Report review #2

This manuscript presents an ensemble-based variational data assimilation framework aimed at calibrating parameters in the ORCHIDEE Land Surface Model using atmospheric CO_2 concentration data. The authors refer to their method as a 4DEnVar approach and demonstrate its performance through synthetic twin experiments. However, after carefully reviewing both the manuscript and the authors' response to my previous comments, I remain increasingly convinced that the methodology implemented in this study more accurately reflects a 3DEnVar formulation rather than a true 4DEnVar. I strongly encourage the authors to revisit my earlier review, in which I raised key concerns regarding the temporal treatment of the observation operator and the background error covariance. In the following comment, I will explain in more detail why the method, as described, lacks the defining features of a four-dimensional ensemble-variational system and should therefore be categorized as a 3DEnVar.

We would like to thank the reviewer for taking the time to read through and comment on this manuscript. Your comments will undoubtedly strengthen the paper and help clarify key points. We have thoroughly revised the manuscript and incorporated all the changes you suggested. Below, you will find a point-by-point response to your comments:

1. In my previous review, I highlighted that one of the fundamental distinctions between 3DVar and 4DVar (and, by extension, 3DEnVar and 4DEnVar) lies in the incorporation of an assimilation window, i.e., whether the assimilation framework explicitly accounts for the temporal evolution of observations and background states within a defined time window. However, in the initial manuscript, there was no mention or implementation of an assimilation window, and the cost function used follows a 3DVar formulation. In the authors' response, they argue that their assimilation window spans two years (2000–2001), equating the entire simulation period with a single "assimilation window". I must emphasize that this interpretation is incorrect and conceptually flawed. An assimilation window refers to the temporal period over which observations and model background states are compared and assimilated within each cycle. In typical 4DVar or 4DEnVar systems, the assimilation window is short (e.g., daily or sub-daily), and the system progresses through multiple assimilation cycles to iteratively improve the estimate of the

optimal analysis over time. In contrast, the authors perform only a single assimilation cycle over a two-year period, without any time-evolving ensemble perturbations or temporally resolved assimilation updates. If one were to accept the authors' definition of an assimilation window, then any 3DVar system operating on a long time series would be mistakenly classified as a 4DVar system, which is clearly not consistent with established data assimilation literature. Moreover, given that the model is run only once over 2000–2001, the proper assimilation window—had this been a genuine 4DVar or 4DEnVar system—should be of at least daily resolution within that two-year period, requiring many sequential assimilation cycles. Therefore, I cannot accept the authors' interpretation that their assimilation window is two years in length. Based on the methodology described and implemented, I am confident that this system is best characterized as a 3DEnVar, not a 4DEnVar.

Thank you for your thorough and insightful feedback regarding the classification of our data assimilation methodology. We greatly appreciate your expertise and the opportunity to refine our manuscript to ensure clarity and alignment with established data assimilation terminology.

We understand your concern that our interpretation of a 4DVar does not correspond to the conventional definition of a 4DVar (or 4DEnVar) system, which typically involves several short assimilation cycles (e.g., daily or sub-daily) in order to account for the temporal evolution of observations and model states. Your comment that an assimilation window refers to the time period during which observations and model background states are compared within each cycle is relevant, particularly in the context of traditional 4DVar systems used in meteorological applications.

In our study, we seek to optimize time-invariant parameters (e.g., Vcmax, Q10) using a variational framework that explicitly takes into account the temporal evolution of observations (atmospheric CO2 concentrations) and model results (CO2 concentrations simulated from the ORCHIDEE model coupled with the LMDZ atmospheric transport model) over a two-year period. These time-invariant parameters are fixed constants used to calculate the interested variable (here the NBP flux, and then the atmospheric concentration). You are therefore correct in saying that the background terms are not evaluated over time (as this is not possible), even if changing the parameter would alter the evolution of the NBP (and then the atmospheric concentration) over time (as this would change the state of the carbon pool, for example).

We also acknowledge that our use of a single assimilation cycle covering the entire two-year period deviates from the conventional 4DVar/4DEnVar framework, which typically uses sequential cycles with shorter assimilation windows.

