Review #1

Evaluating the overall quality ("general comments"),

The authors conduct a twin model experiment to test the ability of two competing parameter estimation techniques (eta-4DVAR vs. 4DEnVar) to constrain 54 parameters within the land model ORCHIDEE. The authors use synthetic observations to generate atmospheric CO2 station data in which to retrieve the 'true' parameters from prior parameter distributions. The authors find that the 4DEnVar approach performs the best in terms of the RMSE statistic of global NBP and in terms of the posterior parameter values relative to the true values. The authors claim this demonstrates strong potential for 4DEnVar to be used with real data, and should be widely applicable to other land surface models.

We would like to thank the reviewer for taking the time to read through and comment on this manuscript, comments that will strengthen the paper and help clarify key points of the paper.

This reviewer found the topic relevant to the current state of earth system science which requires a wide range of observations to calibrate model performance to improve forecasts/projections. The potential for such a problem to be ill-constrained and suffer from equifinality stood out to this reviewer given the use of a single data set (atmospheric CO2) to constrain a multi-dimensional problem. This author recommends a more nuanced discussion of equifinality for this application both in the twin model experiment and for potential applications of using real data (see scientific questions below). This reviewer also would have appreciated a better description of how the parameter values were sampled (perturbed) from their respective distributions – and also to what extent the authors could have presented their results for the 4DEnVar using parameter distributions rather than point estimates. It was somewhat surprising how well the posterior parameter values improved the global distribution of NBP, given the limited range of the CO2 station footprint. More discussion related to the distribution of land PFTs relative to the CO2 station footprint may have been helpful to aid this discussion.

We thank the reviewer for highlighting these very important points, which we have addressed by improving existing figures, adding a new one, and revising or expanding the text accordingly. Please see the specific changes in the manuscript, as detailed in our response under each reviewer's comments. We agree that the

inherent equifinality of such data assimilation problems was not properly addressed in the discussion. We have added a paragraph to the discussion to address this issue and believe that this comment strengthens the article.

Individual scientific questions/issues ("specific comments")

(Please note that changes to the text are referenced by the line number in the new manuscript.)

Lines 45-50: If you calibrate against CO2 observations only, and do not adjust for model biases in land carbon pools – how accurate can your projections/forecasts of the carbon cycle be?

Thank you for raising this important point. While it is true that calibrating solely against atmospheric CO2 observations may not fully address biases in land carbon pools, it still provides a valuable global constraint on carbon fluxes. This approach ensures that the overall carbon budget is consistent with observed atmospheric CO2 concentrations, which is crucial for accurate projections and forecasts of the carbon cycle. To further improve the accuracy of our model, we acknowledge the importance of incorporating additional data sources integrating remote sensing data and in-situ measurements of land carbon pools to better constrain and reduce model biases. This multi-faceted approach would enhance the reliability of our projections and provide a more comprehensive understanding of the carbon cycle dynamics.

In this study, we assume that performing a complete TRENDY simulation, as described in Section 2.1.1, provides a realistic estimate of carbon sinks. However, we would like to emphasise that the main focus of this article is methodological - we present a two-year twin experiment designed to evaluate the feasibility and performance of the assimilation framework using atmospheric CO_2 data.

In the case of assimilation of real observations, the ideal solution would be also to include a pre-assimilation phase in the assimilation, i.e. the complete spin-up + transient simulation, to monitor the effect of the initialisation of the land carbon pools in the atmospheric CO_2 concentration. However, given the cost of such simulations, this may not be feasible. Another potential solution is to run a small transient simulation with the modified parameter before the assimilation windows. For example, when assimilating the year 2000-2005, start the simulation in 1995

until 2005 and only consider the last five years (2000-2005). This way, the carbon reservoir would be modified and influenced by the parameter. It would also allow for better monitoring of the effect of long-term changes and thus increase confidence in the projection.

Line 95: "We demonstrate the potential of 4DEnVar using synthetic observational data and compare its performance with that of 4DVar with finite differences."

Your criteria for testing the differences between the methods is vaguely stated here. Are you judging success based on which method best identifies the 'true parameters'.

