing the experiments with the toy models throughout the literature review, which we also hope was partly achieved by cutting out speculative and superfluous sections. To further help link the literature review and the experimental section we have added an introductory paragraph to (the new) Section 3 ("Demonstration with two simple models and synthetic data") that summarises the issues raised in the previous section and introduces the experiments at the same time. Therefore the following paragraph has been inserted before Section 3.1 ("Methods"):

"The three sub-sections in Section 2 highlight examples within a carbon cycle modeling context of the three main challenges faced when performing a multiple data stream assimilation, namely, i) the possible negative influence of including additional data streams into an optimization on other model variables; ii) the impact of bias in the observations, missing model processes or incompatibility between the observations and with the model, and iii) the difference between a step-wise and simultaneous optimization if the assumptions of the inversion algorithm are violated, which is more likely to be the case with non-linear models when using derivative-based algorithms and least-squares formulation of the cost function. The latter point is important because derivative methods (compared to global search) are the only viable option for large-scale, complex LSMs given the time taken to run a simulation. This section aims to demonstrate these challenges using simple toy models and synthetic experiments where the true values of the parameters are known. Most importantly this framework also allows us to investigate the impact of biases and violation of assumptions related to linearity (as discussed in Section 2.2. and 2.3), which are not always evident with real data and large-scale models. Thus the following sections include a description of the toy models together with the derivation of synthetic observations, the inversion algorithm used to optimise the model parameters and the experiments performed, followed by the results for each test case."

Finally, we have included one sentence in the literature review regarding representativity – if I have understood the suggestion of the reviewer correctly ("The spatial distribution of each data stream is also important, especially for heterogeneous landscapes (Barrett et al., 2005; Alton, 2013)") but we did not discuss this or the localisation problem further as we would like to keep the focus on multiple data stream assimilation and not general DA issues. Indeed, we have modified and added certain sentences throughout to try to reinforce this main focus of the paper (see response to reviewer 1).

»

3) The "advice" list needs to be re-written to provide a bit more order. See comments above.

» We agree and have done this – please see the response above and the new text at the bottom.

4) On page 11 at line 27 there is a statement suggesting that the data streams of s1 and s2 contain enough information to retrieve all the parameters individually for the quasi-linear model. This to me seems to be a flaw in the experimental design. Some of the conclusions from this part of the paper revolve around the linearity of the model, e.g. that differences between the step-wise and simultaneous experiments are minimal because of this. However given that the model is such that either set of observations can be used to determine both parameters it is not possible to say definitively that is the models linearity which is responsible for this. My hunch is that the authors are correct, but what would happen with a more complex linear model where not all parameters

are observable from either one data stream? The only way to demonstrate this is by introducing a new model - which I do not recommend - however I think it is vital that the authors are clear about what can or cannot be deduced from these experiments.

» RESPONSE

We thank the reviewer for pointing out the lack of clarity here. Indeed it is not possible to say definitively that it is related to the model linearity and we also feel fairly sure that if we had a more complex linear model that not all parameters would be observable from one data stream. We have therefore clarified this in the text by changing "under this assimilation set-up" to "under this assimilation set-up with this model", and with the sentence at the end of the section: "However, we cannot definitively say whether this is due to the simplicity or relative linearity of the model – it is possible that observations of variables in more complex linear model would not be able to retrieve the true values of all parameters."

»

I have the following minor comments: 1) The first paragraph of page 4 makes a lot of statements that are not referenced. It would be helpful to the reader who wanted to follow up on some of these aspects to provide references.

» RESPONSE

Thank you for pointing this out. We have tried to provide some references. This paragraph now reads: "Mathematically, the optimal approach is the simultaneous, but computational constraints related to the inversion of large matrices or the requirement of numerous simulations, especially for global datasets (e.g Peylin et al., 2016), and/or the weight of different data streams in the optimisation (e.g. Wutzler and Carvalhais, 2014), may complicate a simultaneous optimisation. On the other hand, in a step-wise assimilation the parameter error covariance matrix has to be propagated at each step, which implies that it can be computed. If the parameter error covariance matrix can be properly estimated and is propagated between each step, the step-wise approach should be mathematically equal to simultaneous. However, many inversion algorithms (e.g. derivative based methods that use the gradient of the cost function to find its minimum) require assumptions of model (quasi-) linearity and Gaussian parameter and observation error distributions (Tarantola, 1987, p195)."

