Review #2

While this manuscript addresses a highly relevant and important topic—adjoint-free data assimilation for land surface model parameter estimation using atmospheric CO_2 concentrations, there is a fundamental conceptual misunderstanding regarding the data assimilation framework employed. The authors repeatedly describe their method as a "4DEnVar" approach. However, after careful review of the methodology and experimental design, it is clear that the implemented framework aligns more closely with a 3DEnVar method rather than a true 4DEnVar. I will provide more specific comments below detailing the evidence for this classification error and offering suggestions for how to appropriately revise the manuscript.

We would like to thank the reviewer for taking the time to read and comment on the manuscript. We acknowledge the concerns raised regarding the terminology used to describe our method. However, we believe there may have been a misunderstanding about the nature of our approach. To clarify, our method is indeed a 4DEnVar and not a 3DEnVar, as it involves time assimilation over a two-year window covering all data.
We realise that our initial explanation may not have been sufficiently clear, and we apologise for any confusion this may have caused. In response, we have provided detailed replies to each of the reviewers' comments below and have revised the manuscript accordingly to improve clarity.
(Please note that changes to the text are referenced by the line number in the new manuscript.)

1. The authors incorrectly label their method as "4DEnVar." In classical data assimilation terminology, the key distinction between 3D and 4D variational methods lies in the incorporation of assimilation windows. A three-dimensional variational (3DVar or 3DEnVar) method assimilates observations as a function of space only,

without explicitly considering the time evolution of the model states. In contrast, a four dimensional variational (4DVar or 4DEnVar) method introduces a temporal dimension by defining an assimilation window, allowing the model to evolve dynamically and best fit the observations distributed across time within that assimilation window. In this manuscript, there is no explicit mention of an assimilation window, nor is there any evidence that the model trajectory evolves and interacts with observations at multiple times during a window. Instead, observations appear to be treated statically, consistent with a 3DEnVar framework. Therefore, the method used in this study should be accurately referred to as 3DEnVar, not 4DEnVar. The manuscript must be revised to correct this misclassification throughout, including the title, abstract, methods, results, and discussion sections.

We fully agree with the explanation given by the reviewer on the difference between the 3DVar and the 4DVar. However we would like to clarify that our approach does fall under the category of the 4DVar as we assimilate pseudo-observations both in space and time. Specifically, we are using the full 2 years of pseudo-observations within an assimilation window of 2 years. Furthermore, a temporal window is essential when calibrating parameters. Since parameters remain constant over time, we need to find the 'best' set of parameters that fits the observations over time. Assimilating a single time step would lead to different parameters for each time step. In order to clarify this point, we have added to L300

> **The assimilation of atmospheric CO_2 concentrations at the 21 stations is performed simultaneously using a two-year assimilation window in order to assimilate all observations and thus monitor variations in carbon fluxes in space and time, as shown in Fig. A4.**

Figure A4 has been added to illustrate the assimilated time series and clarify this point.

Given that we are using two years of observations and that temporal evolution is as

important as the spatial distribution of carbon fluxes, we need to incorporate time into the assimilation scheme and therefore rely on a 4DVar or 4DEnVar method.

2. The description at L56–58 defines general variational assimilation, not 4DVar specifically. 4DVar uniquely involves an assimilation window and time-evolving model trajectory. Please correct this definition.

We apologise for the lack of precision. We have therefore modified it L57 as follows:

> The 4DVar approach involves defining a cost function (which is usually based on a least-square criterion) that computes the difference between observations and model outputs **distributed in space and time** as well as a background term that accounts for prior knowledge of the parameters.

3. Data assimilation can generally be categorized into full-field assimilation and anomaly assimilation. Full-field assimilation adjusts the model state towards observed absolute values, while anomaly assimilation only incorporates the observed anomalies. The authors must clearly specify which approach is used. If full-field assimilation is applied, the impact on climatology should be explicitly evaluated and presented.

We thank the reviewer for raising this point. In our study, the model parameters are adjusted to better match the model output to the absolute values of the observations, rather than considering their anomalies. As such, the assimilation most closely match the definition of a full-field assimilation. This approach is consistent with the goal of our study, which is to evaluate the potential of using atmospheric CO_2 observations to

constrain land surface model parameters through 4DEnVar.

We would like to emphasise that this study is based on a controlled twin experiment, where the primary objective is to assess the method's ability to recover known “true” parameter values and to match synthetic observations. As such, we focus on demonstrating the feasibility and strengths of the approach itself, rather than evaluating its impact on climatological metrics in a real-world context. We agree that when assimilating real observations, it will be important to assess the impact on other aspects of the model outputs. However, since this study is methodological in nature and uses synthetic data, we believe that including such an evaluation would be beyond the scope and would not directly strengthen the main message.

4. The cost function shown in Equation (5) corresponds to the standard 3DVar formulation, as it lacks any temporal dimension or model trajectory integration. A true 4DVar cost function should involve the evolution of the model state over time:

Please revise both the equation and the method name accordingly.

