Biogeosciences Discuss., 10, C9042–C9051, 2014
www.biogeosciences-discuss.net/10/C9042/2014/

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.

D. Santaren et al.

peylin@lsce.ipsl.fr

Received and published: 5 March 2014

For the sake of clarity, we put each reviewer's comment (R) before our responses (A).

****

R: 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.

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.

A: We agree with the reviewer's statements although we believe that the answers to those questions are also partly provided by the optimization process. A direct analysis of the sensitivities of the model outputs with respect to its parameters helps indeed to determine which observations constrain which parameters. However, one difficulty is that a sensitivity analysis has to be handled carefully when dealing with a complex and non-linear model such as ORCHIDEE. This was not the main focus of our work and we thus did not explicitly mentioned any sensitivity test in our work even though we ran a preliminary one (Morris analysis) to determine roughly which parameters should be kept constant (i.e. parameters with a very weak impact on the observations at the considered time-scales). Ideally, we should consider all parameters that have a significant uncertainty but practically it may hamper/slow the convergence of the algorithms. Practically, we restricted our study to nearly the 28 most sensitive parameters but we recognize that a more thorough analysis of the parameter sensitivities to observations would be beneficial and will be central for following studies. Finally, concerning the information content of the data, the computation of a posteriori errors (Eq. 2) provides a way to assess the level of constraint of the observations on the parameters as we illustrate in paragraph 3.3 (Parameter and uncertainty estimates) and discussed in

paragraph 4.1.

****

R: 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.

A: The reviewer assertion is indeed crucial. However, the use of soil of carbon pools data to constrain model parameters is complex and beyond the scope of this paper. The problem is that measurements of soil carbon cannot be reliably compared to the simulated carbon pools by ORCHIDEE (slow, passive, active); there is a fundamental issue to relate the observations, usually limited to the top 20cm or 100cm soil depths, to the simulated quantities that include all soil carbon pools. Prior to the assimilation of these data, we need to derive from the observations the equivalent total soil carbon that should be compared to the modeled pools. Concerning the biomass pools, we could have used accurate measurements of the aboveground carbon pool but it would have involved simulations over 40 years (age of the forest) which would have largely increased the computing cost of an optimization. Moreover, using the above ground biomass implies that we properly simulate the human thinning that occur in this forest every 5-7 years. Such work is only under progress. As we focus the study on the

“fast carbon fluxes”, i.e. from synoptic to inter-annual, but not on long time-scales for which the inclusion of information about the soil/biomass carbon pool size has a crucial impact, we choose to discard these carbon pool data in the optimization process. Nevertheless, we agree with the importance of this type of data for constraining the ORCHIDEE parameters and it will be the subject of a following study about the assimilation of above ground biomass data including disturbance effects (thinning since the plantation, i.e. removal of nearly 5% of the biomass each 5-7 years in the recent past).

****

R: 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).

A: As mentioned above the purpose of our study is to evaluate the structure of ORCHIDEE and its ability to simulate CO2 and water fluxes at different time-scales (from diurnal to annual). We acknowledge that the optimal values of the model parameters may be biased if we have not the right carbon pools, but this bias will have the most impact for long term future simulation, not investigated in this paper. We are thus fully aware that flux observations do not contain enough information to constrain all the parameters of the model and that for a given parameter, multiple optimal values can lead to the same fit to the data (equifinality). We have reinforced the discussion on this topic in the discussion section (around P18033).

****

R: 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.

A: We agree with the reviewer. However, we illustrate this point in paragraph 3.1. where the performances of the genetic algorithm (GA) and the gradient-based algorithm (BFGS) are compared through 10 twin optimizations. To assess the convergence of the gradient algorithm (using BFGS), we started the downhill iterative search from 10 different initial parameters sets that were randomly prescribed within the admissible range of variation of the parameters (P18024 L8->L13). Moreover, we have discussed the convergence efficiency of the GA and BFGS algorithms in the first paragraph of the discussion section (P18032 L17 -> P18033 L22).

****

R: 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.

A: We do not agree completely with the reviewer as we believe that the problem is slightly more complex. Indeed the whole history of the site, including past land cover change and disturbances are also crucial to position the initial carbon pool size. Given

that we can not reconstruct these land use change trajectories and their impact on the soil carbon pools we choose: 1) - to make a spin up of the model as done in most global studies. 2) - to further correct for any past historical effect on the soil carbon pool, we have introduced a coefficient that scales the carbon pools at the initial date of the simulation (KsoilC, eq. A17). In this context the decomposition might be underestimated if we consider that the current soil carbon pools are larger than the true pools (i.e. the spin up brought the pools to a maximum value for a given soil carbon input). Although this is most likely, they might be also overestimated if the past disturbances would have led to larger soil carbon pool than actually (i.e., a fire that led to accumulate highly recalcitrant carbon material in the soil,...). Moreover and importantly the use of a scaling coefficient to initial soil carbon pool allows to partly relieve the constraint of the initial equilibrium for the soil. The case of the above ground biomass remains a constrain, but we believe it is rather small given that a 40-yr old forest already has an above ground biomass close to equilibrium. We however added these points in the discussion to clarify the message.