However, the cost function incorporates observations spread over time and model forecasts, comparing them at several points during this period in order to constrain the parameters. It is this temporal dimension of the assimilation process that led us to classify our approach as 4DVar and 4DEnVar, as it captures the dynamic evolution of the system rather than relying on a single snapshot, as in 3DVar or 3DEnVar.

To avoid potential misinterpretation and to align more closely with the expectations of the data assimilation community, we propose a compromise by adopting the terms Variational Data Assimilation (VarDA) and Ensemble Variational Data Assimilation (EnVarDA) in place of 4DVar and 4DEnVar, respectively. These terms emphasize the variational nature of our approach, which optimizes parameters by leveraging the temporal evolution of observations and model outputs. We have revised the manuscript to reflect this updated terminology and have added a clarification in Section 1 to explicitly describe our assimilation framework L52:

> There is a long history of using data assimilation frameworks to calibrate LSM parameters (Rayner, 2010; MacBean et al., 2022; Raoult et al., 2024b). **Most of the methods used for parameter calibration are derived from Bayesian formulations of inverse problems and defined here as variational data assimilation (VarDA) methods. The VarDA method is inspired by the four-dimensional variational (4DVar) method, which was originally developed in the field of meteorology and Earth sciences (Talagrand and Courtier, 1987; Courtier et al., 1994; Asch et al., 2016) and also employed in atmospheric inversions to correct the surface CO2 fluxes (Chevallier et al., 2005; Basu et al., 2013; Liu et al., 2021). This approach is characterised by the definition of a cost function, which is typically based on a least-squares criterion. This cost function calculates two terms: (i) an observation term that computes the difference between observations and model outputs, and (ii) a background term that incorporates prior knowledge of**

> **the state. The computation of both terms is performed in space and time. We define here the VarDA method, as our focus is not on directly optimizing the prior state. Instead, we concentrate on time-invariant parameters used in the parameterisation that defines the variable of interest, such as the Net Carbon Flux. Therefore, while the observation term of the cost function incorporates time-distributed observations and model predictions - comparing them across multiple time points - the background term only compares prior parameter values once, as these values remain constant over time. Furthermore, with the VarDA method, a single assimilation cycle covering the entire observation period is used, which differs from the conventional 4DVar framework, which generally uses sequential cycles with shorter assimilation windows. In order to minimise this cost function, the VarDA method calculates its gradient with respect to the different parameters to be calibrated.**

and L89

> More recently, an ensemble 4DVar method named 4DEnVar implemented in Pinnington et al. (2020) for LSM parameter estimation has proved very promising. This method uses a small ensemble to circumvent the necessity for a tangent linear and adjoint model. This 4DEnVar method has been used to estimate JULES LSM crop parameters at a single Nebraskan site (Pinnington et al., 2020) and to calibrate pedotransfer functions to improve JULES LSM soil moisture predictions over East Anglia (Pinnington et al., 2021) and the whole of the UK (Cooper et al., 2021). This method was also successfully used by Douglas et al. (2025) to calibrate the parameters of a simple carbon model in a twin experiment. **Although the method was defined as 4DEnVar in Pinnington et al. (2020) and Douglas et al. (2025), we choose to refer to it as EnVarDA to maintain consistency with the definitions previously presented**.

This clarification highlights that our approach optimizes parameters over a time period using a single cycle, with the cost function incorporating time-varying observations and model outputs, distinguishing it from traditional 3DVar/3DEnVar methods while acknowledging its differences from standard 4DVar/4DEnVar implementations.

We have also added:
L205: “where, for example, the concatenated vector y=(y_0,y_1,...,y_{N_t})^T represents all available observations at all times over the time window.”
L261: “where each H(x_i) is a concatenated vector of extracted simulations to correspond with all observations available at all times across the time window.

We believe this revision addresses your concerns while maintaining the integrity of our methodology. Thank you again for your valuable feedback, which has significantly improved the clarity and rigor of our manuscript.

