Reviewer 1

Overall, I applaud the authors on the use of multi-observation streams, including manipulation experiments to explore the extent they can better optimise their LSM. The paper is logically put together and clearly written. I have a few potential issues but hopefully, nothing that would preclude this manuscript from being published with some revision. This is an exciting approach and you could see how it might be extended for a broader range of flux/satellite and manipulation experiments (e.g. https://onlinelibrary.wiley.com/doi/full/10.1111/gcb.16585).

We would like to thank the reviewer for his positive review and enthusiasm for our work. We do indeed hope that this work will open up a lot of exciting new avenues given the wide range of manipulation experiments available. We also thank the reviewer for the citation which has been added to the text (L42).

One issue I have with the aims in the intro relates to the extent it makes sense to optimise against both ambient and elevated conditions (your line 77)? The underlying driving principle of the FACE model intercomparison was that the models should be able to predict eCO2 response given existing parameterisations and underlying theory. If the parameters need to be re-optimised for eCO2 conditions, this is fine, but it implies the models lack the theory to predict these changes. For example, in De Kauwe et al. 2013, we tested whether the stomatal slope changes with eCO2 (Fig 6) and found it did not (i.e. if the theory is right the ambient parameterisation should allow you to predict the eCO2 response), supporting the point I just made. This is not to say it isn't a valid question to ask / test, but I think the text requires more nuance that we currently see. For example, to achieve the improved fit when eCO2 was considered, what params had to change? More on this point below.

The reviewer is correct that, in theory, we should be able to calibrate against ambient conditions and keep the elevated data for the evaluation since a robust model should be able to predict changes under different conditions. However, we know that our models are not perfect - there are a lot of missing processes, for example, ones linked to plant adaptation, as well as issues linked to scale representation, for example, leaf-scale in situ measurements versus kilometer-wide model resolution. Furthermore, our treatment of the nutrient limitation, although state-of-the-art, remains a very rough parameterisation of the nitrogen cycle. Therefore, we can not expect to be able to calibrate under ambient and have the model automatically work under elevated CO_2 conditions.

We strongly agree that we shouldn't have a separate elevated CO_2 parameter set. Note that at the end of the BOTH calibration, we only have one set of parameters - a parameter set that works best under both atmospheric regimes. In addition, note also the fact there is

such large uncertainty in the parameters, it is also possible to have the correct theory (i.e., equations) but the wrong parameters. When optimising the cost function, we find that our parameter space is full of minima (Kuppel et al., 2012; Bastrikov et al, 2018). This means that there can be different parameter sets giving similar fits but with wildly different parameter values. Therefore, it is recommended to perform multi-data stream calibrations to a) help smooth parameter space and b) make sure we don't overfit to one data source. By treating ambient and elevated conditions as two different data streams, we ensure we don't overfit to one regime but get the parameter set giving the compromise. Finally, given we will be using this model in CMIP-type exercises, we need to have the most operational version with a best set of parameters to use. As seen in Figure 5, for one model structure, different parameter values can give large variations in model responses. The proposed methodology in the paper (calibrating using data from both ambient and elevated CO2 simultaneously) finds a parameter set we have more faith in than just calibrating against ambient conditions.

Nevertheless, the reviewer is correct that we do need to be cautious with our approach. The tuning is probably compensating for missing processes and wrong/un-precise parametrisations, as well incorrect model forcing. Parameter estimation can improve the model fit for the wrong reasons, hiding structural problems. However, it can also help us identify structural errors. Most of the time, we use this methodology to do just that - when we are unable to `match the observations using data assimilation, this helps us to identify structural deficiencies and therefore we feed back to the ORCHIDEE model developers to improve these deficiencies. For example, in this work, we found that, with the set of parameters optimised, we were unable to match the seasonality of LAI suggesting we need to reconsider how LAI phenology is modelled in ORCHIDEE. Furthermore, if we had been unable to find set of parameters that worked satisfactorily under both conditions, then this would have suggested a process was missing - like a functional response to CO_2. We have expanded the text on L443 to add this discussion to the conclusions:

> "...Finally, structural changes do need to be made to the model to better capture the inter-annual variability of simulated NPP and LAI.
>
> **Identifying structural deficiencies is the main strength of parameter estimation - we need to be sure that we are not simulating the right model output for the wrong reasons. Indeed, if we had not been able to find a set of parameters that gave a satisfactory fit to both atmospheric regimes, this would have highlighted a missing process in the model. Ideally, we would want to calibrate under ambient and test the robustness of the theory under eCO2; however, given all missing (e.g., adaptation) or highly simplified processes (nutrient limitation), using both conditions is one approach to improve the overall model behaviour while highlighting these deficiencies.**"