We added more detail in L96:

> We demonstrate the potential of 4DEnVar using synthetic observation data **according to different criteria: i) the differences between synthetic observation and simulation of atmospheric CO2 concentration, ii) the spatial distribution of carbon flux as well as their subcomponent, and iii) the recovery of the true parameters used to generate the synthetic observation. We also compare the performance of 4DEnVar using these criteria** with that of 4DVar with finite differences

Lines 135:140: Given the use of pre-calculated transport fields that relate atmospheric concentrations to surface fluxes, you do not make use of a dynamic atmospheric ensemble to generate this relationship based upon actual atmospheric forcing. This seems a bit like using a background climatology to get the concentration/surface flux relation. Also the land seems to be decoupled from your atmosphere – in the sense the dynamics that drive the CO2 concentration/flux relationship is not the same as the actual met forcing driving the land model. This is perhaps not as important given the authors are generating 'perfect' obs, but during implementation for real parameter estimation should this not have an important impact?

The land model is decoupled from the atmosphere, and ORCHIDEE is run offline using meteorological forcing files from a reanalysis of ERA-Interim data every 3 hours. The use of reanalysis meteorological fields is important in order to better respect the temporal dependence of meteorological events. The pre-calculated transport field is generated using the LMDZ model, which is driven by winds from

the same reanalysis meteorological data. This is not the same as using climatological data, as the calculation is performed using reanalysis winds for the same years as those of the simulations. The use of these precalculated transport fields reduces the calculation time when running the same atmospheric simulation several times by only changing the surface fluxes. We add more clarification by adding in L141:

> **The calculated winds (u and v) used to drive LMDZ are provided by ERA-Interim reanalysis meteorological data in order to realistically account for the temporal dependence of meteorological events.**

and L151

> **Nevertheless, it is important to note that the use of these pre-calculated transport fields does not allow for the evaluation of dynamic feedbacks between the surface and the atmosphere that may occur due to parameter changes.**

Line 150-155, Figure 1: One would expect that the CO2 sites were also chosen such that land surface areas sensitive to atmospheric CO2 also coincide with the range of land surface PFT types in this analysis. Any consideration of this?

Line 280-85 – Same question as before, these sites were chosen to be sensitive to land surface fluxes, however, do they provide good sampling of the most important PFTs? Sampling of North America and South America look poor. Sampling of Africa does not include the tropics at all.....

In this study, the authors considered the pre-calculated transport fields for a handful of actual atmospheric stations (21). The locations of the stations are therefore determined by the actual observation networks. It is true that some PFTs are undersampled as there are more stations in northern latitudes, mainly in America and Europe, than in the rest of the world. We selected the stations based on their sensitivity to continental fluxes. Although the number of stations is not ideal, it allows us to monitor terrestrial carbon fluxes. The other stations were mainly located over the ocean and would have been less affected by changes in the terrestrial component given the chosen assimilation window. Only the TrBE (Tropical Broadleaf Evergreen forests) PFT, mainly found in the Amazon rainforest and

Central Africa, and the BoND (Boreal Deciduous Needleleaf forests) , mainly found in Siberia, appear to be less observed. We believe that this is also related to the results presented in Figure 7 for Epsilon 4Dvar. Indeed, we see from the well distribution that the largest differences appear in these two regions, which could be explained by the fact that they are less monitored.

We have added the following text to the manuscript L162

> **The selected stations also provide a good overview of most PFTs. However, as we can see in Fig. 1, two PFTs appear to be less sensitive to the selected stations: TrBE, which is mainly found in the tropical forests of Amazonia and Central Africa, and BoND, which is mainly found in Siberia.**

L486

> **We believe that the different spatial structure obtained by e-4DVar against the synthetic net carbon flux could be explained by the fact that the two PFTs TrBE and BoND are not well monitored, creating a dipole in the Amazonian and Siberian regions to compensate for the incorrect carbon flux corrections in other regions.**

Line 286: How was your prior parameter distributions chosen? From the uniform distribution shown in the figures where you only show upper and lower bounds? Or from a normal distribution as described in Equation 14 and 15?

