Response to anonymous referee # 2

The manuscript by Meunier et al., uses TLS data to inform coefficients of an ecosystem model's allometric equations and initial conditions, quantifies its impact, as well as testing influence of TLS information on model calibration. While the study is well thought out and generally well-written and visualized, there are some issues with both modelling and calibration protocols (in terms of both technicality and clarity). Also the manuscript remains somewhat inconclusive about the superiority of TLS-informed model predictions, or at least if that wasn't the case the manuscript needs to be revised to clearly present it as such. It's a pity Table 5 was not available for the review process. Overall, I think the study would be of interest for the community and worth publishing, however, I would strongly recommend tackling the technical issues raised by both reviewers. Line numbers below refer to the author's preprint.

First, we would like to thank reviewer #2 for their thorough assessment of our manuscript. We believe that the comments that were raised are fair and will contribute to improve the overall quality of the study. Their suggestions fall in line with the ones of the first reviewer and we agree with both of you that despite being interesting and worth publishing (thanks for pointing that out!), the results were probably not presented in the clearest way. We therefore propose to reshape the manuscript around one central analysis, which will replace and complete the previous three. The concepts, data, and model will remain essentially the same but will be analysed and presented in a somewhat different manner. The new analysis will assess the global sensitivity of the model and evaluate its performance when the simulations are constrained or not by TLS data. More precisely, we will run large ensembles under different model configurations (default and TLS-informed),

and partition the overall uncertainty into its components (initial conditions, model structure, and model parameters).

Model runs will be initialized from near bare-ground or from field inventory (default) or directly from TLS-derived size distribution (TLS). For the model structure, we will test the impact of the crown representation (finite or infinite) as well as of simulating more than a single PFT. Additionally, we will increase the number of parameters to be tested to include those that were shown to be significantly contributing to the overall model uncertainty in previous ED2 studies. This global sensitivity and variance decomposition analyses will allow us to investigate the following research questions that were somehow present in the previous version of the manuscript but not clearly identified nor fully answered:

(i) What is the total model uncertainty? What are the contributions of the different sources of uncertainty? And how do they change along the simulation?

(ii) Is the total model uncertainty reduced when constraining the model to TLS observations? Do the primary sources of uncertainty remain the same?

(iii) Does the use of TLS data improve model performance?

To answer the last question, we will use the independent datasets that were previously used for model calibration (eddy-covariance fluxes, LAI and light observations) as well as recurring forest inventories as validation datasets and compare how the ensembles reproduce those. Most of those data were included in Table 5 and we are sorry that it somehow disappeared during the submission process. We are very confident that such a reshape of both the study and the manuscript will significantly improve the technicality and clarity of the protocole and clearly shows the positive impact of the use of TLS data on model uncertainty and performance.

We also thank reviewer #2 for the detailed review, which is very useful to correct all minor points. However, as we expect the manuscript to change significantly because of the reshape of the analyses, some of the comments will no longer be relevant. For that reason, below we only answered the comments that will remain relevant for the new version of the manuscript and the new analysis. Should some of those comments that we left open appear to be relevant during our revision, we will of course take the reviewer's suggestion into account.

Title: As mentioned in the general comment above, the manuscript is rather inconclusive about the reliability of TLS informed model predictions. Even the abstract reports only the sensitivity of the results to model configuration and TLS information. Hence, it feels as if the title would reflect the study more closely if it was revised to something along the lines of "Sensitivity of ED2.2 forest ecosystem simulations to TLS informed/constrained structure and functions" (as also presented by the authors on L95 and L195).

At the moment, we agree that we mainly tested the sensitivity of the model simulations to TLS-informed structure and functions and therefore a title such as the one suggested would be probably more appropriate. Yet, we think (based on the results presented in the first version) that the new version of the model analysis will allow us to be more conclusive about the reliability of the TLS-informed simulations, and therefore hope that we will be able to keep this (or a similar) manuscript title.

L28: "imposed openness" do you mean the FC configuration here? If yes, please revise to explicitly say "model configuration that imposes finite canopy radius dramatically influenced..."

We indeed meant the crown representation here (finite or infinite) and this will be made clear in the next version of the manuscript.

L33-34: After reading the manuscript, I wasn't quite left with a conclusion about the most adequate model structure. If you identified it, why not say it in the abstract explicitly.

