Review Mathew Williams

The comparison between two parameter estimation methods is informative although I am left wondering why these two were chosen and how these methods would compare to others (e.g. Haario, Saksman and Tamminen (2001) An adaptive Metropolis algorithm. Bernoulli, 223-242.). Of course it would be ridiculous to demand more methods to be assessed but it would be good to know why these two methods were chosen for this specific study.

We first choose a gradient-based algorithm given that the study should be put in a more general context of a global carbon cycle multi-data assimilation systems (P.J Rayner et al., 2005, "Two decades of terrestrial carbon fluxes from a carbon cycle data assimilation system (CCDAS)"). With the perspective of doing global scale optimization using either global observations or observations at many sites, the computing time becomes a crucial issue and the gradient methods are well suited compared to random search algorithms (especially if an adjoint model to compute the gradient of the cost function is available). We further choose the BFGS algorithm to implement our variational method, given its wide use in the community and its ability to restrict the parameter search over a given range of variation. Note that we have also used this algorithm with ORCHIDEE in previous studies: Santaren et al. (2007), Kuppel et al. (2012) (see ref. in the paper).
Additionally, we selected a "Monte Carlo" approach to illustrate the potential of global search algorithms using a random generator. We choose the genetic algorithm (GA) among indeed a large ensemble of methods, including various forms of MCMC, simulated annealing,… Our choice was guided by few practical issues, linked to the computing efficiency (i.e. the number of iterations) and the ease of implementation,… With these criteria in mind, we discarded MCMC algorithms, given that their computational expense may be prohibitive with global terrestrial ecosystem models (see "On the capability of Monte Carlo and adjoint inversion techniques to derive posterior parameter uncertainties in terrestrial ecosystem models", Ziehn et al., Global Biogeochemical Cycles, 2012). The GA proved to be relatively efficient and was implemented as an example of Monte Carlo algorithm but not as the unique solution.
We thus agree with the reviewer that the justification of these methods should be reinforced in the paper. We added in the section 2.3.2 about the minimization algorithms the sentences (at P18021L7):
"The GA was chosen among other global search algorithms (Simulated annealing, Monte Carlo Markov Chains (MCMC)) because of its ease of implementation and its efficiency in term of convergence. Particularly, MCMC algorithms were discarded given that their computational expense may be prohibitive with global terrestrial ecosystem models (Ziehn et al., 2012)"

The comparison in model performance when fitted to different time slices is also informative and sends a clear message about the fact that while these models are "process inspired" we should not necessarily expect (and I would say rarely) the parameters in the model to map on to what they are meant to represent in reality (indeed I personally would expect most to NOT map directly on to reality, and if we force them to do so then we should expect the model to weaken elsewhere).

Most biogeochemical process representations are indeed a simplification of the reality and model equations cannot embody the whole complexity of the processes they are meant to represent.

The insight into the benefits in model performance to using a wide range of environmental conditions as training data is excellent – this is exactly what is hoped for. However, I disagree with the assertion that "This suggests that ORCHIDEE is able to robustly predict the fluxes of CO2 and latent heat of a temperate beech forest after optimisation." This is wishful thinking – there is no more evidence presented as to the wider applicability of the model (e.g. evaluate but not train against independent data on different forests for example). All we know is that the model performs to a particular measure of performance under the range of conditions provided for the 4 years for

this specific beech forest. I would expect decreases in performance when applying it to other temperate beech forests and/or under other environmental conditions for similar reasons to why the model performs poorly for the whole dataset when only trained to a year worth of data.

We fully agree with the reviewer that this assertion is largely overstated. Indeed, a parallel study by Kuppel et al. (2012) using the same model, showed that the model parameters derived from single site optimizations can be substantially different than those derived from a multi-site optimization. We have change the text to:
"This suggests that ORCHIDEE is able to robustly predict after optimization the fluxes of CO2 and latent heat of a specific temperate beech forest in France (Hesse site)."

P18015 L23 "enhance the model fit to observations equally?"

We agree and have changed accordingly.

P18016 L1 This question needs rephrasing. Perhaps "Does the information gained from parameter optimization indicate anything new about modeling carbon fluxes in deciduous forests?"

We agree. The question has been replaced by the proposed one.