The new a priori parameter vector was manually perturbed in order to find a new set of parameters giving a simulation of the a priori atmospheric concentration consistent with the observations or simply realistic. We added to the text L296:

> **A new vector of a priori parameters was generated manually, ensuring that it differed from the "real" parameter values while retaining physically meaningful values.**

Figure 3: Showing the ensemble mean parameter behavior doesn't give any information on the ensemble distribution. Maybe I am misinterpreting the implementation of the 4DEnVar method, but can't you show this in terms of the true, prior and posterior *distributions* instead of the ensemble mean behavior?

Figure 6: Same question as before – can you convey this information in terms of distributions (histograms)?

Figures 3 and 6 have been modified to display the standard deviation of the posterior ensemble as an error bar that can be considered as the posterior uncertainty. We prefer not to use histograms, as all distributions are Gaussian. The mean and standard deviation are therefore sufficient to visualise the ensemble. The use of histograms is very useful, but would make the figure difficult to read, particularly in the complex case involving 57 parameters (which would result in 57 histograms). The legends have been modified as well.

We added in L274:

> **A posterior ensemble can be generated as it is described by Douglas et al 2025 by calculating Xa' where**
>
> $$X'_a = X'_b(I + (HX'_b)^T R^{-1} - 1HX'_b)^{-\frac{1}{2}}$$

and in L463:

> **Furthermore, Fig. 3 shows a significant decrease in the standard deviation of the posterior ensemble. This allows us to identify which parameters and therefore which PFTs appear more sensitive. In this case, it seems that the results for the TrNC3 and Crops C4 PFTs are the most uncertain.**

and in L473:

> **The posterior ensemble generated for the 4DEnVar also shows a reduction in uncertainty for all parameters. This uncertainty reduction is not equal for all parameters - a maximum reduction can be seen for the Q10 parameter(reducing the standard deviation by 94 %) and the lowest for the less sensitive $m_{maint.resp}$ parameter (with a 14% reduction for the NC4 PFT).**

Figure 6: (SLA panel) Why do most of the prior values for the parameters all start at the same value? Were they not being perturbed independently?

Indeed, the parameters were not perturbed independently; the new set of *a priori* values used in data assimilation was derived from the 'actual' *a priori* values of the model which may be identical across different PFTs. We applied the same perturbation for these PFTs.

Table A1: What does proportion mean in this context?
Line 430: As far as I can tell, it is still not defined what the 'proportion' of the parameters is. Is it based on global land area coverage, or land coverage that coincides with spatial footprints from your chosen network? These could be two completely different things. Did you do any comparison of the MAD statistic based on % of land area covered by the CO2 network spatial footprint? Are they strongly related? Would be nice to see a land surface map with PFT distribution.

Plant functional types (PFT) in ORCHIDEE and acronyms used in this study as well as their proportion

The proportion is the Global Cover Fraction by the PFT, the remaining proportion being Bare Soil. The title of the Table has been modified to reflect that.
The PFT maps used in this study are available here :
https://orchidas.lsce.ipsl.fr/dev/lccci/orchidee_pfts.php

Table A2: Is the partial derivative averaged over space and time? Therefore the 4dEnvar itself doesn't account for any seasonality (time-variation) in the relationship?

The partial derivative in Table A2 is indeed averaged in space and time; it is used for the e-4DVar approach to select the "best" epsilon values, It is not used with 4DEnVar. In both approaches, assimilation is performed over the entire two-year time window. The temporal variation is therefore taken into account. We have added this clarification to the text of Table A2 and Figure A2.

> **Spatial and temporal average** of the partial derivative for all parameter for each PFT

Figure 7: I was surprised at how well the True – Posterior 4dEnvar net carbon flux (top right panel) performed given the limitation of the spatial footprint influencing the station CO2, thus informing the biogenic contribution to CO2. This is promising, but be aware, that the ability to match the net carbon flux gives no guarantee that the component fluxes are well simulated. An interesting complement to this plot would be to compare the true component fluxes of GPP and ecosystem respiration against the posterior component fluxes of GPP and ecosystem respiration for the complex case.

