Biogeosciences Discussions

Biogeosciences Discuss., author comment AC2
https://doi.org/10.5194/bg-2022-109-AC2, 2022

Reply on RC1

Cédric Bacour et al.

Author comment on "Assimilation of multiple datasets results in large differences in regional- to global-scale NEE and GPP budgets simulated by a terrestrial biosphere model" by Cédric Bacour et al., Biogeosciences Discuss., https://doi.org/10.5194/bg-2022-109-AC2, 2022

We thank the reviewer for his/her time spent on our manuscript and for his/her useful comments and suggestions. We answer them point by point below.

Summary

- **In this paper the authors report on a series of data assimilation experiments in which three different data streams (tower fluxes, MODIS NDVI, atm CO2 concentrations) were assimilated into the ORCHIDEE model individually, in pairs, or all together to assess the impact of multiple constraints on model parameters and predictions. The clearest take-away from the exercise was the importance of initializing soil carbon (C) pools by site or region to account for the non-equilibrium nature of the contemporary carbon cycle.**

We are sorry if this is the sole take-away message retained by the reviewer as this is not the main message we wish to convey. We acknowledge that the presentation of the objectives and conclusion were not clear enough, as the two reviewers regret. We have therefore restructured the introduction and conclusion sections to highlight more the main messages of the study.

From a general perspective, this study addresses the broad community of global process-based terrestrial biosphere models (TBMs), which may be less aware than the Data Assimilation community about the challenges and caveats associated with the joint assimilation of multiple data-streams, and with respect to set-up and model structure. To our knowledge, there are only a few similar studies addressing those questions from a global carbon cycle/TBM perspective.

The main take-away messages of the paper concern:

- the importance of assimilating different and complementary data-streams all together in order to avoid model overfitting and reduce the risk of degrading the model performance with respect to data-stream/variables not assimilated;
- the importance of using *global*scale measurement datasets (here atmospheric CO_2data) constraining the soil carbon disequilibrium in order to optimize the net land sink albeit the remaining challenges.

(It is worth noting that flux measurements at sites are not representative of large regions and do not allow tuning the global land sink. Atmospheric CO_2 data provide such constraint, although indirectly, but their assimilation remains sub-optimal in our set-up and DA framework because of the large initial bias between the data and the model).

In addition to these messages (which were identified in the "key points" of the manuscript), our study also addresses some technical aspects:

- in assessing the impact of the combination of the data–streams assimilated on the optimized net and gross carbon fluxes;
- in exploring different metrics allowing to improve and check the consistency of the DA set-up, and in particular assess the informational content brought by each data–stream.

As it will be detailed later, we have modified the introduction and conclusion sections to clarify the study objectives and so the main messages of the study.

Major comments

- **In the Intro (L103-106) the authors raise the point that joint assimilation is "more optimal" than sequential assimilation but don't really explain the point, despite this being absolutely central to what is novel in this paper relative to previous ORCHIDEE DA papers. This really needs to occur, as from a theoretical perspective there should be NO difference between these two approaches – it's very easy to mathematically prove that under Bayes theorem, joint and sequential assimilations give the same answer. The authors acknowledge this somewhat in the Discussion, but even that doesn't really explain what's going on in their system. Given the numerical approaches used (numerical optimization instead of MCMC, SMC, or EnKF; covariances dropped or coarsely analytical approximated) it's unclear which results represent ORCHIDEE-specific problems and what are general issues the rest of the community needs to worry about (and I'm somewhat inclined to think much of this is an artifact of how the ORCHIDEE sequential DA was implemented). To constructively move forward, please spend more time in the intro acknowledging and explaining these differences and more time in the Discussion reflecting on what the general take-home messages are and a bit of space on other Bayesian numerical methods and speculating about whether they are going to have the same issues.**

We agree with the reviewer that we did not develop on the differences between stepwise and simultaneous approaches (although this was discussed in §4.1). This is mostly because this is not the novelty of this paper and because it has been discussed in more detail in other studies the reference of which are provided in the document. In particular we already explored the differences between the two approaches in MacBean et al. (2016) using a toy-model. And so, with respect to this notion of "optimality", we only referred to the paper of Richardson et al. (2010).

We do agree with the reviewer on the point that sequential and simultaneous assimilations should lead to the same results if the error covariance matrices are properly quantified and propagated during the different steps of the sequential approach. However, this is not the case in practice for complex problems as reported and discussed in Kaminski et al. (2012) or MacBean et al. (2016), and as it was already summarized in §4.1 of the paper, given that the posterior error covariance matrix can not be easily quantified with complex global TBMs (i.e., only some elements of that matrix are usually derived). This is clearly

not a feature that is inherent to our system/model as suggested by the reviewer. Summarizing the above mentioned studies, differences between the stepwise and simultaneous approaches may arise due to incorrect description of the error probabilities. The use of a gradient descent algorithm for optimization, with the risk that it gets trapped in local minima, and equifinality also increase the probability that stepwise and simultaneous approaches diverge. For the stepwise approach, an incorrect calculation of the posterior error covariance matrix at the end of each step will likely result in a loss of information at the next step. An incorrect description of the observation(–model) error distribution can result from a poor characterization of the error correlations or a model-data bias. Using a toy model, MacBean et al. (2016) evaluated the impact of a model bias (linear model) and the model non-linearity (no bias) in the context of calibrating parameters of computationally expensive TBMs at global scales using gradient descent methods (due to the computation cost of "global search" methods) and showed considerable differences between the simultaneous and step-wise approach.