P18017 L1 I'm not sure you define the acronyms NEE and LE before this.

The reviewer is right and we thus have replaced the sentence by:
"Within this period, less than 5% of the net CO2 flux (NEE) and latent heat flux (LE) data were missing and observational gaps mostly occurred during December 2001, February and December 2004 where periods with no data lasted around 10 days."

P18017 L4 u* star (is star redundant?), this is too much jargon and needs a brief definition

We agree. The friction velocity, usually denoted u*, is used as a criterion to discriminate between low and well mixed periods.
We have replaced the sentence by:
"We use original data that were not corrected for low turbulence conditions nor gapfilled, nor partitioned (Reichstein, 2005; Papale et al., 2006)"

P18017 L20 given->for

Change as suggested

P18017 L20-21 this sentence needs to be made more readable... especially the "having a very long turnover rate".

We have replaced the sentence by:
"Model processes are computed on a wide range of time-scales: from 30 min (e.g. photosynthesis, evapotranspiration, …) to thousands of years (e.g. passive soil carbon pool decomposition)."

P18017 L26 delete "and"

Done

P18018 L5 LSMs needs definition

Done: Land surface models (LSM)

P18018 L17 and rest of paragraph. This is a key paragraph that is slipped in here. This should be thoroughly discussed in a paragraph in the discussion. I will come back to new items that I'd like to see in the discussion below.

We agree with the reviewer that the spin up of the model to bring all carbon pools to equilibrium is a crucial hypothesis (although used in most studies) that should be discussed further. We thus added few sentences in the discussion about that particular point to mention that:

- Bringing to equilibrium the soil carbon pools imposes that we also optimize the initial value (at t0) of the active and slow pools in the soil. As for the passive pool, given its turnover time of around 1000 years, we could keep it to its equilibrium value, as it only contributes little to the heterotrophic respiration.
- Ideally we should consider using measurement of soil carbon pools to scale the soil pools at t0, but these measurements only represent the carbon content of the upper soil and their assimilation is not straightforward.
- We made an additional test where we assimilated the NEE observation where the net yearly carbon sink has been removed (With this observation, the forest seems to be at equilibrium). Such test indicates that most of parameters remain relatively unchanged compared to the standard optimization, while only few significantly differ; especially those that are related to the heterotrophic respiration.
- Finally, the above ground biomass pool (AbGB) should also in principle be optimized in order to match the measured biomass of this 30-year old forest.
- An ongoing study with ORCHIDEE is currently investigating these issues and how to properly assimilate measurements of soil carbon and AbGB with FluxNet data.

P18018 L27 how do justify your data probability distributions as Gaussian? Is there any evidence that sampling or observational error is Gaussian? Please justify this assumption. Similar for your parameters. Some parameters may be log-normally distributed (especially rate parameters)

We agree that data/model error probability distributions have few chances in reality to strictly follow Gaussian distributions. An experimental study with two towers has shown that random errors on eddy-covariance measurements may be represented by double-exponential distributions (Hollinger and Richardson, 'Uncertainty in eddy covariance measurements and its application to physiological models', 2005). We agree as well that some parameters are more likely to be log-normally distributed.

The major reason of assuming Gaussian distributions is that the Bayesian optimization problem becomes equivalent to a least squares minimization and thus analytically solvable by minimizing a cost function (Tarantola, 87). This simplification is not needed for the Monte Carlo approach (Genetic Algorithm) but is at the base of our implementation of the variational scheme. We have thus, like in most variational optimization studies, used the Gaussian approximation. Using other distribution with a variational scheme is beyond the scope of this paper. Note finally that, whatever the type of error distributions may be, the minimum of the cost function can be still considered as optimal from the perspective of minimizing root mean square differences but, if errors distributions are not gaussian, optimal parameters cannot be considered as the most likelihood.

However we agree with the reviewer that such choice should be better justified in the text. We have thus added two sentences at the end of line 18, P18019:

"Although we acknowledge that observation and parameter errors do not strictly follow Gaussian distributions, we have used such hypothesis to simplify the Bayesian optimization problem which becomes equivalent to a least squares minimization and thus analytically solvable by minimizing a cost function like Eq. 3. (Tarantola, 87)."