2. In my previous review, I explicitly pointed out that setting all diagonal elements of the R matrix to 0.01 ppm is an overly simplistic and unrealistic treatment of observation error. I suggested that, at a minimum, the R matrix should reflect the variance of the observations at each site, which would introduce basic spatial heterogeneity and improve the physical realism of the assimilation system. However, in the authors’ latest response, no changes were made to the R matrix design. The authors merely acknowledge that the 0.01 ppm setting is simplistic, but continue to use it without further justification or adjustment. Computing observation-based variances—even from synthetic data—is not technically difficult, especially in a twin experiment framework where the synthetic observations are fully defined. A more realistic R matrix would be straightforward to implement and would significantly improve the credibility of the assimilation framework. Therefore, I strongly urge the authors to either: Revise the R matrix to reflect spatially varying variances based on the observation time series, or Provide a quantitative justification (e.g., sensitivity analysis) showing that using a constant 0.01 ppm does not materially affect the assimilation results. Without such a revision or justification, the conclusions drawn from the assimilation experiments may not be robust or generalizable.

We thank you for your feedback regarding the R matrix in our manuscript. Your comments, particularly on the simplistic use of a uniform diagonal R matrix with a value of 0.01 ppm highlight an important aspect of data assimilation that warrants careful consideration.

We fully agree that, in applications involving real observations, the R matrix should incorporate at least spatial and, ideally, temporal variability to reflect observation error variances at different sites, thereby enhancing the physical realism of the assimilation system. In this study, our primary objective was to compare the performance of two data assimilation methods - VarDA and the EnVarDA - in solving

an inverse problem for parameter calibration within a complex atmospheric CO2 concentration modeling framework, specifically using the ORCHIDEE land surface model coupled with the LMDZ atmospheric transport model.

To focus on this methodological comparison, we employed a twin experiment framework with synthetic observations generated by ORCHIDEE+LMDZ without added noise or perturbations. This idealized setup eliminates complexities such as model structural errors or measurement errors, allowing the model to perfectly match the synthetic observations. In this context, we chose a uniform R matrix value of 0.01 ppm for both methods to ensure a consistent and controlled comparison, as the synthetic observations are noise-free and equally reliable across all stations. Assigning differential weights to stations via a heterogeneous R matrix seemed less relevant and sub-optimal in this perfect case, where all stations can theoretically be fitted equally well.

We added L348:

> **The R matrix represents the model-structural and observation errors. In our twin experiment setup, we choose not to add only very small errors in the pseudo-observations to compare both methods in an ideal case. In this context, structural errors in the ORCHIDEE LSM (i.e. missing processes, etc.) and in the transport model (i.e., coarse spatial resolution, wind biases, etc.), or measurement errors are discarded. Indeed, since the pseudo-observations are generated by a simulation, as detailed in Section 2.3, there exists at least one solution where all observations can be matched perfectly. For this reason, we use a simplified R matrix with the same small diagonal terms of 0.01 ppm for all stations. The rationale behind this choice is that, as all stations can be matched perfectly, we do not want to introduce any spatial or temporal preferences.**

It is also clear that any change to the R matrix should change the results, as the opposite seems incorrect, but we believe that characterising the error in a system and testing the potential performance of a system are two different things. In our article, we do not wish to characterise the observation/model error (as there is none in reality), but we wish to test and compare two methods. We are convinced that as long as the same R is used in both systems, we can conclude that EnVarDA has many advantages over VarDA.

Moreover to assess the robustness of our approach and address the potential risk of overfitting due to the simplified R matrix, we evaluated the normalized chi-square statistic for the simple case, following Talagrand and Boutier (1999) and Trémolet (2006), obtaining scores of 0.015 for EnVarDA and 1.2 for VarDA. These values suggest that our error estimates are either appropriately specified or slightly

overestimated, as scores significantly greater than 1 would indicate error underestimation and a potential shift in the cost function minimum (Trémolet (2006)).