Indeed, the previous set of analyses did not allow to identify the most adequate model structure because we lacked independent datasets (eddy covariance fluxes were used to calibrate the model, and LAI observations were derived from the same TLS datasets which served to parameterize the model). Reshaping the analysis as we suggest above will allow us to reach such a conclusion: we will now be able to see the benefits of constraining some of the model parameters and processes to TLS data on the model uncertainty and performance.

L81: Somewhere around this paragraph I would have expected a brief introduction about other (e.g. airborne) lidar studies with TBMs as well. Especially given that studies exist directly with the ED model, Hurtt et al. 2004 (https://doi.org/10.1890/02-5317), 2019 (https://doi.org/10.1088/1748-9326/ab0bbe), Thomas et al., 2008 (https://doi.org/10.5589/m08-036). I think this could benefit the discussion as well, e.g. what did the authors build upon the previous lidar-ED2 studies? or they can draw parallels to this study.

While very relevant, those studies were unknown to us. They will definitely be included in the introduction and discussion of the new version.

L136-140: Appreciated the length authors went with extracting the data. However, this paragraph would benefit from further information on the overall quality of the data: what the frequency of the data is (daily, sub-daily?), how it was filtered, QA/QC'd, how the GPP was derived, what the accuracy of data retrieval from Plot digitizer software is, if there are known issues with the time series that could affect the calibration and so on.

In the future version of the manuscript, we will provide these pieces of information.

L152: I'd like to point out that the authors themselves avoid using the word "validation" here, which again reinforces my comment about inconclusiveness. In case you decide to strengthen the paper's conclusions, at least consider the word "assessment" here.

The trait data will now serve to inform the parameter distribution before the global sensitivity analyses. Model performance will be assessed with other independent datasets (including the eddy covariance fluxes) that were previously used to calibrate some of the model configurations. In that sense, we will use those observations as validation datasets and will be truly able to compare model performance with and without TLS constraints.

L164: Agreed with the other reviewer. Why was all classified as mid-successional pft in ED? I agree that each species, at least the five on Figure 1, needs reasoning as to which PFT they were mapped to and why. Please also provide citations for mappings when possible e.g. see supplementary on https://onlinelibrary.wiley.com/doi/full/10.1111/j.1365-2486.2011.02477.x where Acer is LH, Quercus is NMH. Admittedly, using multiple PFTs would complicate the reporting as authors are currently only concerned with a single set of allometric parameters, but worth exploring. Also, even if the authors decide to continue with a single PFT after revisions, they should emphasize already here that this is an over-simplification which could help prevent misuse by others referring to this study in the future.

The first version of our study mainly aimed at assessing the sensitivity of the ED2 model to parameters and processes that could be constrained by TLS data. At this stage we wanted to simplify the model complexity as much as possible to render a clear and simple message. In that sense, increasing the number of PFTs, while important to simulate the complexity and diversity of acquisition strategies of the tree species, would have in our opinion made our analyses and conclusions fuzzier and more difficult to grasp. We

propose to overcome the issue by incorporating the number of modelled PFTs in the model structure uncertainty and doing so to quantify the uncertainty that is associated with this simplification. The tree species mapping to the model PFT will be achieved using trait data of SLA and wood density, the allometric relationships derived from TLS, as well as previous classifications from the literature, as suggested by the reviewer.

L192: Agreed with the other reviewer. Please provide more details or point to the initialization/settings files of ED2.2 specifically if you have deposited them to the repository cited at the end (you could have a supplementary table telling which initialization/settings files went with which experiment or populate the readme file on the repository) .

In the next revision, we will add the missing information in the main body and include the settings files of the model in the Zenodo repository.

Figure 2 is great, but I'd call Analysis III: Bayesian calibration instead of data assimilation to be more precise, or at least continue using "parameter data assimilation". Also for analysis I, did you use TLS to inform structure directly? Looking at Table 4 it's only allometries. Allometries in return affect the structure but if I saw only allometries in that box, it would have helped me follow the study better.

Indeed it should have been called “parameter data assimilation” or “Bayesian calibration”. This analysis will no longer appear in the revised version of the manuscript and Figure 2 will therefore be updated.

L202-203: What do you mean by "to assess the relative importance of TLS we compared it to field observations"? Does this exercise result in Fig S1 and S2? Isn't it then better to call this ground-truthing or validation of TLS? Please clarify.

Indeed this corresponds to Figure S1 and S2. We will call this “ground-truthing of TLS” in the revisions.

L207: Could you already explain here if 100 years spinup is enough, especially considering that the actual age of the forest is much older? I know of other models running much longer spinups (e.g. 500 years), please motivate the reader if 100 years is appropriate.