P18019 L27 – Why are data uncertainties chosen? Are they not provided with the data themselves (e.g. for a given datapoint do you not have NEE plus and minus a certain amount representing your measurement uncertainty?). Are you ignoring the true data uncertainties? Or are you using the sampling errors to give relative weights within datasets (e.g. an NEE with high uncertainty has less influence than an NEE with low uncertainty) but then applying an additional weighting between datasets (e.g. LE versus NEE) to balance their effects on the parameters?

Data errors represented by the matrix **R** encompass both measurement and model errors. Even though measurements are sometimes provided with an estimation of their random uncertainty, systematic uncertainties (due to horizontal advection, low turbulent mixing, …) are difficult to estimate and never provided with the measurements. More importantly, errors on model representation (i.e., from representativeness issues as well as missing processes) are complex to assess and are considered to be the dominating part of the overall data error. We thus choose a rather simple approach, used in many studies, based on the mean squared difference between the prior model and the observations. The underlying hypothesis is that the larger the model – data mismatch is the larger the error is. This led to define relatively large error values (typically 25% of the maximal amplitude of daily fluxes) that should roughly take into account both observational and model errors. An additional benefit of this approach is that the relative weights of the NEE and LE datasets within the cost function are similar.
Finally we have checked that with our choice (R matrix) the value of the normalized Chi-Square after the optimization (i.e. twice the value of the cost function divided by the number of observations) does not indicate strong error underestimation, (i.e. value not greater than one), a conservative choice.
However, we agree with the reviewer that the text was lacking justification about the choice of data errors. We have thus added the following sentences at the end of the line 2 P18020:
"This rather simple formulation leads to relatively large errors (typically 25% of the maximal amplitude of daily fluxes) but it is consistent with the fact that data uncertainties comprises not only the measurement errors (random and systematic) but also significant model errors (from representativeness issues as well as missing processes)."

P18020 L3 – "chosen to be relatively large"

Sentence changed accordingly

P18020 L3 – What was the basis for selecting these ranges at all? Expert knowledge? Publications? Based on the original hand chosen values used in previous studies? I see you include a paragraph on this in P18023 L18-21. This should be stated before.

The description of parameter errors was indeed not precise enough. The chosen values gather expert knowledge about the different equations used in ORCHIDEE as well as reported values from an ensemble of publications or from databases (such as the TRY database (http://www.try-db.org)). Moreover the crucial point is to use specific bounds for these parameters that have been carefully adjust to the possible physical range for each parameter. The little paragraph is thus replaced by:
"The magnitudes of prior parameter uncertainties (matrix **B**) were chosen i) according to expert knowledge and ii) to be relatively large to minimize the influence of the Bayesian term and thus to provide the maximum leverage of the observations upon the parameters. Moreover, for the minimization algorithms, an additional important information is the range of variation for each parameter; such range is prescribed from expert knowledge and typical observed values (TRY database). As in Kuppel et al. 2012 (see paper references), the prior error for each parameter was thus set to 40% of its prescribed range of variation (Table 1) and we do not consider correlations between a priori parameter values."

P18020 L20 – please explain what you mean by "sorely"

The sentence has been replaced by:
"For complex models, the construction of the TL model is a rather sophisticated operation (even with automatic differentiation tools (Giering et al., 2005)) where threshold-based formulations are not handled and need to be rewritten with sigmoid functions for instance."

P18022 L21 – but is the model linear with Gaussian distributions for data and initial parameters {this should be parameter} errors? This assumption should be discussed in the discussion.

Indeed, one needs to verify the assumption of both the locality and linearity approximation to compute **Pa**. Cross sections of the cost function J(**x**) for the different parameters near their minimum are useful with that respect: if the model is locally linear, the shape of J(**x**) is parabolic with the minimum coinciding with the optimal value of the parameter. We have checked and illustrated the "local linearity" of the ORCHIDEE model in a previous work for a Pine forest (Santaren et al., 2007, Optimizing a process-based ecosystem model with eddy-covariance flux measurements: A pine forest in southern France). We have thus build on this previous result but we should also stress that this study do not strongly rely on the interpretation of the posterior parameters uncertainties. This estimates are only used in a relative way to highlight which parameters are the most constraint. We thus believe that any discussion about this issue would overload the discussion and would require additional tests beyond the scope of the paper to bring new information to the reader compared to Santaren et al. (2007).