We have also changed L77 to be clearer in the introduction:

> "Furthermore, by optimising a land surface model to both ambient and elevated conditions **simultaneously**, we will gain extra confidence in the model projections using this **single** set of parameters. **Although ideally we would want to calibrate under ambient conditions and test the model under elevated conditions, known model structural errors do not guarantee that the model is able to predict changes under different conditions. As such, we provide an alternative approach to model calibration, maximising the available information content of the optimisations.**"

Finally, we acknowledge that information about the posterior parameters is missing from the manuscript. Since both reviewers have highlighted this gap in our manuscript, we have added a section to discuss how the parameters change. More on this below.

A second issue I have relates to the breadth of parameters to be optimised and the data constraint. For example, to what extent can we expect to constrain a range of these quantities (e.g. litterfall, leafage, LAtoSA, etc) from gross or net carbon fluxes? Clearly indirectly a quantity like the leaf-2-sapwood area affects LAI, which affects GPP, but it feels quite downstream and I wonder why we think the gross flux would offer a strong (any) constraint here? Equally, if it turns out that it does (I'm writing this comment in advance of reading the results), what does this really mean? Again not a problem, but I think as a reader I'd appreciate a sentence or two in the methods on rationale or expectations here.

The reviewer is right to highlight this - this is often a problem in land surface model data assimilation. There are a lot of parameters we can calibrate - most of which will have an indirect effect on the flux assimilated. This is why we performed a sensitivity analysis study to first select the parameters. We acknowledge that this crucial step may be a bit lost in the manuscript since we have put this in the model/parameter section of the Methods and Data. As such we have added the following to better highlight the sensitivity study (and its motivation):

> In Section 2.4 "Performed experiments" L292:
>
> "**Before performing the optimisations, we also conducted a sensitivity analysis on the parameters (as described in Sect. 2.1.2 and shown in Table 1). A sensitivity analysis tests how different the model outputs change with respect to different parameters. This was done to ensure that only parameters showing some sensitivity to the model outputs were used in the optimisation and therefore minimising the risk**

of using parameters that are weakly constrained by the fluxes. This is an important step since we want to avoid constraining parameters that will have a small impact on the optimisation but have the potential to significantly degrade the model-data fit against processes not included in the calibration.

Once spun up **and with the list sensitive parameters,** …"

In Section 3.1 "Fluxnet optimisations" L239:

"... and whether they can be optimised using observed data **and the sensitive parameters identified in the Morris experiment (Table 1)**."

Nevertheless, even after the sensitivity analysis, some parameters are kept that are indirect and therefore weakly constrained by NPP and LAI data. We recognised that this is a risk and that we may overtune parameters without other data. Note that in addition to the gross and net carbon fluxes used, we do assimilate against LAI data, which should provide a more direct constraint on the phonology parameters highlighted by the reviewer (i.e., leaf age crit and leaf-2-sapwood). Nevertheless, more data streams constraining specific processes would be better. This study is only a first step toward a more comprehensive complete calibration. In future, we may consider some sort of cascade tuning to use specific data directly related to specific processes at each step, however, this is beyond the scope of this work. We have added the following text to the conclusions to acknowledge this risk (L445):

"**Although we performed a sensitivity analysis to select sensitivity model parameters, a large number of parameters were kept, some of which are indirectly impacted by the processes optimised. This can pose a risk, since changing such parameters may have an important impact elsewhere in the model and, therefore, may result in a degradation in model-data fit against processes not considered in the optimisation. Therefore, this study is only a first step toward a more comprehensive approach with more data streams."**

A third issue and one to tackle in the discussion relates to how best to utilise FACE data. Overall, I applaud the approach taken here by the ORCHIDEE group but equally have concerns that an implication of calibrating to ORNL for *all* broadleaved sites could inadvertently impose a progressive nitrogen limitation (as observed at this experimental site) across sites where this may not be true. I think it is important to predict the response at ORNL for the right reasons (see Zaehle et al. 2013 New Phyt) rather than using this as a basis for all nitrogen-limited sites. I'd go as far as to say it may even make sense *not* to well match the decline in NPP at this site from a standard tuning exercise.