We have also changed this sentence to explain more what we mean: "If these assumptions are violated, or the error distributions are poorly defined, it is likely that the step-wise will not be equal to the simultaneous, and that information will be lost at each step.", to: "If these assumptions are violated, or the error distributions are poorly defined, it is likely that the step-wise will not be equal to the simultaneous, because information will be lost at each step due to an incorrect calculation of the posterior error covariance matrix at the end of the first step.".

»

2) On page 5 I felt a bit more information was required about the model. How is the value of the functions F(t) being evaluated (possibly I have just misunderstood what is going on - so maybe just some clarification is needed).

» RESPONSE

Indeed we have not described how the function F(t) is calculated at all! Thank you for pointing this out. We have added the following sentence in: "The F(t) forcing term is a random function of time ("log-Markovian" random process) representing the effect of fluctuating light and water availability due to climate on the NPP (Raupach, 2007 – Section 5.3)." Also, following some of the comments from reviewer 1, we have added in further clarifications in this section, including for example the model time step in this sentence: "The first term on the right-hand side of Eq. (1) corresponds to the Net Primary Production (NPP) i.e. the carbon input to the system as a function of time, represented by F(t), weighted by factors (the two fractions in parentheses) that account for the size of both pools, in order to introduce a limitation on NPP.", and "It is based on

two equations that describe the temporal evolution (at a daily time step) of two living biomass (carbon) stores, s1 and s2, and the biomass fluxes between these two stores".

»

3) Page 23, line 4, I am not sure what is meant by orthogonal here. Given that S1 and S2 are interdependent on each other in the quasi-linear model the observations of them (assuming the model is correct, which it is in these synthetic experiments) cannot be not orthogonal. Perhaps the word "additional" would be better used here? Either that or I think the choice of "orthogonal" needs to be justified.

» We agree with the reviewer, this was a lax use of the word in this context. We have changed it to "additional".

Typographic and small errors: P03L10: step -> steps P03L19: one -> only P12L11: uniform -> constant (?) P13L11: than -> as P15L4-L11: this sentence needs to be broken up for clarity. P29L05: 2013.). -> 2013).

» RESPONSE

Thank you for these corrections, we have changed them all. Also we have changed P15 L4-11 from: "Most step-wise test cases (particularly 2b-d) do not result in the same parameter values as the simultaneous test case 3a in which all the observations are included (Fig. 4a), highlighting that strong non-linearity in the model sensitivity to parameters together with the use of an algorithm that is only adapted to weakly non-linear problems, as well as the assumption of linearity in calculating the posterior error covariance matrix at the minimum of the cost function, can result in differences between a step-wise and simultaneous approach in multiple – data stream assimilation (see Section 1)." To: "Most step-wise test cases (particularly 2b-d) do not result in the same parameter values as the simultaneous test case 3a in which all the observations are included (Fig. 4a). This highlights that strong non-linearity in the model sensitivity to parameters, together with the use of an algorithm that is only adapted to weakly non-linear problems, can result in differences between a step-wise and simultaneous approach in multiple – data stream assimilation (see Section 1).”

»

F2a: y-label should read "posterior" instead of "post"? F2b: y label should contain "%". F3caption: Equation should be 1-(RMSE_post/RMSE_prior)x100 F4b: as F2b

» Changed, thank you.

References: Trudinger, Cathy M., et al. "OptIC project: An intercomparison of optimization tech- niques for parameter estimation in terrestrial biogeochemical models." Journal of Geo- physical Research: Biogeosciences 112.G2 (2007). Fox, Andrew, et al. "The REFLEX project: comparing different algorithms and imple- mentations for the inversion of a terrestrial ecosystem model against eddy covariance data." Agricultural and Forest Meteorology 149.10 (2009): 1597-1615. Interactive comment on Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-25, 2016.