We have changed the text in the Introduction to:

"**Although with model parameters and observations described by probability distributions, simultaneous and sequential assimilations should theoretically lead to the same result (Tarantola et al. 2005), this is not the case in practice for complex problems. Incorrect description of the error statistics may result in large differences between simultaneous and step-wise approaches (see Kaminski et al., 2012; MacBean et al., 2016). In addition, model non linearities also tend to exacerbate these differences. Simultaneous assimilation is considered to be more optimal in the context of optimizing TBM parameters as it maximizes the consistency of the model with the whole of the datasets considered (Richardson et al., 2010; Kaminski et al. 2012) and avoid incorrect propagation of the error statistics from one step to the other (Peylin et al., 2016). The use of gradient descent approach for optimization, with the risk that it gets trapped in local minima, also increases the chances that stepwise and simultaneous approaches diverge. However,** sequential approaches remain appealing for modelers…**"**

In the Discussion section (§4.3 "Caveats and perspectives concerning the initialisation of the soil carbon pools"), we made a general statement about the importance of identifying model-data biases and ultimately correcting them:

"**From a more general perspective, the detrimental consequences of model-data biases become even more important when assimilating multiple observational constraints because of their interconnected contribution to the model calibration. It should be noted that the impact of systematic model-data errors is not inherent to our minimization approach (gradient-based) and has also been highlighted using random search approaches (Brynjarsdóttir and O'Hagan, 2014; Cameron et al., 2021). Thus, the importance of accounting for bias correction approaches into data assimilation schemes (Dee, 2005; Trémolet, 2006; Kumar et al., 2012) becomes increasingly important as the complexity of models and the amount of observational constraints increase**".

- Brynjarsdóttir, J., & O□Hagan, A. (2014). Learning about physical parameters: The importance of model discrepancy. *Inverse problems*, *30*(11), 114007.
- Cameron, D., Hartig, F., Minnuno, F., Oberpriller, J., Reineking, B., Van Oijen, M., & Dietze, M. (2022). Issues in calibrating models with multiple unbalanced constraints: the significance of systematic model and data errors. *Methods in Ecology and Evolution*.
- Kumar, S. V., Reichle, R. H., Harrison, K. W., Peters□Lidard, C. D., Yatheendradas, S., & Santanello, J. A. (2012). A comparison of methods for a priori bias correction in soil moisture data assimilation. *Water Resources Research*, *48*(3).
- Trémolet, Y. (2006). Accounting for an imperfect model in 4D□Var. *Quarterly Journal*

of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography, *132*(621), 2483-2504.

We compare our different simultaneous experiments to the stepwise result (based on Peylin et al. 2016), of which our study is the continuation, to illustrate precisely that "configuration / inverse setup matters" (different combinations of assimilated data-streams result in different C flux estimates). We acknowledge that we did not study here the different possible configurations of the stepwise approach (i.e. choice of the order of the data-streams that are assimilated) which goes beyond the scope of our paper. If the review feels that comparing the simultaneous experiments investigated in our study to one stepwise result is too confusing, we would remove the latter from the paper.

If the results from the stepwise approach is kept, we would correct the associated results in the paper: We had initially used those from Peylin et al. (2016) which relies only on three years of atmospheric CO_2 data while our study assimilates 10 years of atmospheric data, which complicated any direct comparison between the stepwise and simultaneous approaches. The updated results correspond to assimilating the same 10 years of atmospheric CO_2 data in the step 3 of the step-wise approach. The impact on the main messages of the paper is marginal. The corrections in the paper would hence concern:

- §2.3.3: "**While only 3 years of atmospheric CO_2 data were used in Peylin et al. 2016), the stepwise results presented here really accounts for the same ten years used in the simultaneous experiments (2000-2009) to facilitate the comparison of the approaches (in particular the impact of using the atmospheric CO_2 growth rate over 10 years on the optimisation of the mean terrestrial carbon sink). There are however a few** differences in the set-up compared to the present study"

- §3.1.3: we would remove the sentence "**which is probably due to the longer period of the atmospheric CO_2 data considered (10 years vs 3 years for the stepwise)**".

- §3.2: we would change the sentence " For these experiments that include CO_2 data, the optimized carbon sinks are about -2.4 GtC.yr^{-1} at the global scale, **with the exception of the stepwise approach, which is -1.7GtC.yr^{-1}**" to "For these experiments that include CO_2 data, the optimized carbon sinks are about -2.4 GtC.yr^{-1} at the global scale, **similar to the stepwise approach**"

- §3.2: we would remove the last sentence (" Note that the lower terrestrial sink obtained with the stepwise approach…") which was related to the assimilation of three years of atmospheric CO_2 data.

- §3.2: we would change the text:

"While the three joint assimilation experiments F+CO2, VI+CO2, and F+VI+CO2, lead to similar NEE budgets across regions, the CO2 and F+VI+CO2-2steps experiments result in distinctly different estimates. In the northern extra-tropics, the CO2 assimilation results in the largest C sinks (numbers provided in Supplementary Text S6), while the F+VI+CO2-2steps assimilation leads to the lowest C sink, with a magnitude that matches the stepwise assimilation set-up (Peylin et al., 2016). The reverse is obtained for the Tropics."

to

"While the three joint assimilation experiments F+CO2, VI+CO2, and F+VI+CO2, lead to similar NEE budgets across regions **(with magnitudes comparable to the stepwise assimilation)**, the CO2 and F+VI+CO2-2steps experiments result in distinctly different

estimates. In the northern extra-tropics, the CO2 assimilation results in the largest C sinks (numbers provided in Supplementary Text S6), while the F+VI+CO2-2steps assimilation leads to the lowest C sink. The reverse is obtained for the Tropics."