While the reviewer is correct that the standard 3DVar formulation lacks a temporal dimension, here, y represents a vector of observations and H(x) the model output over the given window. Therefore, the assimilation over time is made implicit by this formulation.However, we recognise that there may be some confusion about this and have changed the manuscript L185.

> **Here, the model operator output H(x) and the observation y are defined in time and space. All observations are concatenated into a large vector of observations y, in order to represent all observations available in a given time window. The same operation is performed for the output of operator H(x).**

We have added the classic 4DVar Cost feature and explain our simplification in L200:

> the 4DVar cost function:

$$J(\boldsymbol{x}) = \frac{1}{2}(\sum_{t}^{N_t} \boldsymbol{\mathcal{H}}_t(\boldsymbol{x}_t) - \boldsymbol{y}_t)^T \mathbf{R}_t^{-1}(\boldsymbol{\mathcal{H}}_t(\boldsymbol{x}_t) - \boldsymbol{y}_t) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x_b})^T \mathbf{B}^{-1}(\boldsymbol{x} - \boldsymbol{x_b})$$

where t refers to time steps 0,...,N_t. Since the parameter must be constant over time, we consider only a single time window that includes all observation vector y (in time and space). We therefore simplify to the compact form the initial 4DVar cost to the compact form:

$$J(\boldsymbol{x}) = \frac{1}{2}(\boldsymbol{\mathcal{H}}(\boldsymbol{x}) - \boldsymbol{y})^T \mathbf{R}^{-1}(\boldsymbol{\mathcal{H}}(\boldsymbol{x}) - \boldsymbol{y}) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x_b})^T \mathbf{B}^{-1}(\boldsymbol{x} - \boldsymbol{x_b}).$$

5. Lines 310-315: The R matrix represents the observation error covariance (not both the model/observation error) and should account for spatial heterogeneity in observation uncertainty. Setting all diagonal elements to 0.01 ppm is overly simplistic. Even in a simple setup, R could be statistically computed based on the variance of the observation. Please consider a more realistic design or justify this simplification.

In our case, model structural errors include both the structural errors associated with the ORCHIDEE model for the computation of the net carbon fluxes and the transport model error. In this context the model errors are likely the dominant part of the R matrix, given that measurements of atmospheric CO2 are usually relatively precise. The R matrix represents errors linked to the comparison of y and H(x) excluding the model parametric error that is accounted for in the B matrix. Therefore, the error in the model structure and the observation operator are essentially taken into account here. so that R contains model and observation errors in this case, and B only parameter errors.

Furthermore, we agree that setting all diagonal elements to 0.01 ppm is simplistic, but since we are in a conceptual framework relying on twin experiments, we have considered the model to be perfect, allowing for a very low error to be used. We would also add that this observation error is close to the difference in atmospheric CO_2 data between observation and model simulations obtained after optimisation (especially for 4DEnVar), as illustrated in Figure 4. We have added the following to

the text L329:

> **Indeed, since no error was included in the pseudo-observation and as we are in a Twin experiment, a simplistic R matrix was used.**

We also added to the text some references and justification in L334.

> **The configuration of the R and B matrices was based on previous data assimilation studies with ORCHIDEE and a simplified carbon model (Kuppel et al., 2012 and 2013; MacBean et al., 2016; Peylin et al., 2016; Bastrikov et al., 2018). These studies employed diagonal matrices for R and B to assimilate in situ observations, while Peylin et al. (2016) specifically used them for atmospheric CO_2 observations.**

6. Lines 364-366: Please clearly define how RMSD is computed, and indicate whether systematic bias is removed prior to calculation.

we have clarified how we computed the RMSD in L386 :

> The results in terms of 1) mean reduction in the mean square difference (RMSD) **calculated between the pseudo-observation and the simulation over the two years of the assimilation window for the 21 atmospheric stations**,

We also added an Appendix named **Metrics calculation**:

> **The RMSD and MAD are calculated as follows:**

$$RMSD = \sqrt{\frac{1}{N_t}\sum_{t=0}^{N_t}(\boldsymbol{\mathcal{H}}(\boldsymbol{x}_*)_t - \boldsymbol{y}_t)^2},$$

$$MAD = \frac{1}{n_{param}}\sum_{i=0}^{n_{param}}|\boldsymbol{x}_{*i} - \boldsymbol{x}_{\boldsymbol{true}i}|,$$

where x* can be either xb or xa. The Pearson correlation coefficients were computed using the Numpy Python library with the 'corrcoef' function. The paired t-tests were computed using the 'stats.ttest_rel' function from the Scipy library.

7. The evaluation relies almost entirely on RMSD and MAD. Please consider providing additional diagnostic metrics, such as correlation coefficients between posterior and true fields, or skill scores, to offer a more complete picture of assimilation performance.

We agree with the reviewer that a range of metrics is necessary to evaluate performance. In addition to the evaluation of the NBP already in the manuscript, we computed to Pearson correlation coefficients of the NBP in time and in space and add this to the manuscript in L 442:

The Pearson correlation coefficient between the 'synthetic' NBP and the prior NBP is 0.87 in time and 0.17 in space. The posterior NBP obtained by the 4DEnVar method shows a Pearson correlation coefficient against the 'synthetic' NBP of 0.99 in time and 0.98 in space. In comparison, the posterior NBP obtained by the e-4DVar method has correlation coefficients of 0.98 in time and 0.84 in space.