P18025 L1-2. This may not be due to the data assimilation framework but simply the fact that you lack adequate data constraints. This needs to be addressed in the discussion.

We agree with the reviewer that the sentence was somehow misleading. The term "data assimilation framework" was supposed to include not only the minimization algorithms or the choice of parameters to be optimized but also the type of data used for the optimization. We have thus rephrase the sentence to be more explicit about the main cause which is indeed the lack of adequate constraints:
"The chosen data assimilation framework and more specifically the type of assimilated data (NEE and LE daily means) does not allow to distinguish several parameters that appear to be correlated".
Moreover, we mention in the conclusion that additional type of data should be used to constrain more parameters than when using only NEE and LE data. We think that this argument should be put forward in the conclusion rather than in the discussion section.

P18025 L7-8. I don't understand what you mean by "The importance of the equifinality problem is directly related to the posterior uncertainty of the parameters." Do you mean that the more equifinality you have (parameter correlations?) the more uncertainty you have in the posterior parameter estimates?

The sentence was indeed not clear enough but it is exactly what we mean. To clarify the message, we have rephrased the sentence:
"As Santaren et al. (2007) illustrated for the ORCHIDEE model, the more uncertain an optimized parameter is, the most likely the equifinality problem could affect this parameter."

P18025 L16-19. This is hardly a surprising conclusion – indeed you can say a lot about parameter dependencies/correlations/equifinality from a variance-covariance matrix. This is not a result, it is common knowledge. I'd rather you say how you intend to use this information in interpreting the parameter estimates when using the real data.

We agree that this conclusion refer to common knowledge. However, we want to insist on the utility of the posterior variance-covariance matrix to gain information on the retrieved parameters, as it is done in section 3.3. We thus slightly rephrase the sentence to: "In summary, we stress again the need to use the information from the posterior variance-covariance matrix and especially the correlations between parameter errors to highlight the level of constraint on each parametrization (see section 3.3 with real data). »

P18027 L2 in->on

Done

P18028 L23 – what is a "proper simulation"? Please be more informative.

We wanted to mean a correct simulation. "proper" is thus replaced by "correct"

P18028 L24 At->On

Done

P18031 L16 an->a

Done

P18032 L21 to converge->from converging

Done

P18033 L14 "A lot of effort has to be put". Please reword.

Sentence replaced by: "An important programming work is required especially if the model contains parameterizations with thresholds."

P18033 L24 "The lack of convergence problem usually" please reword

Sentence replaced by: "The convergence of the minimization algorithms is less effective when the number of optimized parameters increases."

P18038 L5 Table 2-> Table 1

Indeed, we have changed accordingly.

P18038 "where vegetation is treated as a single equivalent surface for the carbon cycle." Please reword

Sentence replaced by: "ORCHIDEE is a "big leaf" model that maps the processes and properties of a whole canopy onto a single representative leaf [Jarvis, P.G., Scaling Processes and Problems, Plant Cell and Environment, 18 (10), 1079-1089, 1995.]".

P18042 L4-L7 – the sentence starting "Soil litter laid..." is a bit garbled and would be better being reworded

We propose a new sentence : see point below

P18042 L7 increasing-> decreasing

Done

P18042 L8 – do you mean 3 soil carbon pools?

It's indeed 7 pools. In order to clarify, the sentence was rewritten:
"Soil litter of the forest floor is partitioned into 4 soil carbon pools (structural/metabolic litter above/below ground) with structural and metabolic pools characterized by different turnover times. Soil organic matter is distributed among three soil carbon pools of increasing turnover (active, passive and slow)."

P18056. You need axis labels (BOTH AXES) – remember units on the y-axis.

Done

P18057. You need axis labels (BOTH AXES) – remember units on the y-axis. Also I don't understand your error bars – They appear to be symmetric about the mean – consistent with Gaussian distributions, but then this doesn't make sense when the error bar extends beyond the parameter maxima and minima or go past zero – what's going on here?