-§3.2: we would change the text:

"the stepwise and F+VI+CO2-2steps assimilations follow the typical partitioning pattern of TBMs' behavior, with a stronger C sink in the tropics than in the northern hemisphere. On the opposite, the three two or more data stream experiments F+CO2, VI+CO2 and F+VI+CO2 lead to an approximately equal C sink in the northern hemisphere and tropics."

to

"the F+VI+CO2-2steps assimilation follows the typical partitioning pattern of TBMs' behavior, with a stronger C sink in the tropics than in the northern hemisphere. On the opposite, the three two or more data stream experiments F+CO2, VI+CO2 and F+VI+CO2 and **the stepwise** lead to an approximately equal C sink in the northern hemisphere and tropics"

-§3.3: we would remove the reference to Peylin et al. (2016) and removed the sentence "at the last step where three years of atmospheric CO_2 data were assimilated"

- we would update Figures 3, 4 and 6, as well as Supplementary Text S1 and Figure S1.

- Intro fails to include a clear statement of objectives, research questions, and hypotheses. The whole time I'm reading the rest of the paper I found myself asking "But why? What's the question you're trying to answer with this analysis?"

We have restructured the last part of the introduction in order to make the objectives and research questions more clear:

"**By conducting different assimilation experiments in which each data stream is assimilated alone or in combination (for all combinations of datasets), the research questions that we address in this study are:**

- **What impact does the combination of different data streams assimilated have on the reduction in model-data misfit, and to which extent are the model predictions improved (or degraded) with respect to the other data-streams that were not assimilated?**
- **What is the impact of the choice of the combination of assimilated data-streams on the values and uncertainties of the optimized parameters, and on the predicted spatial distribution of the net and gross carbon fluxes at regional and global scales? How do the derived carbon budgets compare with independent process-based model and atmospheric inversion estimates from the Global Carbon Project's 2020 Global Carbon Budget (Friedlingstein et al., 2020)?**
- **How does a large model–data bias related to incorrect initialisation of soil carbon disequilibrium impact the overall optimisation performances within a Bayesian assimilation framework relying on the hypothesis of Gaussian errors?**

Our analysis of the useful informational content provided by different data-streams on global C fluxes with our set-up is supported by methodological aspects aiming at:

- **Checking that the error statistics on parameters and observations are correctly assigned, and ultimately enhancing the realism of the prior error statistics by making them consistent with the differences between prior model simulations and observations;**
- **Quantifying the observation influence of each of the three data streams on the joint assimilation in which all three datasets were included in the optimization.**"

- **In terms of many of the high-level conclusions of the paper, it's not clear what here wasn't predetermined by structural choices of the Model Set-up (e.g. differences in which parameters were constrained depending on which data were assimilated). In particular the choice to spin-up carbon pools to equilibrium would automatically make it impossible for the model to match C observations unless C pool initial conditions were included in the calibration (the paper focuses on soil C, but undoubtedly this parameter is also indirectly absorbing some of the effects of the non-equilibrium vegetation C pools). This structural choice, more than anything about the DA itself or the choice of data streams or their information contribution, predetermines the main conclusion of the paper of the importance of assimilating soil C. Similarly, if you only let NDVI constrain phenology (not LAI, photosynthesis, etc) then it's obvious that NDVI isn't going to improve predictions of C fluxes or concentrations, and it's not surprising that it can actually make those things worse (since the default parameters will have compensating errors baked into them to get the right NEE with the wrong phenology, such that "fixing" the phenology will break NEE).**

The reviewer is right in stating that some of the study findings are due to our modeling and assimilation set-up. One aim of this study is precisely to point out that 1) the configuration, and 2) the combination of data to be assimilated, do matter for simulating regional/global C fluxes using a land surface model. We are addressing people from a large community working with global process-based TBMs who may not be fully aware of those caveats, and not only to the data assimilation community.

The reviewer is right stating that the correction of the soil C pools indirectly absorbs the effects of the non-equilibrium vegetation C pools. The optimization of the KsoilC parameter was meant initially to correct soil C pools disequilibrium while for the vegetation C pools, the optimization of the model parameters partly correct for the vegetation C pool disequilibrium (i.e. for forests). The correction of the vegetation C pools should be done with existing information on forest age and forest management practices. However, such correction was beyond the scope of that paper and thus our soil C correction somehow also partly correct for vegetation pools in the case of forest ecosystems.. We acknowledge that we put too much emphasis in the paper on the "correction of the soil C pools" when what we meant was the "correction of the soil carbon imbalance" (Carvhalais, 2008). It is worth noting that the topic of initialisation of the soil C pools is also a point that is crucial for reviewer 2.

We do not assimilate soil C pools related datasets because 1) having a more "data-driven" initialisation of the soil C reservoirs would only partially address this disequilibrium and because 2) we would anyway expect large biases between such datasets and our organic soil carbon model (please refer to our response #4 to Reviewer 2). We have also improved §4.3 in the Discussion to justify why we do rely on soil C products.

On the point raised by the reviewer on the fact that constraining phenology only using NDVI will not improve predictions of C fluxes, we actually showed the opposite in MacBean et al. (2015) in which the assimilation of satellite NDVI data improved the timingof the simulated GPP at the global scale. As for the atmospheric CO_2 concentrations, the study of Kuppel et al. (2014) showed parallel improvements in the seasonality of simulations of LAI/NDVI and atmospheric CO_2 concentrations, resulting from multi-site assimilations of *in situ* flux data. These are strong evidences that the phenological information conveyed in NDVI data is beneficial in improving the seasonality of the modeled GPP (even though NDVI informs more on the photosynthetic capacity of the terrestrial ecosystems than on their actual photosynthetic activity).

The level of useful constraints provided by one given data-stream is a function of its associated errors, of the model structure itself (i.e., if the relevant processes are included or not) and to the error of the observation operator. The large differences observed between different LAI products (Garrigues et al., 2008) make NDVI/FAPAR a better option for now to constrain the timing of GPP with our modeling framework (see discussion in Bacour et al. (2015), §4.2).