We agree that the performance of True - Posterior 4dEnVar was significant on both net and gross (not shown in the first version of the paper) carbon fluxes. We have now included in the appendix the analysis on GPP (figure A4) which shows consistent quality of the fit of 4DEnVar posterior to the True observation.

We believe that this result is mainly due to the fact that we are in a twin experiment with no model-data bias. In this case, the model is considered 'perfect' and can fully recover the synthetic observation, which we believe greatly simplifies the problem. We also believe that the poorer performance of e-4DVar is mainly due to the fact that 1) the method is more sensitive to local minima, 2) the estimation of the tangent linear model relying on a finite difference approach is more uncertain (strongly dependent on the value of epsilon),and 3) The BoND PFT does not change because the PFT is not sufficiently sensitive and is therefore compensated by other PFTs/regions.

However, both of these problems seem to be solved by 4DEnVar. The approach is less sensitive to local minima because it generates an ensemble that is less affected by local minima. We added the figure of the GPP estimates in the appendix (Figure A3) and added this text in discussion L491

> **Fig. A3 shows the differences in spatial distribution of gross primary production (GPP) between the "synthetic" fluxes and the prior/posterior estimate of the two methods, as well as their global yearly budget. We can see that GPP obtained with the 4DEnVar method is slightly better than the e-4DVar method for the global budget and better matches the spatial distribution of the synthetic flux.The e-4DVar method appears to compensate for the lack of change between the prior and posterior GPP across most of the Northern Hemisphere.**

and L497

This experiment of calibrating a large number of parameters represents a more realistic case, even if we consider **a very low** model/observation error. **The results demonstrate the good performance of 4DEnVar, which, even in a 'perfect' model situation, i.e. a model that can perfectly simulate observations, can assimilate observations while being less impacted by local minima. However, this may not be the case when using actual observations and introducing more complex modelling/observation errors.**

Line 431: I think a more nuanced discussion of equifinality is required here and/or in the Discussion. In addition to the pure number of parameters attempted to calibrate simultaneously – equifinality can arise for a number of different reasons – 1) a single parameter type being compensated within the large list of PFTs, 2) the station CO2 concentration is influenced through the NBP, which is a confluence of both photosynthetic and respiration processes, which can easily compensate for each other to provide a net biogenic carbon flux consistent with station CO2 data. I understand that this is a twin model experiment, so the following do not necessarily contribute here, but if this setup were to be applied to real data additional equifinality challenges present themselves including 1) the model state itself (carbon, water nutrient pools) have not been constrained by any data, thus parameters will compensate for biases due to model state initialization problems 2) The biogenic fluxes (controlled by parameters) would seem to contribute just a portion of the land-atmosphere carbon exchange which includes other large fluxes from fossil fuel, fires and ocean fluxes which would have to be measured accurately-- 3) atmospheric model transport errors, influencing the relationship between land carbon flux and station CO2 data.

We agree with the reviewer and propose adding the following text:

L507

In this twin experiment, both methods have to deal with the inherent equifinality of atmospheric concentration assimilation. This equifinality occurs when parameters compensate for each other, resulting in either an incorrect spatial distribution of NBP or inaccurate estimates of subcomponents such as GPP and total ecosystem respiration (TER), but

still allowing for a match with observations. Although both methods considered in this study successfully recovered the global budgets for NBP and GPP, the e-4DVar method did not obtain the correct spatial distributions of NBP and GPP (see Figures 7 and 8). This is not the case for the 4DEnVar method, which better recovered the 'true' spatial distributions of NBP and GPP. We believe that this equifinality could increase the number of local minima, further disrupting the performance of the e-4DVar method. We also believe that the ensemble nature of the 4DEnVar method provides a more comprehensive view of the parameter space, making it less sensitive to local minima and therefore to equifinality issues.