We agree that within a strict probability framework, the errors bars should not extend beyond parameter maxima or minima or go past zero for positive parameters.
First, we should notice that the use of Gaussian probability distribution together with minimum and maximum values is not strictly rigorous, given that the Gaussian probability should not be bounded. However such set up is commonly used, provided that the bounds are relatively large compare to the wide of the Gaussian distribution. In our case, like in Kuppel et al. 2012 [see ref in the text], we defined the prior errors (one sigma of the Gaussian PDF) as 40% of the range of variation for each parameter. The posterior errors are thus smaller than this value but given that the mean parameter estimates is not anymore centered on the range of variation, we can thus have one side of the error bar that reach one limit. Finally, we should stress that there was a small mistake on the calculation of the posterior errors for few parameters that led to the strange behavior noticed by the reviewer.
The proper computation does not change the conclusions arising from the a posteriori errors assessment. The corrected figure is now inserted in the revised manuscript.

P18058. Axis labels. Label legend

Done

P18060 Label x axis. More importantly, you should convey the observation errors. Either shade in the background or simply include a representative error bar somewhere on the graph. It is not acceptable to show these plots without conveying observational error. This will allow us to gague the relative adequacy of the different model prediction lines.

We agree, a shaded area representing data errors has been added.

P18061. Same comment as immediately above.

Done

Figs. 5, 6 and 8: I'm disappointed that there is not propagation of parameter uncertainty through the

predictions. Given you have parameter uncertainty estimates, these should really be propagated into the predictions to convey your uncertainty in those predictions. I'm not insisting that you do this though.

We agree with the reviewer that a discussion on the error propagation would be useful. However, we have not included such discussion yet given that:
- We are currently invetigating in depth the computation of the error with the linearity or without the linearity assumptions. Such work is more comprehensive but still beyond the scope of the paper.
- We will in a future study compare the approach of the paper (linearity and gaussian assumption for the errors) with a more general approach based on a MonteCarlo estimation of the errors.
Note however that we have mentioned in the conclusion the uncertainty propagation as a future research direction.

Discussion My most substantial criticism is that the discussion is not thorough enough. For me, it appears too much like a reiteration of results and not enough genuine discussion.

We agree that the discussion could be more in depth. We have thus tried to discuss some of the points raised by the reviewer, although not going into long argumentation in order not to overload the paper and not to increase substantially its size.

What data would you recommend being collected to obtain better constraints on parameters? The simplest thing, if the data exists, is to include a wider variety of data on different aspects of what the model is designed to represent. It is asking a lot to constrain so many parameters with just two sorts of data, even if they do vary over multiple timescales (of course it's also asking a lot to ask you to somehow magic up a wider variety of relevant datasets).

We agree that the use of solely NEE and LE daily fluxes is not sufficient enough to constrain the parameters that we selected. For instance the computation of the respiration fluxes (maintenance, or heterotrophic respiration) are base on first order kinetics and corresponds the product of the pool size with a decay rate and with temperature/water scaling functions (Eq. A17). The measurement of the fluxes at night thus only infers on the overall respiration flux but could not infer separately on the pool size or the decay rate or the Temperature/Water functions. With this in mind, we thus highlighted the need for measurement of carbon pools, including above ground biomass and soil carbon pools. To our minds, these discussions were belonging to the conclusion, where we indeed mention that the use of "Observations of soil carbon content, respiratory fluxes from flux chamber measurement, above ground annual wood increment or LAI would help the optimization framework to determine more accurately the parameters" We have thus partly introduced and reinforced this point in the discussion section.

Would simplifying or reparameterising the model be a sensible future step to avoid wasting time with an underconstrained model?

It's indeed a very interesting question that should be answered by comparing the predictive abilities of ORCHIDEE and a simpler model. However, ORCHIDEE has been developed from the perspective of including process-based equations to describe the different processes; most simple approaches that could rely on statistical formulations have been generally discarded. The drawback of this approach is indeed the impact of under-constrained equations. However, we expect that on a longer-term perspective this approach is more promising and that it will provide the best tool to test specific hypothesis linked to climate change, land use change, fertilization,.. Overall, we believe that this analysis is unfortunately beyond the scope of the paper.

Does it even matter that some of these parameters are unconstrained (in terms of the predictive

performance)?