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 for the global budget and better matches the spatial distribution of the synthetic flux. The e-4DVar appears to compensate for the lack of

change between the prior and posterior GPP across most of the Northern Hemisphere.

We have also added Figure A4, which shows the time series for the different stations. We calculated Pearson's correlation coefficients for each station, but they were all between 0.98 and 1. This high value is mainly due to the fact that we are in a twin experiment mode, where 'synthetic' observations are generated by the model. We felt that it was not necessary to include them in the manuscript.

8. The differences in RMSD reductions between different methods and configurations are discussed, but no statistical tests are provided. Please include simple significance tests (e.g., paired t-tests) to assess whether the differences in RMSD reductions are statistically meaningful across stations.

We thank the reviewer for this suggestion. We have performed a paired t-test as suggested by the reviewer, which confirms that the differences in RMSD reduction are significant. We added the following text in L423:

Since the posterior RMSDs obtained were close, we performed a paired t-test (Student, 1908) between the two posterior RMSDs to determine whether they were significantly different. We obtained a t-value of -2.125 between the posterior RMSDs obtained by 4DEnVar and e-4DVar, with a p-value of 0.046. This confirms that the average posterior RMSD obtained by 4DEnVar is significantly lower than the posterior RMSD obtained by e-4DVar, with a confidence level of 95%.

9. The comparison between 4DEnVar and ε-4DVar is repeatedly emphasized, but ε

4DVar is not equivalent to standard 4DVar. Please emphasize earlier and more clearly that the ε-4DVar results are only a rough approximation and that conclusions should not be generalized to comparisons with a full 4DVar system.

We do agree that the ε-4DVar is not equivalent to standard 4DVar and we actually explicitly say it in the manuscript L530

> The results obtained here for the ε-4DVar are not equivalent to a standard 4Dvar using a tangent linear and adjoint model. Therefore, we can draw no conclusions on the comparison between the 4DEnvar and standard 4DVar methods as was highlighted in Liu et al. (2008)

In order to clarify we added in L229 the following sentence:

> **Due to this approximation, ε-4DVar is therefore not entirely equivalent to standard 4DVar.**

10. All evaluations are performed against the same synthetic dataset used for assimilation. For robustness, a portion of the synthetic observations should be withheld during assimilation and used for independent validation.

We agree with the reviewer that this approach is relevant as a sanity check for data assimilation, and in particular when "real" observations are used. Because we here use synthetic data over a limited time window (2-years) and a limited number of stations, we have rather chosen to perform a complementary evaluation with respect to NBP and the subcomponents carbon flux (GPP) in Fig 7 and Fig 3A. This evaluation is described in the discussion section, L491

> **Fig. 3A shows the differences in spatial distribution of gross primary production (GPP) as well as their global estimates. We can see that 4DEnVar better matches the global estimates and the spatial distribution of the synthetic flux. It seems that e-4DVar has to compensate for the fact that the BoND PFT does not change**

between the prior and the posterior and seems to fall into a local minimum.

References

Pinnington, E., Quaife, T., Lawless, A., Williams, K., Arkebauer, T., and Scoby, D.: The Land Variational Ensemble Data Assimilation Framework: LAVENDAR v1.0.0, Geoscientific Model Development, 13, https://doi.org/10.5194/gmd-13-55-2020, 2020.

Liu, C., Xiao, Q., and Wang, B.: An ensemble-based four-dimensional variational data assimilation scheme. Part I: Technical formulation and preliminary test, Monthly Weather Review, 136, https://doi.org/10.1175/2008MWR2312.1, 2008.

Bastrikov, V., Macbean, N., Bacour, C., Santaren, D., Kuppel, S., and Peylin, P.: Land surface model parameter optimisation using in situ flux data: Comparison of gradient-based versus random search algorithms (a case study using ORCHIDEE v1.9.5.2), Geoscientific Model Development, 11, https://doi.org/10.5194/gmd-11-4739-2018, 2018.

Peylin, P., Bacour, C., MacBean, N., Leonard, S., Rayner, P., Kuppel, S., Koffi, E., Kane, A., Maignan, F., Chevallier, F., Ciais, P., and Prunet, P.: A new stepwise carbon cycle data assimilation system using multiple data streams to constrain the simulated land surface carbon cycle, Geoscientific Model Development, 9, https://doi.org/10.5194/gmd-9-3321-2016, 2016.

Kuppel, S., Chevallier, F., and Peylin, P.: Quantifying the model structural error in carbon cycle data assimilation systems, Geoscientific Model Development, 6, https://doi.org/10.5194/gmd-6-45-2013, 2013.

MacBean, N., Peylin, P., Chevallier, F., Scholze, M., and Schürmann, G.: Consistent assimilation of multiple data streams in a carbon cycle data

assimilation system, Geoscientific Model Development, 9, https://doi.org/10.5194/gmd-9-3569-2016, 2016.