Biogeosciences Discuss., 10, C7817–C7821, 2014
www.biogeosciences-discuss.net/10/C7817/2014/
© Author(s) 2014. This work is distributed under
the Creative Commons Attribute 3.0 License.

Biogeosciences
Discussions
Open Access

Interactive comment on "Ecosystem model optimization using in-situ flux observations: benefit of monte-carlo vs. variational schemes and analyses of the year-to-year model performances" *by* D. Santaren et al.

E. Weng (Referee)

weng@princeton.edu

Received and published: 10 January 2014

This paper presents a study comparing two data assimilation methods on optimizing a subset of parameters of a well established land model, ORCHIDEE, with four years high frequency eddy flux data (carbon and water exchanges between land surface and atmosphere) of a forest site. It found that the Monte Carlo approach (GA) is better than the gradient-based algorithm and assimilating more data (four years' data) made the model fit better than assimilating only one year's data. Overall, this paper is well-written. But I have couple of concerns regarding the design of this research, the

protocol of data assimilation, and the explanations to the results.

1. The design of this research

Before we estimate the parameters of a model, we must ask at least two questions: what information does the model need to estimate its parameters and what information can the data provide for constraining the parameters of this model? These two questions determine what results you can expect and, technically, what parameters should be chosen (relaxed) in the data assimilation procedures, because given model structure (or formulation) and data, the probability density functions (PDFs) of parameters are determined. The data assimilation procedures are just to find out what they are.

For example, as for a comprehensive model like ORCHIDEE, the system equation can be written as the following differential equation (this equation can help us to analyze the model and tell what data are needed):

$dX(t)/dt = AX(t)+BPs$ (Eqn 1)

$X(0)=x0$

where $dX(t)/dt$ is the net change of ecosystem carbon, vector $X(t)$ is the carbon in different pools; matrix A contains the parameters (and functions) governing carbon transfers among C pools and decomposition processes; Ps is photosynthesis, which is from a photosynthesis model and may interacts with leaf pool for most models. x0 is the initial value. For a differential equation, the solution is a bunch of equations if no x0. One needs the initial value to pinpoint one of them. So, for such a model, it may need both fluxes data and pool data to constrain the key parameters about carbon and water dynamics.

Usually, the fluxes data can only constrain the parameters related to response functions to make the model simulate seasonal or diurnal patterns fitted. If only flux data were used, as in this study, many parameters may covariate with different state values. That is, you can fit the flux curves, but the pools can go wild. Many parameters related to

the basal rates may vary with the assumptions of state variables (woody biomass and soil carbon pools). For some parameters, you may find it's no way to constrain them. This analysis can help choose the parameters to be optimized and explain the results (I'll discuss this point further below)

As for the data assimilation methods, we should expect that the gradient algorithm is unable to explore the highly dimensional PDFs of the parameters. Because the model is complex and there are many local minima that the gradient algorithm can't jump out. I suggest the authors to run the gradient algorithm for a couple of times with different initials so that we can know what happened there.

2. Data assimilation protocol

I want to discuss two points in this part: Initial states of plant and soil carbon pools and the parameters chosen for optimizing.

The authors used "equilibrium states" obtained from a 5000 yr model run as the initial states of the system (Lines 17∼19, page 18018). This is consistent with my expectation. Since not any data about plant biomass and soil carbon were used here, they have to find a way to initialize the model. But the forest at this site is young (40 yr-old European Beech) and the ecosystem has a high NEP (550 g m-2 yr-1). If the equilibrium states for carbon pools are directly applied in the data assimilation step, the decomposition rates must be underestimated to get a negative NEE.

The authors introduce a parameter (KsoilC, in equation A17, page 18042) to solve this problem. If I understand it correctly, this parameter scales the soil carbon pool down to it current state (lower than the equilibrium). So, according to my analysis above, the prior range of KsoilC should between 0 and 1, dependent on how far it is from the equilibrium. But, in Table 1, its range is set as 0.25∼4 and the posterior value for GA is higher than 1. It means the initial parameter values greatly underestimate soil carbon pool.

There are must be some biomass and soil data at this site. I suggested the authors use these data to define initial states. That will make the results more robust.

Let's go to Table 1 to see what parameters are estimated. From this table (pages 18053∼18054), we can see most of them are related to response functions (to temperate and moisture), in addition to photosynthesis parameters. The parameters related to basal rates and turnover of carbon pools (including mortality rate of woody biomass) and allocation are not here. There is only one parameter to define soil carbon state, KsoilC, as mentioned above. And, I don't find the parameters related to maintenance respiration. I think it might be very low for sapwood and roots, so that leaves respiration can represent all the components of it.

So, these "state" related parameters must be fixed in the model. And, the optimized parameters are conditioned on those state variables. Once you change them to another set of values, the optimized parameters must be systematically varied with them.

3. Results explanation

About the evaluation of data assimilation methods (Lines 26, Page 18024∼ line 8, page 18025; Section 4.1 minimization algorithms, page 18032∼18034):

I think it is a problem of model vs. data information, though the effectiveness of methods is an issue. Here, the key issue is that eddy-flux data can't constrain a comprehensive model like ORCHIDEE. Actually, even highly simplified ecosystem models (as defined by Eqn 1) can't be constrained by eddy-flux data if the models have C pools.

About parameter uncertainty (Sections 3.3 Parameter uncertainty estimates and 4.2 Parameter optimization):

It should be mentioned that these parameters are conditioned on those fixed ones, which may also lead to unrealistic parameters values in optimization.

By the way, as for modeling carbon and water dynamics at site level, ORCHIDEE model is not special comparing with others. These models share very similar formulations in

simulating these processes. So, they share the same successes and have similar problems.

About data information (lines 10∼19, page 18036):

A more detailed analysis here about how the parameters affect the simulations would be much better. Anyway, the model is not a black box, we know how and why. I also think the data information should be eventually saturated with time. The four years data are not long enough. The authors may try the sites with longer time series data (e.g., Harvard forest).

Interactive comment on Biogeosciences Discuss., 10, 18009, 2013.