- Garrigues, S., Shabanov, N. V., Swanson, K., Morisette, J. T., Baret, F., & Myneni, R. B. (2008). Intercomparison and sensitivity analysis of Leaf Area Index retrievals from LAI-2000, AccuPAR, and digital hemispherical photography over croplands. *agricultural and forest meteorology*, *148*(8-9), 1193-1209.
- Kuppel, S., Peylin, P., Maignan, F., Chevallier, F., Kiely, G., Montagnani, L., & Cescatti, A. (2014). Model–data fusion across ecosystems: from multisite optimizations to global simulations. *Geoscientific Model Development*, *7*(6), 2581-2597.

- **The set of assimilated data is notably out-of-date. The La Thuile flux data set (2006) was replaced by the FLUXNET2015 data set 7 years ago. The MODIS 5 data set (2005) is 3 versions behind the current collection (6.1, out since 2017), and the specific product used is considerably coarser than the actual data (0.72 degrees vs 250m) and then subsampled to just 15 pixels per PFT. In both cases the newer versions of NEE and NDVI are considerably larger in size that what was used and fix known errors in the previous versions. Similarly, the atmospheric CO2 constraint ends in 2009 and there is no mention anywhere in the entire paper of the satellite CO2 data sets that have absolutely revolutionized atmospheric inversions. There's a passing mention at the end of the paper (Line 867) that no data after 2010 is used in the analysis, but the authors don't say anything anywhere about WHY this is so and what impact this has on the inferences being made. Why is so much data being left on the table? And if you're only going to use a subset of NDVI, why not include 250m data from the FLUXNET sites so that you gain the benefits of co-located constraints?**

The reviewer's remark is fully legitimate and we acknowledge we have failed in making it clear why these datasets are used instead of more up-to-date/alternate ones.

The core objectives of the study are to investigate the complementarity of different data-streams and the challenges in assimilating them simultaneously into a complex process-based TBM to optimize its simulation of carbon fluxes from regional to global scales, more than achieving an up-to-date re-analysis of the carbon cycle. The latter task would indeed require more recent data as well as more informative observations (as for instance solar-induced fluorescence -SIF - over NDVI). This is indeed out of the scope of our study and will be the subject of future work.

The more recent FLUXNET dataset (with more sites/years) has not changed the mean informational content related to photosynthesis and plant phenology as the one used in this study. This also holds for atmospheric CO_2 concentration data. For the spatial resolution of NDVI, our choice is also limited by the spatial resolution of available meteorological forcings. We use here ERA-Interim meteorological fields provided on a regular 0.72° grid.

In addition, we also seek to compare with the stepwise approach of Peylin et al. (2016) of which our study is the continuation. This implies that we use the same observation data (hence the same space-time resolution). Another more pragmatic reason relates to the study timeline: Its development has actually started in parallel to the one of Peylin et al. (2016) at a time where the three data-streams were not so "obsolete". The inherent technical (including the computation time) and scientific challenges make that the results are being published only now, which stresses even more the obsolescence of the three data-streams. However, we are still using a long time period of data for the three data-streams and adding more recent data would not change the main outcomes of our study on their complementarity. In addition, we could also probably investigate the impact of different observation record lengths, but again, this isn't the point of this particular study.

As for the assimilation of space-borne XCO_2 retrievals (from OCO-2 or GOSAT, for instance), this requires the full coupling with an atmospheric transport model (and its adjoint) which is also more computationally expensive. This technical coupling between ORCHIDEE and LMDz is currently being implemented within our assimilation framework and the scientific analysis of the information provided by gridded XCO_2 data to constraint NEE at the global scale will be the subject of future work.

Finally, the use of more recent data than those assimilated in this study will allow us to evaluate the predictive performance of the model over the recent period in the future.

- **In the two step optimization, F+VI+CO2-2steps, it sounds like the CO2 data is being used twice, in both the first and second steps (i.e. double dipping). This is not OK and artificially inflates both overall sample sizes and the influence of the CO2 data. On a more minor point, how are you doing this without incurring all the disadvantages of sequential assimilation that are the point of the paper?**

We appreciate the 2steps approach is confusing in the context of the assessment of simultaneous assimilations. Indeed the 2steps approach consists in a stepwise assimilation where atmospheric CO_2 data are used at each step, as it is described in §2.3.3. We discuss in the paper (§4.2) on the fact this approach is not optimal: it was designed to overcome the issue of correcting the large bias in the soil C imbalance which precludes significant changes in the model parameters other than in the multiplicative factor of the soil carbon pools. The results obtained with the 2steps approach are therefore meant to illustrate how the informational content of the data-streams relative to C fluxes is enhanced once soil carbon stocks are more "realistically" modeled.

We have added in the text (§2.3.3): "... We did this to correct for the initialisation of the soil carbon **imbalance** following model spin-up and **illustrate how the informational content of the three data-streams relative to the surface carbon fluxes can be enhanced once soil carbon disequilibrium is more "realistically" represented".**

Regarding assimilation sequentiality:

1) We expect that the effects of the sequential approach are minimized by tuning mostly

one parameter (KsoilC) in the first step;

2) Our paper does not focus on the pros/cons of the stepwise vs simultaneous approaches and all the more on pointing the benefits of the simultaneous over the stepwise. Other studies have discussed this in more detail (Richardson et al., 2010; Kaminski et al., 2012; MacBean et al., 2015; Peylin et al., 2016). Our work mostly assesses the different C budgets that can be obtained within a simultaneous assimilation framework depending on the combination of the datasets used, and compare the results to one benchmark product which has been obtained with a stepwise approach (and the assimilation of the same datasets).

We offer the possibility to discard the results related to the 2steps approach (although mostly illustrative), and also to the stepwise, if it is too confusing.