****

(NEW ADVICE SECTION 4 – remains in Section 4 after the toy model experiments) 4 Perspectives and advice for Land Surface Modellers Although it is clear that in many cases, increasing the number of different observations in a model optimisation provides additional constraints, challenges remain that need to be addressed. Many of these issues that we have discusssed are relevant to any data assimilation study, including those only using one data stream. However, most are more pertinent when considering more than one source of data. Based on the simple toy model results presented in here, in addition to lessons learned from existing studies, we recommend the following points when carrying out multiple data stream carbon cycle data assimilation experiments: âĂć If technical constraints require that a step-wise approach be used, it is preferable (from a mathematical standpoint) to propagate the full parameter error covariance matrix between each step. Furthermore, it is important to check that the

order of assimilation of observations does not affect the final posterior parameter values, and that the fit to the observations included in the previous steps is not degraded after the final step (e.g. Peylin et al., 2016). âĂć Devote time to carefully characterising the parameter and observation error covariance matrices, including their correlations (Raupach et al., 2005), although we appreciate this is not an easy task (but see Kuppel et al., 2013 for practical solutions). In the context of multiple data stream assimilation, this should include the correlation between different data streams, though note that this is not possible in a step-wise assimilation. âĂć The presence of a bias in a data stream, or an incompatibility between the observations and the model, will hinder the use of multiple observation types in an assimilation framework. Therefore it is imperative to analyse and correct for biases in the observationsand to determine if there is an incompatibility between the model and data. Alternatively, it may be possible account for any possible bias/inconsistency in the observation error covariance matrix, R, using the off-diagonal terms or inflated errors (Chevallier, 2007), or by using the prior model-data RMSE to define the observation uncertainty. âĂć Most optimisation studies with a large-scale LSM require the use of derivative-based algorithms based on a least-squares formulation of the cost function, and therefore rely on assumptions of Gaussian error distributions and quasi model linearity. However, if the these assumptions are not met it may not be possible to find the true global minimum of the cost function and the characterisation of the posterior probability distribution will be incorrect. This is a particular problem if the posterior parameter error covariance matrix is then propagated in a step-wise approach, although these issues are relevant to both step-wise and simultaneous assimilation. Therefore it is important to assess the non-linearity of your model, and if the model is strongly non-linear, use a global search algorithms for the optimisation – although at the resolution of typical LSM simulations ($\geq$0.5x0.5°) this will likely only be computationally feasible at site or multi-site scale.

In addition to the above points, we have investigated the impact of a difference in the number of observations in each data stream in this study. Test case 3b, in which only one observation was included for the s2 data stream instead of the complete timeseries, shows that a substantial difference in number of observations between the data streams can influence the resulting parameter values and posterior uncertainty (compare test cases 3a and b in Fig. 2 for the simple C model and Fig. 4 for the non-linear toy model) as each data stream will have a different overall "weight" in the cost function. Xu et al. (2006), among others, have mentioned the possible need to weight the cost function for different data sets. Different arguments abound on this issue. Some contend that the cost function should not be weighted by the number of observations because the error covariance matrices (B and R) already define this weight in an objective way (e.g. Keenan et al., 2013), and we would agree with this assertion. It should not be necessary to weight by the number of observations in the cost function if there is sufficient information to properly build the prior error covariance matrices (B and R). It is always useful to investigate the issues such as those discussed here by setting up synthetic experiments, as in this study, to understand the possible constraint brought by different data streams, and the impact of a possible bias and observation or observation–model inconsistency. Note also that performing a number of tests starting from different random "first guess" points in parameter space can help to diagnose if the global minimum has been reached, as outlined in Section 2.1.6 and discussed at the beginning of the results (Section 3.2). Furthermore, several diagnostic tests exist to help infer the relative level of constraint brought about by different data streams, including the observation influence and degrees of freedom of signal metrics (Cardinali et al., 2004). Performing these tests was beyond the scope of this study, particularly given that the simple toy models contained so few parameters, but such tests may be instructive when optimising many hundreds of parameters in a large-scale LSM with a number of different data streams. Aside from multiple data stream assimilation, other promising directions could also be considered to help constrain the problem of lack of information in resolving the parameter space within a data assimilation framework, including the use of other ecological and dynamical "rules" that limit the optimisation (see for example Bloom and Williams, 2015), or the addition of different timescales of information extracted from the data such as annual sums (e.g. Keenan et al., 2012).

Finally we should also seek to develop collaborations with researchers in other fields who may have advanced further in a particular direction. Members of the atmospheric and hydrological modelling communities, for example, have implemented techniques for inferring the properties of the prior error covariance matrices, including the mean and variance, but also potential biases, autocorrelation and heteroscedasticity, by including these terms as "hyper-parameters" within the inversion (e.g. Michalak et al. 2005; Evin et al., 2014; Renard et al., 2010; Wu et al. 2013;). Of course this extends the parameter space – making the problem harder to solve unless sufficient prior information is available (Renard et al., 2010), but such avenues are worth exploring.

Interactive comment on Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-25, 2016.