The 100 year-long run does not correspond to a true model spinup but rather to the approximate age of the forest (it is the last time since large-scale disturbance occurred to Wytham Woods to the best of our knowledge). So if the model was perfect, the size distribution as simulated after 100 years should resemble the actual observation. We will make this clear in the revision

L214: Looking at Table 4, how about NBG-infinitely wide-TLS setup? See comment below regarding having another control for impact of TLS informed allometries.

We did not include the NBG-infinitely wide-TLS setup because TLS data allowed us to constrain the crown area allometry. Yet, in our new analysis the NBG-infinitely wide TLS will be necessarily included in the ensembles.

L221: Why not explicitly state in what order these changes and combinations were introduced as this might also help following the incremental effect discussion. Listing configurations for 16 runs is not that much, could be also in the supplementary.

L226-228: Agreed with the other reviewer on quantification of indirect effects. I think listing all the configurations for the mentioned 16 runs will help. I assume authors performed a factorial design here but it is not clear which combinations went with which.

For both previous comments: the new global sensitivity analysis will not use this factorial design anymore so we will get rid of those issues while keeping the main results of the sensitivity analysis.

L231: "parameter optimization by Bayesian data assimilation" -> Authors could consider using "Bayesian parameter data assimilation" here as well to be more clear. Or better yet, "Bayesian calibration of model parameters".

Indeed Bayesian calibration or Bayesian parameter data assimilation sound better.

L235: Looking at Table 4, it feels like there needs to be another intermediate setup: inventory-finite radius-default, is there a particular reason why authors omitted this configuration? Also this sentence on L231-L233 suggests this configuration was included but Table 4 does not mention this configuration: "The model configurations included a default model version (default allometric parameters, infinite crown area), and a finite crown representation (default allometric parameters, finite crown radius), *which were both initialized with field inventory data*" I believe, according to this sentence, Table 4 second to last column should read "inventory" for initial conditions, please clarify. Overall, I think if there were 4 configurations in total it would be more systematic where only one thing would change at a time, 1: inventory-infinitely wide-default 2: inventory-finite radius-default 3:tls-finite radius-default 4: tls-finite radius-tls

This intermediate setup will be included in the global sensitivity analysis and should not have been omitted in the first place.

L237-242: As much as I liked the process-based perspective, a sensitivity analysis (running the model with varying parameter combinations drawn from their priors to see how much change they cause on model outputs) would also be warranted here to formally show these parameters are indeed constrainable by the fluxes. Also the authors might be missing some other important model parameters (although there may be many parameters that can be calibrated as authors suggest in the discussion, models are typically most sensitive to maybe a dozen or so). I.e. calibration might be pushing SLA and Vcmax to different values in the

parameter space under different configurations, but in fact if other parameters were included in the calibration it may have not been the case. Besides, other aspects of a proper calibration protocol are skipped here. For example, after determining to target these two parameters, authors could vary these parameters in their prior ranges and plot a likelihood surface (if they had done a global sensitivity analysis this would have come for free). This would have revealed the trade-off (negative correlation) before the calibration and would further implicate the need for either more informative priors (see below), or even not targeting one of these parameters in the calibration. I would have understood if authors, so to speak, enforce equifinality and use TLS to resolve it, but that has not been the case in the end (authors only report differences, don't really conclude -validate- which was more accurate). Instead, authors exacerbate the equifinality issue by choosing correlated parameters and uninformative priors only to confirm low identifiability (L390) and mention TLS' potential to discriminate without actually doing so. To sum up, I have three suggestions for the authors: 1) perform a global sensitivity analysis to at least identify other important parameters, even if they decide not to calibrate them it could help discussion, 2) try to repeat the analysis with more informative priors, 3) elaborate on their calibration results (some suggestions below) and strengthen their conclusions (be less vague).

We globally agree with this comment and the proposed new global analysis should address most of the reviewer's issues. Note that we will now include more parameters than just SLA and V_{cmax} based on previous ED2 studies and will use more informative priors based on available trait data. Furthermore, the model performance with and without TLS data will be evaluated with the independent datasets mentioned above.

L246: GPP is not measured but a derived (modeled) quantity, at least as opposed to other carbon (net ecosystem exchange) and water (latent heat) fluxes. How the uncertainties were affected in this case, how was that accounted for in the calibration?

We will add a discussion about the uncertainties associated with the way GPP is modelled from the raw eddy-covariance data. In addition, we will provide more information about the origin and the treatment of the original data in the material and methods, as suggested by the first reviewer.