- **The approach used to specify the prior parameter uncertainties seems backwards and clearly represents double dipping (using the data being assimilated to specify the priors). Priors are supposed to represent the information you have about parameters BEFORE the data being assimilated is seen. I think if you want to use model outputs to put constraints on priors (e.g. the emergent constraints stuff Mat Williams has done with CARDAMOM) those constraints need to come from different data.**

The approach of Desroziers was precisely developed following the observation that the definition of *a priori* error statistics based on expert knowledge (i.e. how to set the parameter error statistics before assimilation) was flawed. The test actually is used for checking the consistency of both **B** and **R** (and their relative weight), but by no means allows to tune the *a priori* errors by itself. This is still done by users based on the analysis of the different metrics. This consistency test ultimately provides an enhanced expert knowledge on the different error statistics and on the *a priori* model-data mismatch. Because characterizing the **B** error covariance matrix is difficult and is a problem that is shared among all TBMs and data assimilation systems, we believe that the implementation of the Desroziers test (commonly used in the atmospheric DA community) would benefit a broad research community.

Note also that (as suggested by the Reviewer at point #21 below) one way would be to include the prior parameter uncertainties in the cost function itself and thus optimize both the parameters and their uncertainties at the same time as was done for atmospheric inversions by Michalak et al., (2005). However, this complexifies substantially the inverse problem (in particular, we discuss about the limitations in computational time in our answer to point #21) and our approach is thus more practical although indeed there is to a certain level some double dipping.

- **Results and Discussion section is WAY TOO LONG AND REPETITIVE. Section 4, which is just more Discussion, should be merged into the Discussion.**

We have followed the reviewer's suggestion and have restructured sections 3 to 5 to make them more concise and avoid redundancy as much as possible. We believe this has also allowed clarifying the take home messages of the paper.

We detail below the main changes operated:

- Section §3:

- We have changed the title "Results and Discussion" to "Results"
- The discussion on the partitioning of the land C budget between tropics and northern extra-tropics has been moved from §3.2 to §4
- The part of the analysis of the influence of each data-stream performed in §3.4 focusing on the discrimination between PFT / atmospheric stations has been moved to the supplementary materials (supplementary text S7).

- Section §4
  - We have changed the title "Summary and Outlook" to "Discussion"
  - We have switched §4.2 ("Caveats and perspectives concerning the initialisation of the soil carbon pools ") and §4.3 ("Realism of the regional to global-scale C fluxes")
  - Section §4.2 ("Realism of the regional to global-scale C fluxes") has been reorganized and now includes the discussion on the partitioning of the land C budget between tropics and northern extra-tropics.
- Section §5
  - We have moved what was the first paragraph of §4 here.

Specific points:

- **L34: "also given the technical challenges" feels tacked onto the end of the sentence. Not really explained and doesn't really provide any information. Either explain or drop**

We understand the Reviewer's point of view. However, this end of sentence is only provided in the abstract where a justification is limited due to the word limit, and also because this is not really central information in our study. As the related arguments are developed further in the Introduction section, we believe that this part can remain as it is, unless the Reviewer strongly objects.

- **L97: This statement should acknowledge (cite and discuss) Trevor Keenan's "Rate my data" work, which took a really deep dive into the value of multiple constraints (albeit at a single site)**

L97 is intended to be rather generic and the associated explanations and references are indeed developed in the following paragraph. The work of Keenan et al. is actually referenced later in L107 (and also before L89…), together with other studies which have assessed the benefit of assimilating multiple data-streams.

- **L107: on the issue of Likelihood weights, you should take a look at Oberpriller's paper http://doi.org/10.1111/ele.13728**

We thank the reviewer for suggesting this paper of Oberpriller et al. (2021) which is relevant study to cite with respect to both the weighting of assimilated data-streams but also the impact of model-data bias on the optimisation of the model parameters and the reliability of the resulting model predictions.

We have added a reference to Oberpriller et al. (2021) near L107 and have also slightly modified the text:

"Both approaches however face similar challenges, like defining the model-data uncertainty and **(see, e.g., Richardson et al., 2010; Keenan et al., 2013; Kaminski**

et al., 2012; Bacour et al., 2015; Thum et al., 2017; Peylin et al., 2016) hence the weight of that each dataset has on the optimization outcome **(although specific weighting approaches may be envisioned, as in Wutzler and Carvalhais et al. (2014) or Oberpriller et al. (2021)).**"

We have also added a reference in the following sentence.

- **L113: First, you raise the issue of systematic errors, but then the approach you use only accounts for random I.I.D Gaussian errors, so I'm not sure why you're bringing this up. Second, see papers by Istem Fer and Marcel Van Oijan (separately, not coathors) for examples of the formal accounting of systematic error during calibration**

The majority of data assimilation studies in our TBM community rely on the assumption of Gaussian errors which implicitly neglect any possible model-data biases (i.e. bias-blind assimilations). The point here was precisely to recall this strong assumption which may have a detrimental impact on the optimisation when it is not met. In addition, as our study also addresses this point of model-data bias (wrt atmospheric CO_2 data), we believe it is important to introduce this aspect.

We acknowledge that the way the model-data bias is accounted for in our study is not optimal. We aim to improve its treatment in the future (wrt atm. CO_2 data, not only by performing a more consistent initialisation of the soil C pools with an extended transient run, as it is explained in answer #3) and will therefore investigate more rigorous approaches. We would therefore be very much interested in these papers, if the Reviewer could provide us with the references.

- **L165: Why was a linear relationship between NDVI and FAPAR assumed (and what was this calibrated against?). A quick google search turns up lots of papers that suggest a nonlinear, convex relationship between these two variables.**

The linear assumption between NDVI and FAPAR holds for most biomes as demonstrated in Myneni et al. (1994) or Fensholt et al. (2004). FAPAR and NDVI are normalized in our observation operator (as in Bacour et al. (2015) or MacBean et al. (2015), hence no calibration required) which limits the impact of non-linearities. This is something we did not indicate in the text and have since corrected. Also, because NDVI data only constrained phenology parameters in our set-up, any non-linearity between NDVI and FAPAR would have a negligible impact on the results.