These findings support the validity of our conclusions within the controlled twin experiment framework. We recognize, however, that a more realistic R matrix incorporating spatially varying variances is critical for the use of real-observation, as we have observed in an ongoing work with real observations. We have also explicitly added the point raised by the reviewer in the discussion L565:

> However, the assimilation of real observations is not straightforward. The use of real data must be followed by characterisation of the model/observation errors. **Indeed, the matrix R must reflect modelling/observation errors at each site, which would introduce spatial heterogeneity, as each station may have different modelling errors, mainly structural errors from both the transport model and the ORCHIDEE flux model, or measurement problems. A good characterisation of the matrix R is of paramount importance, as it can have a considerable impact on the results obtained.** If the model/observation errors are incorrect, the EnVarDA method can give infeasible posterior parameter values, i.e. outside the imposed parameter bound- aries (and therefore give non-physical parameter values). Furthermore, even with feasible posterior parameter values, the parameters obtained may be beyond the assumption of linearity made by the use of linear combinations in Eq. 18 and therefore do not improve the associated simulation. Nevertheless, several techniques seem promising for managing these limitations.

We believe this approach will address your concerns without necessitating a complete re-optimization of all experiments, which would be computationally intensive and time-consuming given the scope of the study.

3. The manuscript lacks a clear and detailed explanation of how the background error covariance matrix B is constructed. This is a fundamental component of any variational data assimilation framework, as it governs how prior uncertainty is propagated into the analysis. However, the manuscript appears to apply the same simplified approach to the background error covariance matrix B as it does to the observation error covariance matrix R —namely, by assigning constant diagonal values of 0.01. This practice is scientifically inappropriate. The background error covariance matrix and the observation error covariance matrix represent distinct sources of uncertainty and must be treated separately. Using the same constant value for both implicitly assumes that the model and observation uncertainties are identical in magnitude and structure, which is both unrealistic and unjustified—even in a twin experiment setup. Even

in idealized experiments, a scientifically grounded design of B is expected. For example, B could be derived from ensemble statistics, parameter perturbation experiments, or climatological variances. These are standard practices in both 3DVar and EnVar systems.

We thank the reviewer for their comment and apologise if our original explanation was unclear. We fully agree that the two matrices **R** and **B** must be treated separately and confirm that this is what we have done in the article. Note that we couldn't have done otherwise, because **R** relates to the error in atmospheric concentrations (in ppm), while **B** relates to the error in the model parameters, each of which has its own range of variation and unit. As discussed in the comment above **R** is indeed a diagonal matrix with an error of 0.01 ppm. But we clearly presented - also in section 2.3.3 - how **B** is designed : "The **B** matrix contains the background errors associated with the prior knowledge of the model parameters. We set an error corresponding to 30% of the parameter range for the simple case and 20% for the complex case (as we use larger parameter ranges)." To improve clarity and avoid potential confusion, we have revised Section 2.3.3 to more explicitly separate the descriptions of **R** and **B** into two distinct paragraphs (see Lines 330–345 in the revised manuscript):

> **To implement the two data assimilation methods, ϵ-VarDA and EnVarDA, we define two error covariance matrices: R and B. These matrices are configured to be diagonal, as we are assimilating "synthetic" observations, and are common to both methods to ensure comparable experiments. Their configurations are informed by previous data assimilation studies using ORCHIDEE and a simplified carbon model (Kuppel et al., 2012, 2013; Bastrikov et al., 2018; MacBean et al., 2016), with Peylin et al. (2016) specifically applying diagonal matrices for atmospheric CO2 observations.**
> **R Matrix**
> [...]
> **B Matrix**
> **The B matrix represents the background errors associated with prior knowledge of the parameters. We set the error to 30% of the parameter range for the simple case and 20% for the complex case (as we use larger parameter ranges for this case). The background errors of each parameter can be seen in Fig. 3 for the simple case and in Fig. 6 as well as Table A2.**

4. To improve clarity and transparency, I strongly recommend that the authors include a dedicated "Experiment Design" section in the manuscript, preferably early in the Methods section. Currently, the description of the different twin experiments is scattered and somewhat

difficult to follow, especially with regard to the distinctions between test cases, the naming conventions used, and the variables being optimized. Additionally, I suggest including a summary table that clearly outlines the different experiments conducted.

