Geosci. Model Dev. Discuss.,
doi:10.5194/gmd-2016-25-RC1, 2016

Geoscientific Model Development Discussions

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

Anonymous Referee #1

Received and published: 8 April 2016

General comments:

The paper addresses the question of the assimilation of multiple data streams to estimate model parameters and initial conditions together with their uncertainty using variational method for simple C cycle models using synthetic observations.

The paper is organized around two parts. A first part presenting a variational data assimilation (DA) experiment for a simple, yet non trivial, quasi-linear model for the carbon cycle, and a non-linear toy model using multiple (2) data streams. The DA method, 4DVAR, and the experimental setup are succinctly but clearly described. The results of the experiments are extensively exposed, but - while VAR provides (via adjoint techniques) a set of tools to analyse the DA problem - not explained. The second part is devoted to a rather long but factual literature review of the studies using multiple

data streams to constrain LMS in general and their carbon component in particular.

While the paper illustrates some of the challenges of the model-data fusion problem, it does not describe any new idea, concept or tool, and thus does not represent a sufficiently substantial advance in modelling science. The advices presented at the end of the paper describes the golden rules for any DA experiment and the manuscript would benefit from a strict application of these advices. As it stands the paper only reproduces what other studies have done: performing the assimilation of multiple data streams following different scenarii.

I would recommend to shorten the literature review and to insert it before the experimental study and to perform a thorough analysis including sensitivity analysis, non-linearity issues, conditioning of the problem, information content. Due to its apparent complexity and because of "the burden" of coding and maintaining an adjoint, VAR is not the most popular method within this field, however it offers a frame- work where diagnostic and prognostic tools can be clearly (and sometimes analytically) defined, the capabilities of VAR deserve to be fully exploited in the scope of this paper.

Specific comments:

- Page 2: "Observations allow us to understand the system up until the present day, but they cannot tell us about the future (...). They also cannot distinguish between the complex interactions that may occur between different processes". I strongly disagree with this statement, observations do carry information about the future through the deterministic processes that, we believe, govern our world.

- Pages 5-6, lines 12-18: I found the input/output terminology on page 5, line 17-19, a bit misleading. A brief summary of dynamics of the model as described in the work of Raupack 2007 could be useful. The models and the dynamic variables they describe try to encompass different time scales from diurnal to potentially much longer time scales, and the variables themselves are likely to differ by several order of magnitudes. A discussion about the implication of the different typical scales could enlighten some

of the challenges. In the description of the experiments details concerning the time step size, observation window and observation frequency could be useful.

- Page 6, line 28: "including measurement and model errors", how to include model error without a weak constraint formulation?

- Page 7, line 6: "strong linear dependence of the model to the parameters", 4DVAR is the perfect framework where this issue should and could be investigated as advised in the section "advice for LMS modellers".

- Page 7, line 2: statement page 7 line 2 requires the model/observation operator to be linear as is discussed on page 17 line 20-31.

- Page 9 : concerning the experiment where only one observation for s2 is considered, worth mentioning that it corresponds (does it?) to the situation where only one estimation, say for soil C stock, is available. In this case is it used as a prior for s2 or as an observation later in the time window thus allowing the model to create correlation with other variables and parameters?

- Page 11, lines 14-17: discussion about "good or moderate reduction in RMSE for variables not included in any assimilation (...)" why is the reduction so poor for this flux? can this be expected from a model sensitivity analysis.

- Page 18, lines 16-19: "Rather if the model sensitivity to the parameters is very non-linear, multiple combinations of parameter values may exist that result in a similar reduction of the cost function (multiple minima), but provide a different fit to each data stream". This is exactly a crucial aspect that the paper should focus on, simplified and toy models are meant for this.

- Page 18, line 20: information content not defined, and more generally the expression "enough information" appear twice in the text but never made explicit.

- Page 18, lines 23-29: how to find the "troublemaker" and "peacemaker"?

- Page 26, lines 16-17: biases and inconsistencies, and other problematic features, could be addressed prior to optimisation in the context of the linearisation of the model.

- Page 29, lines 25-26: "it is crucial to understand the assumptions and limitations related to the inversion algorithm used" yet I feel that the paper did not provide the analysis, though possible with VAR, that would have helped understanding these assumptions and limitations in the case of the "simple" models presented here.

Technical corrections: - On page 1 line 25: "data stream" instead of "data steam". - On page 3 line 30: "matrices" instead of "matrixes". - In Table 1 : for the non-linear toy model the observation uncertainty for s2 is set to 0.5 whereas it is set to 5 for the simple carbon model, shouldn't it be 5 instead of 0,5?

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