It is true that we do need to be careful since we do not want to impart the wrong information on the Fluxnet sites. The best way to do this would be to include a lot more FACE sites to capture different conditions. Especially, if we could optimise by grouping sites based on different levels of nitrogen limitation, then if the posterior parameters were found to be similar then the model processes allow for these differences. This study demonstrates the untapped potential of FACE data - but is not without its limitations. We have expanded on the need for more FACE sites in the conclusion (L447):

> "...Fluxnet-only optimisations likely overestimating the CO$_2$ fertilisation effect. **However, we do need to be cautious in assessing these results since we are only using one FACE site for each PFT meaning we are likely tuning to the specificities of that site. For example, ORNL shows a progressive nitrogen limitation but this is not expected over all sites. Ideally, we would include a lot more FACE sites to capture different conditions. Especially, if we could optimise by grouping sites based on different levels of nitrogen limitation, then if the posterior parameters were found to be similar then the model processes allow for these differences.**"

The fourth issue that I felt was lacking from the results was the link between what was tuned and the emergent results. To improve the fluxes are a similar set of params being adjusted for PFT group? Or to obtain better fluxes are the tuned params bespoke per site? Apart from the discussion about Ksoil I'm entirely unclear how and by how much the parameters have changed, this relates to my "second issue" above. As a reader, we do not get any sense of which of the params in table 1 are most important, effectively the source of the improvement is opaque. I'm left to ponder if the tuned params made logical sense or if they reflect a repartitioning of model error or uncertainty.

As mentioned in response to the "second issue", we have added the following section which adds more transparency about the parameter changes.

> 3.2.3 Posterior parameter values
>
> In Fig. 3, we consider how the different parameters have changed by comparing results from the Fluxnet-only optimisation and the optimisation including the FACE data. For the temperate broadleaf summergreen parameters it is hard to distinguish significant patterns. The posterior values sometimes move in the same direction for both optimisations (e.g., K_N, FCN_{wood}, R_{root} and z), other times the parameters are pulled in different directions (e.g., $CN_{leaf,max}$ and R_{leaf}). When considering the most sensitive parameters, both optimisations agree with the direction of change, with the exception of NUE_{opt}. For the photosynthetic parameters A_1 and B_1, we see that

these are changed during the Fluxnet-only optimisation but are less important in correcting the fit when the FACE data is included in the optimisation. Overall this highlights that adding the FACE data can pull the parameters in a different direction than when using only Fluxnet data. This indicates that there are likely compensating effects (or different repartitioning of model error). It is possible that with more data sets, the results would be more robust but this would need to be confirmed with the use and assimilation of additional data streams.

In contrast, for the temperate needleleaf parameters, nearly half (11/23) of the parameters were unchanged (or only slightly) during the Fluxnet-only optimisation. However, when optimised with the additional FACE data, these parameters changed greatly. These parameters include a number of the most sensitive parameters suggesting these are especially important in capturing the model response under both atmospheric regimes for needleleaf sites.

Although looking at these parameter values can be very informative - we must remember that there are complex interactions between the parameters and processes that will not be evident by just looking at these values.