We thank the reviewer for these comments and apologise for the lack of clarity. We have modified the structure as follows:

2.3 Experiment Design

2.3.1 Twin Experiment Description

To test the data assimilation methods presented in Section 2.2, we conducted a so-called twin experiment to evaluate their efficiency in calibrating parameters involved in calculating NBP fluxes in the ORCHIDEE LSM model. This experimental framework reduces complexities associated with model-data errors, focusing on the performance of the assimilation methods. The known 'true' parameters being the default parameter values of the ORCHIDEE model are used to generate the synthetic observations. New values of a priori parameters are manually generated, ensuring physically meaningful values that differ from the 'true' parameters both presented in Table A2. The assimilation methods are then applied to assess how closely they converge toward the known solution (standard parameter values). The synthetic observations of atmospheric CO2 concentrations from the 21 continental stations are assimilated simultaneously over a two-year window (2000–2001) to monitor spatial and temporal variations in carbon fluxes, as shown in Figure A4. A limited period was chosen for practical reasons to avoid computationally expensive simulations.

2.3.2 Generation of Synthetic Observations

To generate synthetic observations for the twin experiment, we simulate net biome productivity (NBP) fluxes at the global scale using the ORCHIDEE LSM with default parameter values, referred to as the 'true' parameters (see Table A2). These NBP fluxes represent the net carbon fluxes of the land component, calculated as the difference between emission fluxes (heterotrophic and autotrophic respiration, and disturbance fluxes due to land-use change) and sink fluxes (primarily due to photosynthesis). The concentration given by the surface fluxes (the simulated NBP fluxes, along with other fluxes described in Section 2.1.4) are

transported using pre-calculated transport fields of the LMDZ model over the period 2000–2001. We then extract atmospheric CO2 concentrations at 21 continental atmospheric stations, shown in Figure 1, which are highly sensitive to continental carbon fluxes, providing significant constraints on the parameters. This process enabled the generation of synthetic observations of monthly average atmospheric CO2 concentrations at these 21 stations over the two-year period. It is important to note that the steps taken here to generate the synthetic observations are exactly the same as those used to perform a simulation. This means that there is at least one solution where the model can perfectly match the synthetic observation.

2.3.3 Simplified case

[...]

2.3.4 Complex case

[...]

2.3.5 Error covariance matrices

[...]

2.3.6 Tuning ϵ for gradient calculation

[...]

2.3.7 Defining the impact of the configuration

[...]

Each section was revised to reflect every aspect of the experiment we performed. We specifically split the first section into two parts to explicitly describe how the Synthetic Observations are generated, thereby removing any potential doubt. No additional information was added to the text but the text was revised to eliminate any potential confusion. We did include a list of all the experiments performed in the last section L380:

> To assess their impact, we launch the twin experiment using different configurations:
>
> - for the simple case:

    - 5 different values of ϵ for the ϵ-VarDA based on the sensitivity test presented in Section 2.3.6 ;
    - 5 different ensemble sizes in the EnVarDA;
- for the complex case:
    - 5 different ensemble sizes in the EnVarDA;
    - **1 values of ϵ for the ϵ-VarDA.**

5. In the author's response, it is stated that the experiments represent a full-field assimilation, implying that the assimilation directly updates the full state variables and can correct potential model biases. However, this characterization is not substantiated in the manuscript. There is no analysis or discussion demonstrating how the assimilation affects model biases—either in the prior fields, posterior fields, or fluxes. A full-field assimilation experiment should, by definition, lead to noticeable improvements in the state estimation compared to the biased model trajectory. To support this claim, I strongly recommend that the authors: 1. Include an explicit evaluation of model biases before and after assimilation, especially in CO_2 concentrations or fluxes (e.g., NBP); 2. Quantify the impact of assimilation on these biases.

Thank you for your feedback and for highlighting the need for a clear demonstration of how our assimilation approach addresses model biases in CO_2 concentrations and fluxes, such as Net Biome Productivity (NBP).

In our study, we focus on calibrating parameters within the ORCHIDEE LSM that govern key processes such as photosynthesis and soil carbon decomposition, which in turn influence exchange fluxes like Gross Primary Production (GPP) and NBP. These parameters are optimized using atmospheric CO_2 concentration data, while forcing variables (e.g., temperature, wind, precipitation) are prescribed from ERA-Interim reanalysis data to ensure accurate timing of meteorological events. As such, our assimilation does not directly update the state variables themselves but indirectly improves the model's representation of CO_2 concentrations and fluxes by optimizing the parameters that drive these processes.