We have corrected the text by providing the reference to Myneni et al. (1994) where the linear relationship is mentioned, and moved the reference to MacBean et al. (2015) later. We have also added "**In addition, we consider normalized data in our assimilation scheme**" after the description of our observation operator for NDVI.

- Myneni, R. B., & Williams, D. L. (1994). On the relationship between FAPAR and NDVI. *Remote Sensing of Environment*, *49*(3), 200-211.
- Fensholt, R., Sandholt, I., & Rasmussen, M. S. (2004). Evaluation of MODIS LAI, fAPAR and the relation between fAPAR and NDVI in a semi-arid environment using in situ measurements. *Remote sensing of Environment*, *91*(3-4), 490-507.

- **L198-207: how did you account for the uncertainties introduced via all these other data products?**

We have used the same set-up as in Peylin et al. (2016) to prescribe the model (ORCHIDEE and LMDz)-data errors when assimilating atmospheric CO_2 data (see L350). The difficulty to characterize the model uncertainties associated with all model inputs is general to both the atmospheric transport model (point of the reviewer) and to the land surface model (vegetation map, meteorological forcing, etc.). Our approach to characterize the error covariance matrix, although imperfect (it does not account for error correlations), aims at including all these individual errors in the global error budget.

- **L213: "locations are"**

We have corrected the text accordingly.

- **L253: misfit: this is more commonly called a cost function**

We choose to use the term introduced by Tarantola (1987 and 2005) because the function really quantifies the misfit between the model and the data sets.

The term is not so unusual in the literature, as for instance (among others):

- Evans, G. T. (2003). Defining misfit between biogeochemical models and data sets. Journal of marine systems, 40, 49-54.
- Scholze, M., Kaminski, T., Rayner, P., Knorr, W., & Giering, R. (2007). Propagating uncertainty through prognostic carbon cycle data assimilation system simulations. Journal of Geophysical Research: Atmospheres, 112(D17).
- Vo, H. X., & Durlofsky, L. J. (2015). Data assimilation and uncertainty assessment for complex geological models using a new PCA-based parameterization. Computational Geosciences, 19(4), 747-767.
- Cameron, D. A., & Durlofsky, L. J. (2014). Optimization and data assimilation for geological carbon storage. Computational Models for CO2 Sequestration and Compressed Air Energy Storage, 357-388.

- **L260-61: semantics note: observation error doesn't include model structural error, those are different concepts**

We do not agree with the reviewer here. The inclusion of the model parameterization errors in the **R** matrix (which is commonly designed as the "observation error covariance matrix" - or using a similar terminology (Dee et al., 2000; Bouttier and Courtier, 2002; Scholze et al., 2017)) is widely reported in the literature (Bouttier and Courtier, 2002; Rayner et al. 2005; Sacks et al., 2007; Xu et al., 2006; Knorr et al., 2010; Kaminski et al., 2012; Scholze et al. 2017) when modelization uncertainties can be described using Gaussian statistics (Tarantola, 2005), which is consistent with our assumptions.

- Bouttier, F., & Courtier, P. (2002). Data assimilation concepts and methods March 1999. *Meteorological training course lecture series. ECMWF*, *718*, 59.
- Dee, D. P., & Todling, R. (2000). Data assimilation in the presence of forecast bias: The GEOS moisture analysis. *Monthly Weather Review*, *128*(9), 3268-3282.
- Kaminski, T., Knorr, W., Scholze, M., Gobron, N., Pinty, B., Giering, R., & Mathieu, P. P.

(2012). Consistent assimilation of MERIS FAPAR and atmospheric CO 2 into a terrestrial vegetation model and interactive mission benefit analysis. *Biogeosciences*, *9*(8), 3173-3184.
- Knorr, W., Kaminski, T., Scholze, M., Gobron, N., Pinty, B., Giering, R., & Mathieu, P. P. (2010). Carbon cycle data assimilation with a generic phenology model. *Journal of Geophysical Research: Biogeosciences*, *115*(G4).
- Rayner, P. J., Scholze, M., Knorr, W., Kaminski, T., Giering, R., & Widmann, H. (2005). Two decades of terrestrial carbon fluxes from a carbon cycle data assimilation system (CCDAS). *Global biogeochemical cycles*, *19*(2).
- Scholze, M., Buchwitz, M., Dorigo, W., Guanter, L., & Quegan, S. (2017). Reviews and syntheses: Systematic Earth observations for use in terrestrial carbon cycle data assimilation systems. *Biogeosciences*, *14*(14), 3401-3429.
- Sacks, W. J., Schimel, D. S., & Monson, R. K. (2007). Coupling between carbon cycling and climate in a high-elevation, subalpine forest: a model-data fusion analysis. *Oecologia*, *151*(1), 54-68.
- Tarantola, A. (2005). *Inverse problem theory and methods for model parameter estimation*. Society for industrial and applied mathematics.
- Xu, T., White, L., Hui, D., & Luo, Y. (2006). Probabilistic inversion of a terrestrial ecosystem model: Analysis of uncertainty in parameter estimation and model prediction. Global Biogeochemical Cycles, 20(2).

- **L263: notational consistency: why would you abbreviate the prior as chi instead of writing this term out in quadratic form like all the other terms?**

We agree with the reviewer this is confusing and have hence updated equations 1 and 2.

- **L272: missing a noun between "parameters" and "is calculated"**

Indeed, the term H^0_{ORCH} was missing (some mathematical symbols have disappeared when generating the PDF file before submission). We have corrected the text.

- **L273: Acronym TAF undefined**

We have added "Transformation of Algorithms in Fortran" whose acronym is TAF.