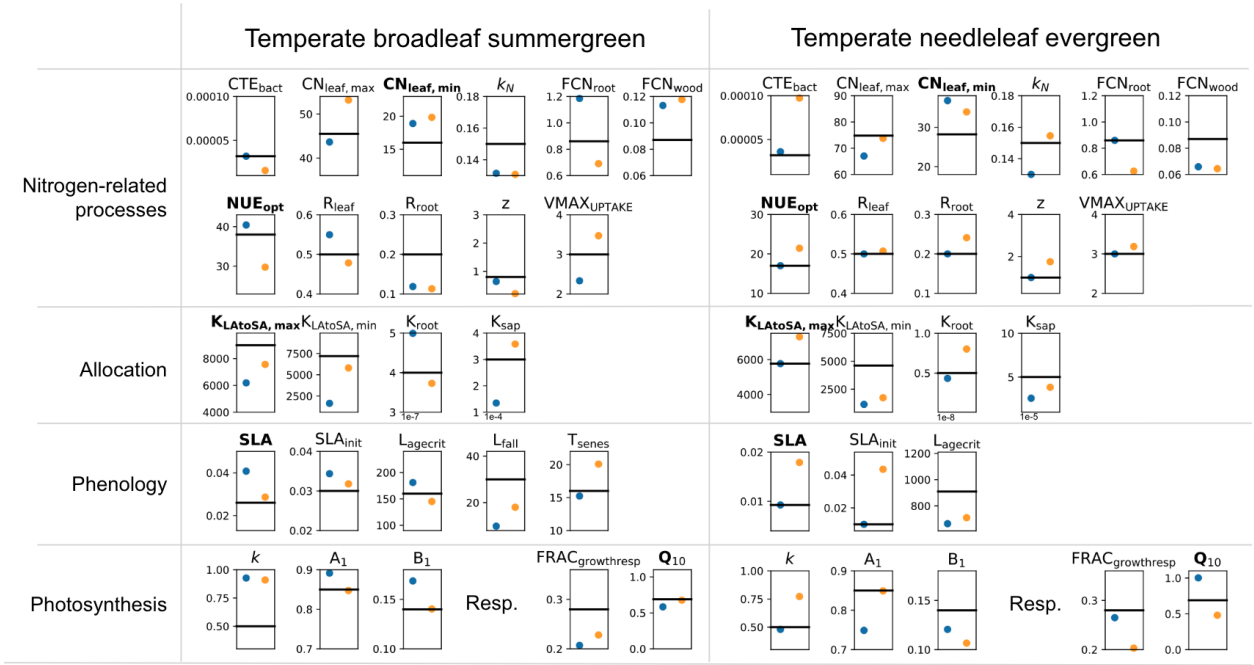


Figure 3: Posterior parameter values of the Fluxnet-only optimisation (blue, FIx_{GN}) and the optimisation incorporating FACE data under both atmospheric regimes (orange, FIx_{GN}-BOTH). The horizontal line represents the prior value of each parameter and each box spans the range of variation

allowed for each parameter during the optimisation. Parameters highlighted in bold correspond to the parameters identified as most sensitive in the preliminary Morris experiment (see Table 1).

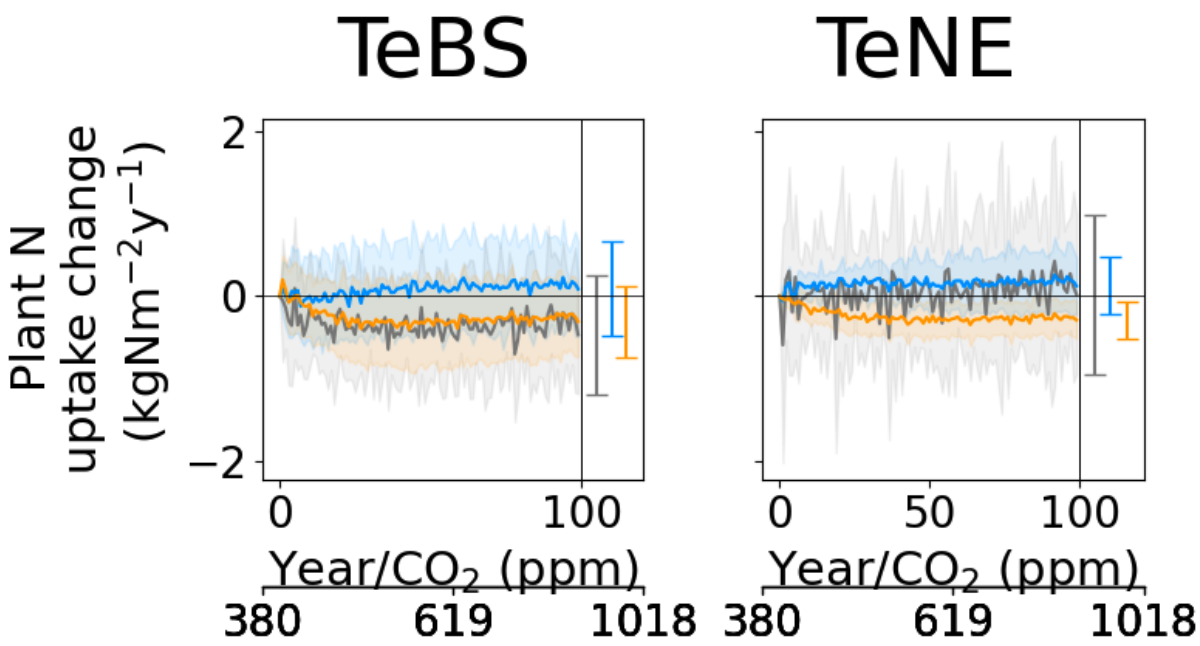
Fifth, in the discussion of Figure 5, you might be able to link to the Walker 2015 paper (https://agupubs.onlinelibrary.wiley.com/doi/10.1002/2014GB004995), there they also imagined what would happen to models if ORNL and duke continued for 300 years. This is similar to your idealised experiment. In particular, that paper extracts the change in N availability over time, which could be something you potentially examine in your analysis. Currently, when you optimise against FACE vs Fluxnet, you get a much lower GPP response, which leads me to wonder how you've affected N cycling cf. Walker et al.