L553

The assimilation of real observations of atmospheric concentrations may also increase the equifinality mentioned in Section 4.1 for several reasons, such as: i) Incorrect initial conditions of the carbon pools, which can impact respiration; ii) Wrong estimates of other flux components, such as ocean or fossil fuel components; iii) Structural errors in either the land surface model or the transport model. The issue of incorrect initial conditions can be addressed by starting the simulation a couple of years before the assimilation window. This allows for the correction of the initial carbon pool and better accounts for the effects of the new parameter values on the carbon pool. To handle other components, such as ocean components, the same assimilation can be repeated using different estimates of the ocean flux. Ideally, an ocean model could be included in the optimization to calibrate both land and ocean components, as is done in atmospheric inversion. The advantage of the 4DEnVar method is that it only requires forward simulations. Therefore, no code adaptations are needed, making it easier to use different transport models. This should help detect and address structural errors. The equifinality can also be reduced by assimilating multiple data streams simultaneously, as done in Peylin et al. (2016) and Bacour et al. (2023), to calibrate both GPP and NBP at the same time.

Line 512: Given the significant challenges related to equifinality mentioned above, I am not sure this setup shows "great potential" to constrain parameters. I might be more realistic and state that it demonstrates that 4DEnvar shows more potential than eta-4DVar.

We continue to believe that the 4DEnVar method has strong potential. Mainly because the technical implementation of this is simpler than that of the standard 4DVar method and the size of the ensemble required is reasonable, with the methods giving satisfactory results (good reduction of the model observation mismatch). We believe that many of the reviewer's very relevant comments are inherent to atmospheric concentration assimilation and would therefore still be present with other assimilation methods. We mostly agree with the reviewer and have replaced "**great**" with **'good'** to be more moderate.

Purely technical corrections

Abstract:

Awkward: "These models rely on parameterisations that necessitate to be carefully calibrated". These models rely on parameterizations that require careful calibration.

Thank you for the suggestion; We have applied this correction.

Introduction:

Can you describe in terms more accessible to the general community what isotropic means in this context?

"corrections to CO2 surface fluxes are isotropic in time and space."

We added in the text L32

> This statistical optimisation generally assumes that the corrections to CO2 surface fluxes are isotropic in time and space. **This suggests that errors in surface fluxes are only correlated in space by distance between points, and not by direction. Furthermore, these errors are not strongly correlated in time.**

Line 115: I wouldn't use the terminology 'assimilation' routine to describe photosynthesis or carbon uptake routine. Assimilation is often used within data assimilation context, a component of this analysis, which is not what this is describing.

We agree with the reviewer and have changed 'The carbon assimilation' by 'The carbon fixation'.

Section 2.1.4. I think it's also worth mentioning that you didn't optimize the prior biogenic fluxes by constraining them with carbon pool observations (LAI, biomass, soil carbon etc).

We presented this aspect in section 2.1.1 describing the spin up, transient and historical simulation carried out to equilibrate the soil carbon pool that are computed dynamically in ORCHIDEE, which also equilibrate other carbon pools such as the leaf area index (LAI) and biomass. We believe that, thanks to the other changes requested by the reviewer, we have presented these aspects more clearly.

Shouldn't Figure A1 include the land biogenic fluxes for the truth simulation, just for relative perspective? After all the parameter optimization is based on the influence of biogenic fluxes on the atmospheric CO2.

This figure is intended to show the other components of the global C budget that are used as input to LMDZ (described in §2.1.4) that have not been optimised in our study. We present the actual simulation of the net land biogenic fluxes in Figure 7. We believe that including these elements in Figure A1 could cause confusion as to what has been optimised in our framework and what has not. We therefore prefer not to include terrestrial biogenic fluxes in Figure A1.

Line 305: LAImax: The absolute max value that LAI can be? Can you clarify? Does this mean carbon cannot be allocated to leaf carbon once achieving this level?

LAImax is the maximum value that LAI can reach for each PFT, which stops allocation to the leaves once this value is reached. We added this clarification L323

Once the LAI reaches LAImax, no carbon is allocated in the leaf.

Line 444: LAI or LAImax ?

We have corrected the manuscript to add LAImax.