- **L344: The method described for estimating R sounds ad hoc and seems to constitute double dipping (using data that is later assimilated to set priors). More to the point, why not just estimate the observation error parameters are part of the calibration like any normal statistical model?**

The definition of the observation errors based on the residuals between observations and the simulated quantities allows us to account, in a rather simple way, both for the error in measurements and in the model which dominates the error budget (Kuppel et al., 2013). The approach was justified in Bacour et al. (2015): For fluxes, the measurement error is usually small as compared to the model error, and has correlation structure that is negligible on a daily timescale (Lasslop et al., 2008). Model errors are rather difficult to assess and may be much larger than the measurement error itself: Kuppel et al. (2013) showed that the model error in ORCHIDEE dominates the error budget: for NEE for

instance, it is on the order of 1.5-1.7 gC/m²/day when the measurement errors is between 0.2 to 0.8 gC/m²/day (Richardon et al., 2008). It is however very difficult to properly estimate model errors and the use of the RMSD between the prior model and the observations usually provides a reasonable approximation (Kuppel et al. 2014). In addition, the application of the diagnostics of Desroziers allowed us to check that the observation error covariance matrix was consistent with the error covariance matrix on parameters.

As for the tuning of the observation error during the calibration process (as in Michalak et al., 2005, or Sacks et al., 2006; Renard et al., 2010), the large calculation times of ORCHIDEE (in particular when it is coupled to LMDz for the global scale simulations) precludes exploring this approach (at least with our set-up and the choice of the data that are assimilated). The problem of equifinal solutions and risks of convergence issues would also be increased with many other parameters (at least one for each site/pixel/station considered) to optimize.

- Lasslop, G., Reichstein, M., Kattge, J., & Papale, D. (2008). Influences of observation errors in eddy flux data on inverse model parameter estimation.*Biogeosciences*, (5), 1311-1324.
- Michalak, A. M., Hirsch, A., Bruhwiler, L., Gurney, K. R., Peters, W., and Tans, P. P.: Maximum likelihood estimation of covariance parameters for Bayesian atmospheric trace gas surface flux inversions, J. Geophys. Res., 110, D24107, https://doi.org/10.1029/2005JD005970, 2005.
- Renard, B., Kavetski, D., Kuczera, G., Thyer, M., and Franks, S.W.: Understanding predictive uncertainty in hydrologic model-ing: The challenge of identifying input and structural errors, Water Resour. Res., 46, 1–22, doi:10.1029/2009WR008328, 2010.
- Sacks, W. J., Schimel, D. S., Monson, R. K., & Braswell, B. H. (2006). Model□data synthesis of diurnal and seasonal CO2 fluxes at Niwot Ridge, Colorado. Global Change Biology, 12(2), 240-259.

- **L349: I can't figure out what was actually done here based on this description**

We agree with the reviewer that the sentence was not clear enough and we have thus corrected the text to :

"For atmospheric CO_2 measurements, we followed a different methodology given the large discrepancy in the modeled *a priori* concentrations **with respect to the observed concentration (i.e., large bias that increases over time due to biases in the land net carbon sink (too small))**. **The errors were determined at each site as the standard deviation of the observed temporal concentrations (Peylin et al., 2005, 2016)**, **to capture the general feature that model-data mismatch is likely large for sites and months with large variations in daily concentrations. Although crude, such an hypothesis has been used in many atmospheric CO_2 inversions and in our case it combines all structural** errors of the terrestrial **and transport** model**s**."

- **L395: (1) "large value of expresses a" seems to be missing a word or two. (2) a strong underestimation of observation error seems to be a problem that needs to be addressed more.**

Indeed, a symbol is missing again here. We have corrected the text to: "while the large value of CHI^2 expresses…" . The reason why we deliberately underestimated the

observation error for atmospheric CO_2 data is actually discussed in the following sentence.

- **L417: In optimization, don't we expect the norm of the gradient to be 0?**

In ideal configurations (Gaussian error distributions, error statistics on model parameters and observations perfectly described, linear model, optimization algorithm not trapped in local minima), this is true that the norm of the gradient should approach 0 at the solution. For such complex problems this is not achieved in practice, and even for the case of assimilation with atmospheric transport models which are more linear than terrestrial biosphere models (Chevallier et al., 2007).

The main issue with our optimisation scheme using a gradient descent algorithm is rather that the solution corresponds to a local minimum of the misfit function.

- Chevallier, F., Bréon, F. M., & Rayner, P. J. (2007). Contribution of the Orbiting Carbon Observatory to the estimation of CO2 sources and sinks: Theoretical study in a variational data assimilation framework. *Journal of Geophysical Research: Atmospheres*, *112*(D9).

- **L422: Small notational preference: can you just call this RMSE like everyone else?**

The term RMSE is not exclusively used by literally "everyone" in the research field and RMSD is also commonly used. Given that the metric is used to quantify the mismatch between model simulations and observations which have their own uncertainty, we prefer to use RMSD over RMSE (which implies an error-free reference data). The definition of this metric is provided in the text to avoid confusion for the reader.

- **L435: This statement might benefit from adding qualifiers about this being difficult to estimate under the linear tangent optimization approach you are using. Other approaches to Bayesian computation don't necessarily have this restriction.**

We have added "**for complex process-based terrestrial biosphere models**" as other Bayesian approaches that do not face this issue are not easily applicable to TBMs from a technical point of view (computational expenses, although emulators are being developed to alleviate this limitation).

- **L483: (1) undefined acronym. (2) This feels like the addition of new Methods in the Results – please go back and mention this in the Methods. (3) I think the paper would benefit from a figure that actually shows these seasonal cycles, and the model's errors in replicating various parts of these cycles and trends, since they're mentioned so often. I was hoping that was what Fig 2c was going to be, but not so much.**

1) Actually, we did not find any definition to CCGCRV, even in the reference web site (https://gml.noaa.gov/ccgg/mbl/crvfit/crvfit.html)

2) We choose to provide all information on the processing of atmospheric CO_2 time series in the supplementary materials rather than in an additional section in the Methods of the main paper.