First, we have to recall that all parameters have a prior uncertainty that come from the results of in-situ or field measurements/experiments. They should not be considered as totally unconstrained. However, we investigated this point in the discussion (P 18033 L23 -> P18034 L 16): fits to the data are analyzed when the optimization is performed only with a restricted set of the most sensitive parameters. The dramatic deterioration of the optimal model performances compared to the case when all the parameters are considered illustrates the need of including the so called "unconstrained parameters" to guarantee the predictive performances of the model.
Nevertheless, we acknowledge that this point was not clearly highlighted and we have slightly rewritten the concerned assertions within a new discussion paragraph entitled "Which parameters to optimized ?".

In what ways would you expect the current version of ORCHIDEE to be structurally adequate or inadequate to predict in terms of carbon flux dynamics in different sites or under different conditions?

In this paper, we have investigated only the temporal aspect of the prediction, i.e. same site but different conditions. This was the aim of the section 4.2 of the discussion. We show there that the cross-validation experiments allow to highlight the structural deficiencies of ORCHIDEE in modeling the evapotranspiration fluxes during dry periods or the leaf onset processes. Nevertheless, we point out that these deficiencies do not prevent ORCHIDEE to correctly simulate the different fluxes at the seasonal and annual scales. We have slightly reinforced this discussion.
Concerning the predictive capability at different sites, this was the specific subject of a parallel study with ORCHIDEE (Kuppel et al. 2012, see ref in the paper). Particularly, they have shown, for a group of several sites of temperate deciduous broadleaf forests, that the multi-sites optimization leads to model-data RMS reductions at each site comparable to the corresponding single-site optimization. We have thus introduced in the discussion the results of this finding to complement our discussion on the temporal prediction for a given site.

Are there other ways to deal with the equifinality problem that could be attempted? For example, cleverly re-parameterising the model (e.g. fitting the product of two parameters rather than the parameters separately) can help – as can using schemes that automatically estimate the parameter covariance matrix and use that in MCMC parameter estimation (various adaptive MCMC methods do this).

Other ways to deal with the equifinality problem could be indeed attempted. Concerning the optimization process itself, the use of more performing algorithms and the reparameterization of the model could help. Also, the form of the cost function could be changed to extract more precisely the temporal information of the data and hence reduced the impact of the equifinality. For example, we could split the cost function into terms representing the data-model mismatch at different frequencies (diurnal, synoptic, seasonal, annual ...), the optimization will likely have a different effect on parameters that are associated to different frequencies. Probably, the KsoilC parameter (Eq A17) which scales the annual value of the heterotrophic respiration and the Q10 parameter which strongly impact the seasonal cycle of the heterotrophic respiration would not be so strongly correlated as with the current data assimilation framework (Fig. 3). However, we did not include these aspects in the discussion because we did not investigate them thoroughly and we believed that including them within the discussion section would unnecessarily lengthen the paper.
Concerning the suggestion to re-parameterise the model by fitting the product of two parameters, we believe the attempt may decrease the equifinality problem but not necesseraly lead to a better predictive model. This is not guaranty. It will just reduce the number of degree of freedom in the optimization problem and may lead to underestimate the overall model uncertainty after data

optimization.
Besides, as we mention in the conclusion, additional type of data (Biomass, LAI, ...), by providing new constraints on the parameters, would help in dealing with equifinality. Using longer data sets of NEE/LE and from another observational sites would as well decrease the undetermination of certain parameters due to equifinality.

I simply encourage you to put more value in the discussion for a general reader – few are going to want to use the paper to decide whether to use ORCHIDEE for this specific site – they are more interested in the general insights, the general applicability of the model, the lessons learned that should be considered in other studies. How would you take what was done here to learn about how to fit ORCHIDEE to many more sites to end up with a data-constrained global model? Please provide the reader with a bit more generality to your conclusions.

We will take this last suggestion into consideration and we will revise the discussion in order to derive more general conclusions to the reader on the potential of ORCHIDEE and on the lessons learnt with our optimization approach that other groups might use.
In concrete terms, the changes that we have made to the discussion section include:
1- 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.
2- In this new paragraph, we are discussing the use of new types of observations to constrain more efficiently the parameters of the model,
3- In the new paragraph, the impact of the equilibrium hypothesis on the optimized parameter values is discussed.
4- In the paragraph "Parameter optimization and model validation", we added some sentences about the predictive ability of the model at different sites.