****

R: 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.

A: We agree that the range of variation of KsoilC was artificially too large. We defined this range to keep a certain symmetry in the scaling of the carbon pools by this parameter (i.e. 0.25: divide by 4) and to explore the as large as possible parameter space. However we agree that this was not appropriated and we have tested the optimization with a more realistic range: The test results with the new restricted range

of variation for KsoilC do not significantly differ as it could already be anticipated with the previous results (all KsoilC were in the range (0.2, - 1.2)). Therefore, given that it would require a lot of computations to redo all the optimizations of the study and that the conclusions would remain unchanged, we have kept the original results. Note that our primary focuses are to assess the ability of the model to simulate CO2 and water fluxes at several time-scales after optimization and how different minimization algorithms (GA vs BFGS) perform with a complex model. We thus focus less on the values of some parameters but more on the potential of embedded equations. Moreover, given the large correlations between the parameters that were selected, the CO2 flux observations cannot help to fully distinguish between KsoilC and other parameters like the dependence of the heterotrophic respiration to temperature (Q10). Therefore, the optimization of KsoilC can still provide values above 1 that are most likely unrealistic, and such behavior would then be compensated by other parameters such as Q10. To clarify this point, we added in the discussion section a study of the impact of the equilibrium hypothesis on the optimized parameter values

****

R: 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.

A: We agree with the reviewer's suggestion but, as we have explained above, we did not assimilate this type of data due to both technical and scientific complications that need to be solved prior to any proper use of these data. This is currently investigated with the inclusion of a realistic thinning, correction of the upper soil carbon measurements to account for the whole soil carbon content,... All this work is beyond the scope of this paper.

****

R: 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 temper-

ate 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.

A: The reviewer raises the important issue of selecting parameters for the optimization process. It's right that the optimizations were performed with a restricted number of parameters to keep reasonable times of computation. Parameters were selected according to a sensitivity test based on their impact on the synoptic, seasonal and interannual NEE flux variations. In this case the "state" related parameters, although important, do not show the highest impacts. However, we agree that the a posteriori values of the parameters that are optimized depend on the parameters that are not included in the optimization process. To limit this dependency, we invert as many parameters as possible given the temporal resolution of eddy-covariance fluxes; we mention this point in paragraph 2.3.4 (P18023 L5->L17). Moreover, the conclusions of the paper do not rely on particular values of certain parameters but on the ability of the model to represent short-medium term flux variability. Such ability may be even larger than shown in the paper given that some parameters were kept constant. Note that parameters related to maintenance respiration have been included within the data assimilation system (Parameters MRoffset and MRslope Table 2, Eqns. A14 et A15).

****

R: 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.

A: We agree that our explanation is incomplete. By "data assimilation framework", we wanted to mean the minimization algorithms, the formulation of the cost function, the choice of parameters to be optimized and the type of data used for the optimization. We have rephrased the sentence: "The data assimilation framework and especially the type of assimilated data (NEE and LE diurnal fluxes) does not allow to distinguish parameters that are correlated."

****

R: 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.

A: We agree with the reviewer's suggestion and we have rewritten part of the discussion to emphasize this aspect. Note that we partly addressed this point with the sentence: "Although only few parameters may be well constrained by the observations, removing parameters from the optimization process could bias the estimated values of the remaining ones." (P18033L28). However, to clarify the message, we have introduced in the discussion section a new paragraph entitled "Which parameters to optimize ?" to highlight the discussion about the choice of the parameters and the importance of taking into account unconstrained parameters.

****

R: 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.

A: Through the cross-validation experiments, we investigate in section 3. how parameters inverted from one year or the whole period of observations affect the simulations. We have rephrased the beginning of the section to highlight this aspect.

****

R: 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).

A: Indeed, it will be very interesting to extend the study to longer series and to other sites. This is partly done in Kuppel et al., 2012 (see ref in the text). However, one particularity of the Hesse site and the period that we choose is that it contains very different climatic conditions, with an exceptionally dry and warm summer in 2003 and a relatively mild and wet summer in 2002. In this context, our study already relies on diverse meteorological conditions that help to assess the potential of the model at one site under various climate conditions. Note finally that studies at different sites with longer records are in progress.

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