We thank the reviewer for the suggestion. Although it is hard to relate the different "GPP vs. CO2" responses obtained for the different sets of optimizations to potential impacts on the N cycle, and so directly compare to Walker et al, we are able to include the evolution of plant N uptake in Figure 5. The difficulty in performing such a comparison arises because similar changes in responses might be attributed to different parameter adjustments, and different parameters do not impact the N cycle in the same way. Nevertheless, by examining the time evolution of plant N uptake in Figure 5 we can provide valuable insights into how the N uptake varies according to different optimization approaches. Figure 5 and Table 3 have been expanded to include Plant N uptake change as follows, and the text expanded to discuss these results:

	TeBS			TeNE		
	Prior	Optimised		Prior	Optimised	
		Fluxnet only	Fluxnet & FACE		Fluxnet only	Fluxnet & FACE
GPP ($kgCm^{-2}y^{-1}$)	5.22 + 0.55	3.58 + 1.95	4.03 + 0.99	2.85 + 0.73	3.93 + 1.73	4.59 + 0.66
Transpiration ($mm.y^{-1}$)	417.29 - 83.43	342.93 - 47.06	365.57 - 87.52	242.62 - 80.05	313.35 - 79.92	357.47 - 119.82
WUE (%)	123.46 + 100.21	104.64 + 85.25	110.13 - 72.14	114.76 + 104.48	95.17 + 78.5	127.85 + 102.21
Plant N uptake ($kgNm^{-2}y^{-1}$)	7.72 - 0.47	2.17 + 0.09	1.86 - 0.31	7.24 + 0.02	1.14 + 0.12	1.32 - 0.29

Table 3. Mean values at the start of the 100-year experiment and net change by the end of the experiment for the prior model and the optimised models.

"...The optimisations lead to higher starting values (except for WUE), and including data from FACE is found to induce a larger increase. **When considering the total plant N uptake for both vegetation types, the prior starts with a high value and the optimised runs with much lower values.**"



> "**For plant N uptake, changes over the 100-year period are small in magnitude but vary between the different optimisations. The prior model shows a range of responses over the different sites (evidenced by the large spread), overall decreasing the rate of N uptake for TeBS and increasing the rate of uptake for TeNE. The Fluxnet-only optimisations lead to a slight increase in plant N update by the end of the century for both vegetation types. In contrast, the FACE optimisation lead to a decreased plant N update. This is especially notable for TENE where the spread of responses over sites is greatly reduced. In all cases, changes to the rate of change occur during the first half of the simulations, plateauing to a constant value for the rest of the runs.**"

We have also added the following to the conclusions to highlight the perspective of constraining more directly the N cycle in future optimisations (L441):

> These variables are more directly linked to productivity than leaf area index, the variable we use in our optimisations. **In future optimisations, we might also want to include more 'nitrogen-type variables in the cost function, such as leaf nitrogen content.**

Finally, the question that I wondered as I was reading the results: "to what extent the calibration against flux + FACE improves the estimated fluxes in Figure 1 (or not)?". I think section 3.2 & Fig 4 addresses this question but I got a bit lost as to whether the multi-site calibration was then being used to re-examine the flux predictions or not.

This is indeed what we show in Section 3.2.3 and Fig. 4. We agree that this section would benefit from stronger callbacks to Figure 1 and text clarify this. As such, we have added the following:

To L343:

"In other words, this helps us see to what extent the calibration with FACE data changes the estimated fluxes shown in Fig. 1."

Caption of figure 4

"This figure complements Fig. 1, here showing the optimisations using the FACE data fit the timeseries compared to the prior model and the optimisation (GN) without FACE."

Minor

=====

- Line 54. "experiment" - suggest you change it to experiment"s" or "one such type of experiment", currently the text implies there is only a single FACE experiment

Agreed, this has been fixed in the text.

- Line 57-59. I didn't really get the point of these two sentences, they aren't linked at all to the text. There have been various other reviews of FACE which you don't list but could do. I think you either need to link this to the existing text or omit these lines.

We have chosen to simply omit these lines to avoid confusion.

- Line 203. While I do take the argument being made here, I do wonder if it makes sense to calibrate against all of the fluxes. I feel like you would calibrate against TER and NEE and perhaps the assessment would be on how GPP would change. I don't necessarily have a recommended change or response needed here, but it feels intuitively wrong to calibrate against all three fluxes. The authors are far more the experts in this field though so it is up to them of course.

Apologies for this not being clearer - we never optimise against all three. We either do TER and NEE or GPP and NEE. However, the three fluxes are used in the evaluation. This has been clarified in the text:

"**In each case, two fluxes are used in optimisations.**"

Martin De Kauwe