3) We have followed the reviewer's suggestion and added in the supplementary materials a figure that compares the observed atmospheric CO2 time series to the model simulations (prior and optimized values for different experiments - for the raw data and de-trended data) at a few illustrative sites.

- **L529: Is this higher improvement due to the double dipping?**

This higher agreement with respect to NEE and FAPAR data, as discussed later in the paper in §4.2, is due to the correction of the bias in the modeled trend in the first step of the two-step approach (where the reviewer sees some double dipping) which ultimately allows a stronger correction of the model parameters related to photosynthesis, respiration, phenology, in the second step. We have added a reference to §4.2 in the text.

- **L550: "ORCHIDEE led with LMDz to overestimates…" I had to read this multiple times to figure out what you were trying to say. Even now, it only makes sense to me if I mentally cut out "with LMDz"**

We agree with the reviewer that the sentence is not clear. We wanted to point out that it is the coupling of ORCHIDEE (which simulates the surface fluxes) with LMDz (atmospheric transport) that permits the calculation of the trend in atmospheric CO_2 concentrations. We have reformulated the sentence to "the fluxes simulated by ORCHIDEE and transported by LMDz overestimate" .

- **L556: undefined acronym**

This is a mistake indeed. The definition of the acronym was provided later in L568. This has been corrected.

- **Pg 20: seems a bit bold to dive into the tropics vs mid-latitude debate given all the uncertainties in this analysis. But if you do, I'd recommend including Schimel et al 2015 https://doi.org/10.1073/pnas.14073021**

We have added a reference to this work: "Conversely, TBMs estimate a larger C sink over the tropics (Ahlström et al., 2015; Sitch et al., 2015), **possibly due to CO_2 fertilization (Schimel et al., 2015)**, than the inversions, which estimate an approximately net neutral C sink (Peiro et al., 2022)"

- **L638: what do you mean that you include all the experiments in the GPP posterior? Each experiment should have it's own posterior and you haven't discussed any sort of Bayesian Model Averaging to be able to combine them**

We agree with the reviewer that the sentence is not clear. We just meant "over all

assimilation experiments". We have removed the sentence.

- **L799-802: Why is this not feasible? How would you make it more feasible?**

The point of the reviewer addresses the propagation of the errors on parameters from one step to the other in a stepwise assimilation approach. We discussed this point earlier in answering the first major comment of the reviewer. In addition, we have now provided some elements in the introduction on the expected difference between simultaneous and stepwise approaches. Therefore, we have changed the text here to: "However, given that this is **difficult** in practice, **and** because of model non-linearities, stepwise/joint approaches lead to different optimized models **(Kaminski et al., 2012; MacBean et al. 2016)**".

- **L867: yes, but WHY?**

We have addressed previously the question of the space-time coverage of the data assimilated in this study. The aim of section 4.3 (including L867) is precisely to clarify the scope of our study which is more about investigating the challenges in simultaneously assimilating different data-streams (whatever their temporal coverage) into complex terrestrial biosphere models to optimize regional to global C fluxes that achieving an up-to-date re-analysis of the global of the C cycle with recent observational data. Actually, our study is what should be done before any such re-analysis, which will be the scope of another work.

We have updated the text here as follow (accounting also for the restructuring of the discussion section):

"... **Especially since we focused on a limited dataset both in terms of temporal coverage** (no atmospheric CO_2 data nor satellite data after 2010, no *in situ* flux data beyond 2007) **and of informational content. Indeed** we did not assess the potential of other data that can bring relevant **(and possibly more direct)** additional **constraints** on the dynamics of terrestrial carbon **stocks and** fluxes and stocks, such as aboveground biomass (Thum et al., 2017) or Solar Induced-Fluorescence (Bacour et al., 2019) which have already been investigated with ORCHIDAS, and with an updated version of the ORCHIDEE model. **The expansion of the assimilated datasets will be the subject of future work.**"

We also recall that we already clarified the main objectives of the paper in the introduction.

- **L869: missing a word in "fluxes stocks"**

Thank you for pointing out this error. We have corrected the text to "fluxes and stocks".

- **Interesting that the Discussion doesn't really mention how other teams are addressing some of the same problems you are struggling with (CARDAMOM, DART, PECAN, etc)**

We added in the second paragraph of the introduction a reference to the DART and PECAn assimilation frameworks:

"Since the first global scale Carbon Cycle Data Assimilation System (CCDAS) […], **and in parallel to the development of community assimilation tools (as DART (Anderson et al., 2009) or PECAn (Dietze et al. (2013))**, other modeling groups have developed their own global scale carbon cycle DA systems, in particular for ORCHIDEE…"

- Dietze, M. C., Lebauer, D. S., & Kooper, R. O. B. (2013). On improving the communication between models and data. *Plant, Cell & Environment*, *36*(9), 1575-1585.
- Anderson, J., Hoar, T., Raeder, K., Liu, H., Collins, N., Torn, R., & Avellano, A. (2009). The data assimilation research testbed: A community facility. *Bulletin of the American Meteorological Society*, *90*(9), 1283-1296.

However, we believe that the approaches by CARDAMON, DART, PECAN, are not directly comparable as they do not use the same type of process-based global land surface model with the same set of observations and in particular with atmospheric CO_2 concentrations that impose to run / optimize the model globally. CARDAMON uses a much simpler C cycle model which allows very different solutions to comprehensively assimilate various data-streams. These solutions are not easily implementable with our global TBM and to our knowledge none of the global TBMs team (i.e. those models participating for instance to the TRENDY model inter-comparison for the Global Carbon Budget) have currently assimilated the three data streams proposed in this paper in a simultaneous approach.

- **Figure 6 is completely illegible. This font size is WAY too small.**

Point noted. We have increased the font size to make it more readable.