L254: Sampled how? From marginal or joint posterior distributions? Please clarify.

From the joint posterior distribution but this is no longer relevant in the new analysis that we propose.

Table 3 and Figure S3 Vcmax units are different from L144 and Fig 5, please reconcile.

We will make sure all units are consistent in the new version (V_{cmax} has $\mu mol_C\ m^{-2}\ s^{-1}$ units)

L255 and Table 3: Why were the priors chosen to be uniform? Are values like 5 really equally likely as 30-40 or is 60.5 impossible for Vcmax? I believe given many observations and prior knowledge about these parameters more informative priors could have been chosen, which in return could have reduced the equifinality problem. Please consider distributions other than uniform.

These prior distributions will now be constrained by trait data, whenever those data exist.

Figure 3: Looking at the figure, hard to tell without playing with the raw data, but it almost looks like there could be two lines fitted to Acer (late hardwood) and others (mid hardwood) reinforcing the point about exploring two PFTs above.

We will test that and classify the tree species into the PFTs accordingly (and also based on trait data and previous classifications published in the literature such as the one provided by the reviewer).

L263: If the fitted parameter values of the lines on Figure 3 are on Table 2, please already say so in this paragraph. Also Fig3 caption can refer to Table 2.

We will make these links in the new version.

L267: Although, maybe it is worth noting that both are performing badly at the tails. Please consider providing the residual plots for Figure 3 fits in the supplement.

Good suggestion, we will provide the residual plots for the allometries in the future version of the supplement (with one and several PFTs).

L270: I don't mind these figures being in the supplement, however, I felt like this should have been the first thing reported in the results. How well TLS do with respect to inventory, before moving on to allometries.

Agreed. And based on reviewer #1' suggestion this will probably move to the main text.

Figure 4: It is surprising how big of a difference infinite versus finite crown configurations makes. Was this documented before in previous ED2.2 studies? Is it appropriate to spinup both configurations for 100 years? Can it be that the FC configuration needs to be run longer? Also I believe each of these bars represents a single realization from the model, is that right? I would be curious to see if slightly different NBG initializations in an ensemble mode could have provided a different picture (i.e. Fig 4 but with error bars where some configurations may manage to get into the right ballpark). Plus, NBG ensembles (with different initial conditions and Vcmax/SLA parameter combinations) could provide an additional quantification of the uncertainty reduction (their ensemble widths can be compared to their IC counterparts which were constrained by TLS). Besides, how would the tree size distribution look like if the authors have used more than just one PFT in these simulations (further point, as noted by the other reviewer, it was not clear if the seedlings were MH-only)?

All these issues will be solved with the new analysis (including one *vs* multiple PFTs, ensembles from near bare-ground conditions, and the quantification of the uncertainty reduction). With the new ensemble runs, we will be able to assess if slightly different NBG initializations could have provided different pictures of the tree size distribution. The large impact of finite *vs* infinite crown representation was documented only recently in a publication (Shiklomanov et al., 2019). It is a well-known 'issue' in the ED2 community but probably not outside (yet see Fisher et al., GCB, 2018).

L281-282: This is a clear result, however, (looking at Table 4, as mentioned above) I would be curious to see a NBG-infinitely wide-TLS configuration with TLS informed allometric coefficients. Does the infinitely wide configuration not use the same allometric equations? These models can be highly non-linear, and the response of NBG-FC with and without TLS allometries could be different than the sensitivity of NBG-infinitely wide with and without TLS allometries. And if it is the same result, it would only strengthen this finding.

The new model analysis will account for that missing configuration.

L293, Figure 5: After referring to Fig 3, please try to make it clear here that you are back to referring to Fig 5 here. E.g. "The large variability around those mean relative changes (Fig 5, error bars) ..." Also state in Fig 5 caption what do error bars represent.

Figure 5 will no longer exist in the revision.

L297: Yet, I'd somewhat expect larger aboveground woody biomass could also result in bigger trees -> less understory PAR. Does it imply problems in model structure?

It definitely relates to problems in the model structure (the woody tissues do not account for light interception). It will be discussed in the revision.

L329: Unfortunately, there is no Table 5.

We are sorry that this table somehow disappeared during the submission process. It will be included in the next revision.