We acknowledge that the use of the term 'full-field assimilation' in our previous response could have been confusing. These terms are not used in the context of LSM parameter calibration (see, for example, the review articles by Raoult et al. 2024, Macbean et al. 2022, or Rayner et al. 2010).

To address your specific recommendations:

1. **Evaluation of Model Biases Before and After Assimilation**: We believe that our manuscript already includes a comprehensive evaluation of biases in CO_2 concentrations and fluxes. For CO_2 concentrations, we quantify the improvement in model performance through the Root Mean Squared Difference (RMSD) scores, as presented in Figures 4 and 5, which compare prior and posterior simulations against synthetic observations at 21 atmospheric stations. Specifically, Figure 4 (and Figure A4 in the revised manuscript) shows the time series of CO_2 concentrations, illustrating the reduction in discrepancies between model outputs and observations post-assimilation. For fluxes, Figure 7 and Figure A3 in the revised manuscript provide spatial differences in NBP and GPP fluxes, respectively, between prior and posterior estimates compared to “true” synthetic fluxes. These figures demonstrate the impact of assimilation on reducing biases in both CO_2 concentrations and fluxes. We have also added a new analysis following Hodson et al. (2021) and Geman et al., 1992, who proposed to decompose the mean square difference (MSD) into bias and variance L451

   > **We computed the mean squared difference (MSD) between the synthetic observations concatenated across all stations and the prior simulation, as well as the two posterior simulations. Following Hodson et al. (2021) and Geman et al. (1992), we decomposed the MSD into bias and variance terms as presented in Section A. The prior MSD is 11.49 ppm2 and is reduced to 0.04 ppm2 using the EnVarDA method and to 0.08 ppm2 using the VarDA method. The decomposition of the prior MSD indicates a squared bias of 4.96 ppm2 and an error variance equal to 6.53 ppm2. The same decomposition for the posterior simulations yields a squared bias of 0.006 ppm2 and an error variance equal to 0.03 ppm2 for the EnVarDA method, and a squared bias of 0.002 ppm2 and an error variance equal to 0.07 ppm2 for the VarDA method.**

   And L502

   > **Furthermore the MSD score is better for the EnVarDA method (0.04 ppm2 using the EnVarDA method and to 0.08 ppm2 using the VarDA method) and the MSD decomposition (Geman et al., 1992; Hodson et al., 2021) highlights that EnVarDA better reduces the error variance, whereas the squared bias reduction is slightly better for the VarDA method. However, squared bias values below 0.01 ppm2 are negligible. While the mean RSMD reduction and MSD scores are similar for the complex case, the MAD scores in parameter space are different.**

2. **Quantification of Bias Impact**: The manuscript quantifies the impact of assimilation on biases in CO_2 concentrations and fluxes in the “Results” and “Discussion” sections. For CO_2 concentrations, we report a mean RMSD reduction of 91.3% for 4DEnVar and 92.3% for e-4DVar across all stations (L 425–430, Page 17). For NBP and GPP, we discuss the recovery of global budgets and highlight challenges in achieving correct spatial distributions due to equifinality, particularly in 485–497 and 510–517 (Pages 21 and 23). These sections detail how both methods successfully reduce biases in global NBP and GPP budgets, although 4DEnVar performs better in capturing spatial patterns due to its ensemble-based approach, which mitigates issues related to local minima.

We believe these analyses directly address the impact of assimilation on model biases, as requested.

References

Asch, M., Bocquet, M., and Nodet, M.: Data Assimilation: Methods, Algorithms, and Applications, vol. 28, 2016.

Courtier, P., Thépaut, J., and Hollingsworth, A.: A strategy for operational implementation of 4D-Var, using an incremental approach, Quarterly Journal of the Royal Meteorological Society, 120, https://doi.org/10.1002/qj.49712051912 , 1994.

Talagrand, O. and Courtier, P.: Variational Assimilation of Meteorological Observations With the Adjoint Vorticity Equation. I: Theory,Quarterly Journal of the Royal Meteorological Society, 113, https://doi.org/10.1002/qj.49711347812 , 1987.