L331-334: IC-TLS uses both the DBH distribution and the allometric coefficients informed by TLS. So I assume it was able to capture Figure3 - leaf biomass relationship very well? Sounds like it also produced LAI values in the right ballpark. The link between leaf biomass and leaf area is through the SLA in the model (L238: SLA is used to convert the leaf biomass into leaf area), and SLA posterior of IC-TLS is agreeing with the CWM. So it almost looks like this configuration gives the right answers for the right reasons for these variables and parameters. And if IC-FC is producing the same leaf area as IC-TLS but have different SLA, it must be missing leaf biomass? I'm trying to see if authors could discriminate a bit more explicitly about performances of different configurations here, please consider elaborating as such.

In the new version of the manuscript, we will make sure which configurations perform better for every single independent observation (eddy covariance fluxes, leaf area, basal area, growth etc.). Doing so, it will be clear how TLS data improves model performance and reliability.

L343: Does this contradict or how is it related with the finding mentioned before on L281-282: when NBG is concerned model structure had a bigger impact than allometries on tree size distribution? Also although I couldn't see Table 5, I have a feeling that it will be hard to digest. I recommend authors consider a figure instead or in addition.

This sentence (L343) does not contradict previous findings because in that case we were investigating the allometric parameters only and we did not compare the relative process and parametric errors. This is something we will investigate with the new analysis. The

figures and tables are very likely to change significantly but we will make sure they are digestible or we will increase the number of graphic elements accordingly.

L381-382: Could you be more specific here? The statement is too vague. NBG with TLS informed allometry didn't do any better for capturing the tree size distribution (they also did bad for the ecosystem variables L346). So informing allometry alone was not enough. Is TLS more useful when it is used for prescribing the initial conditions or what? How does this agree with studies in the literature? E.g. does this mean initial condition uncertainty is a bigger problem than allometric uncertainty?

In the new version, we will be able to demonstrate the benefits of TLS to more adequately represent the modelled ecosystem. As explained above, to do so we will compare the performance of the different tested configurations against independent validation datasets. By partitioning the overall uncertainty, we will also quantify the relative benefits of prescribing the initial conditions, the model structure or the allometric equations.

L385-386: I don't know if it is striking, but it was expected given the trade-off and uninformative priors.

Fair enough. This problem will be solved by informing the priors with the trait data in the new analysis.

L388-389: "Very different" but how? Again very vague. Was a particular one any better?

This will clearly emerge from the new analysis.

L395-396: But did it in this study? Does this mean the authors trust IC-TLS posteriors on Fig 7 more? Also please see my comment above for lines 331-334, and try to be more specific. If you did discriminate between equifinal model versions, say it here which did better.

We will in the new version (see our comments above).

L407: But aren't there more formal ways to deal with this? E.g. one could start the model from the past (when flux data is available) with more uncertain IC even if they don't know about the forest structure and composition a decade ago, and then calibrate the model with past data, continue simulations in time and assimilate more recent inventory data to constrain the states? In fact, the Thomas et al 2011 paper cited by the authors already have some useful values for conditions 10 years ago. Furthermore, the Butt et al. citation implies there was a tree census in 2008? (I merely clicked the link)

The recurring forest inventories will be added to the list of validation datasets used to discriminate the model performance.

L412: While it is true that it would increase the overall complexity of the study, I'm not sure it sufficiently justifies simulating one PFT when at least Acer and Quercus are concerned. Could more informative priors be chosen when more distinct PFTs are used? There were also numerous occasions mentioned above where using multiple PFTs could potentially remedy some of the shortcomings. At least without demonstrating it, I'm afraid this argument remains unconvincing.

This will be solved by (i) testing more than one PFTs and (ii) using more informative priors for some of the traits

L418: This statement, although true, seems rather irrelevant for the conclusion of the present study as both SLA and Vcmax are measurable. In general, until the last three sentences, the conclusion reads like an introduction and needs to be tailored towards the study more. I'd recommend starting from what you demonstrated, then telling what the implications of your findings are, how well your results aligned with your prior expectations, if your methodology was adequate, if you got new insights / new ideas for future steps and so on.

We will make sure the conclusion is tailored towards our findings. As we expect those to be significantly impacted by the new analysis, we prefer not to give too much detail here about its future content.

L422-425: Apologies for repeating myself but I overall think the reporting was rather inconclusive as to whether the TLS informed model was indeed more reliable or able to discriminate between equifinal model versions. In other words, yes, TLS-informed results were different but were they more realistic? What was the independent validation? Which configuration got the right answers for the right reasons? Reader has to work really hard to figure it out. You could further provide your concluding recommendation regarding how TLS is best utilized.

All those fair points should be hopefully addressed by the new conclusions emerging from the global analysis.