Chevallier, F., Fisher, M., Peylin, P., Serrar, S., Bousquet, P., Bréon, F. M., Chédin, A., and Ciais, P.: Inferring CO2 sources and sinks from satellite observations: Method and application to TOVS data, Journal of Geophysical Research Atmospheres, 110, https://doi.org/10.1029/2005JD006390 , 2005.

Basu, S., Guerlet, S., Butz, A., Houweling, S., Hasekamp, O., Aben, I., Krummel, P., Steele, P., Langenfelds, R., Torn, M., Biraud, S., Stephens, B., Andrews, A., and Worthy, D.: Global CO2 fluxes estimated from GOSAT retrievals of total column CO2, Atmospheric Chemistry and Physics, 13, https://doi.org/10.5194/acp-13-8695-2013 , 2013.

Yannick Trémolet. 2007. Model-error estimation in 4D-Var. Quarterly Journal of the Royal Meteorological Society 133, 626, 1267–1280 https://doi.org/10.1002/qj.94, 2007.

O. Talagrand and F. Bouttier, Internal diagnostics of data assimilation systems,Conference Paper, https://www.ecmwf.int/en/elibrary/76594-internal-diagnostics-data-assimilation-systems, 1999.

Raoult, N., Douglas, N., MacBean, N., Kolassa, J., Quaife, T., Roberts, A. G., Fisher, R. A., Fer, I., Bacour, C., Dagon, K., Hawkins,L., Carvalhais, N., Cooper, E., Dietze, M., Gentine, P., Kaminski, T., Kennedy, D., Liddy, H. M., Moore, D., Peylin, P., Pinnington, E., Sanderson, B. M., Scholze, M., Seiler, C., Smallman, T. L., Vergopolan, N., Viskari, T., Williams, M., and Zobitz, J.: Parameter Estimation in Land Surface Models: Challenges and Opportunities with Data Assimilation and Machine Learning, ESS Open Archive, https://doi.org/10.22541/essoar.172838640.01153603/v1 , 2024b.

MacBean, N., Bacour, C., Raoult, N., Bastrikov, V., Koffi, E. N., Kuppel, S., Maignan, F., Ottlé, C., Peaucelle, M., Santaren, D., and Peylin, P.: Quantifying and Reducing Uncertainty in Global Carbon Cycle Predictions: Lessons and Perspectives From 15 Years of Data Assimilation Studies With the ORCHIDEE Terrestrial Biosphere Model, Global Biogeochemical Cycles, 36, e2021GB007 177, https://doi.org/https://doi.org/10.1029/2021GB007177 , e2021GB007177 2021GB007177, 2022.

Rayner, P. J.: The current state of carbon-cycle data assimilation, https://doi.org/10.1016/j.cosust.2010.05.005 , 2010.

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.

Kuppel, S., Peylin, P., Chevallier, F., Bacour, C., Maignan, F., and Richardson, A. D.: Constraining a global ecosystem model with multi-site eddy-covariance data, Biogeosciences, 9, https://doi.org/10.5194/bg-9-3757-2012 , 2012.

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.

Douglas, N., Quaife, T., and Bannister, R.: Exploring a hybrid ensemble–variational data assimilation technique (4DEnVar) with a simple ecosystem carbon model, Environmental Modelling & Software, 186, 106361, https://doi.org/10.1016/j.envsoft.2025.106361, 2025.

Pinnington, E., Quaife, T., Lawless, A., Williams, K., Arkebauer, T., and Scoby, D.: The Land Variational Ensemble Data Assimilation Framework: LAVENDAR v1.0.0, Geosci. Model Dev., 13, 55-69, 10.5194/gmd-13-55-2020, 2020.

Geman, S., Bienenstock, E., and Doursat, R.: Neural Networks and the Bias/Variance Dilemma, Neural Computation, 4, https://doi.org/10.1162/neco.1992.4.1.1, 1992.

Hodson, T. O., Over, T. M., and Foks, S. S.: Mean Squared Error, Deconstructed, Journal of Advances in Modeling Earth Systems, 13,

https://doi.org/10.1029/2021MS002681, 2021.