Response to anonymous referee # 1

The concept for this study is strong, using TLS to constrain forest structure and function in the ED2.2 model, follows a decade and a half's work on using remote sensing to constrain predictions made by ecosystem models (in reducing process and initialization errors). While this idea is worth publishing, the execution is not clear, the structure of the study needs improvement, and the actual constraining of the ED2.2 model is adequately done. Concerning this last point. Essentially you want to know how well your TLS-constrained ED2.2 simulations has fared compared to Ground-based-initialized ED2.2 simulations and compared to bare-ground simulations. To assess the improvement you need to compare all 3 of these simulations to observed data (like GPP, plot basal area changes, and/or growth and mortality dynamics). You need to do this for both TLS-structure and TLS allometric improvements.

We first would like to thank Reviewer #1 for their positive review of our work. We were pleased to see that they share our opinion that TLS has strong potential to successfully constrain forest models and that such an idea is worth publishing In Geoscientific Model Development. After reading both reviews, we decided to thoroughly reshape our study in order to improve the structure of the manuscript and enhance the execution of the research idea. The concepts, data, and methods remain essentially the same but were reshaped and presented in a different manner.

As highlighted by both reviewers, what we are interested in is (i) the sensitivity of the ED2 model to parameters and processes that can be constrained/defined via TLS-derived observations and (ii) how model simulations constrained with TLS data compare to ground-based initialised simulations, bare-ground model runs and independent observations. In order to investigate this and address the main comments of both

reviewers, we replaced the current set of analyses by an overall global sensitivity and variance decomposition analysis applied to different model configurations (near-bare ground, inventory-initialised and TLS-informed). This new analysis encompasses the three included in the previous version of the manuscript and extends them. For each model configuration, we ran ensemble simulations (N = 500) and decomposed the total variance into its components (model structure and parameters). This allowed us to answer three research questions that were already present in the previous version of the manuscript but were not explicit, and so only partly identified and answered. Those research questions are:

(i) What is the total model uncertainty of the ED2 model for the specific site that we try to mimic in silico, and what are the contributions of the different sources of uncertainty?

(ii) Is the total model uncertainty reduced when constraining the model with TLS observations and dothe primary sources of uncertainty remain the same?

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

Model ensembles were started from TLS-derived inventories (TLS configuration), from near bareground conditions (NBG) or field inventories (Census). The importance of the model structure was tested in the different configurations by changing the crown representation (finite or infinite), the number of simulated PFTs (see below for details), the type of radiative model used and the trait plasticity (the summary of configurations is listed in Table 2 of the manuscript) . The impact of model parameters was quantified by changing a larger set of parameters than just SLA and $V_{c,max}$ as suggested by the second reviewer. We included the parameters that were shown to drive a significant fraction of the model uncertainty in previous studies (see e.g. Dietze et al., 2014 and Shiklomanov et

al., 2019) and not only the ones for which we have trait data available. Hence, we could separate the parameter uncertainty into the contribution of TRY-constainable parameters, allometric parameters and other ED2.2 parameters. In the TLS configuration, the allometric parameters were constrained to TLS data while the prior distributions of those parameters remained uninformed in the NBG and Census configurations.

Those ensembles served to quantify the total model uncertainty and partition it into its components. Doing so, we also could estimate the benefits of TLS to reduce the total model uncertainty and how its drivers change when some processes and parameters were constrained by TLS data. Finally, model outputs were compared to the datasets previously used to calibrate some of the model configurations (mainly eddy covariance fluxes, and LAI). Doing so, we were now directly able to compare the performance of the different configurations.

There are many more comments in the text attachment below.

We thank reviewer #1 for this detailed review, which is very useful to correct all minor points. As the manuscript changed significantly, we do not provide here a detailed response for every single comment. Instead, below we repeated and answered only the comments that remained relevant for the new version of the manuscript and the new analyses.

Detailed comments and response

L25: "... at a temperate forest site" As it is a single site, isn't it meaningful to say what it is?

We now name the site (Wytham Woods, UK) starting from the abstract

L27: "... productivity ..." gross primary productivity? net primary productivity? leaf area production?

The global sensitivity and variance decomposition analyses were performed on several model outputs, including gross primary productivity.

L29: "the imposed openness" do you mean the gap fraction here?

By "imposed openness" we meant "the representation of the canopy (infinite or finite crown areas)". We do not use this term anymore in the new version of the manuscript.

L35-39: "We conclude … and reduce their overall uncertainties" Have you validated you new TLS-configured ED2.2 runs? Do you have tested reductions in uncertainty?

Validating the TLS-informed runs with independent datasets (including eddy-covariance data, basal area changes, etc.) was one of the objectives of the new analysis and is now presented in the new version of the manuscript. We showed that the model uncertainty of the model outputs is reduced when simulations are constrained with TLS data, and even more when both TLS and TRY data are used. The TLS also exhibits good performance for reproducing the seasonal cycle of GPP and predicting the ecosystem LAI.

L165-166: "All trees inventoried in Wytham Woods were classified as mid-successional temperate deciduous trees" Are you sure. Sycamore can be considered a shade-tolerant tree (late successional), maybe hazel. You need to go through each and every species to see which are early, mid and late successional species. Identifying them all as mid-successional needs a good explanation.

As the manuscript was initially seen as a simple sensitivity analysis of the ED2 model to the parameters and processes that could be constrained to TLS data, we wanted to simplify the analysis as much as possible and hence decided to simulate a single PFT. Incorporating more than one PFT would have made the whole analysis way more complicated to present, as well as for rendering the outputs to the readers. Yet, with the new global analysis, it was possible to take this complexity into account by simulating one or several PFTs and include that uncertainty into the model structure error. The tree species mapping to the model PFT was achieved using previous classifications of the literature and the allometric relationships derived from TLS. Our analyses showed that due to the compensatory effect, the contribution of the number of simulated PFTs to the overall model uncertainty was only a small fraction of the total variance (see Figure 4 in the new version of the manuscript) .

L191: "i) near-bare ground initial conditions (i.e. seedlings only" Is this with Mid-Successional Hardwoods seedlings only?

In the previous runs, it was indeed with Mid-Successional Hardwoods only. In the new runs, it is with Mid-Successional Hardwood trees or a combination of Mid- and Late-Successional Hardwood trees (see comment above).

L193-194: "that were allowed to fuse along the simulation" Unclear. Delete?

This part of the sentence was deleted.

L194-195: "Simulations were run … dataset (Viovy 2018)" We need more information on the initializations. How long was the bare-ground simulation run for? How did you cycle the years of the met data? Which PFTs did you use? What soil data were used at the site? What is the resolution of your met data? What are the structural attributes of the ground vs TLS initial forests...i.e. Basal Area, DBH, LAI, TLS occlusion etc.

Those pieces of information were provided in the new version of the manuscript. The NBG simulations were run for 100 years (the approximate age since last large-scale disturbance) and forced with the corresponding years of the CRU-NCEP dataset. For the soil data, we used soil texture analyses from the literature. We also now compare the tree size distribution of the inventory vs TLS (Supplementary Figures S4 and S5).

L214-216: Can you have results for this analysis and comparisons to ground-inventories?

In the revised version, all near bare-ground runs are compared to ground inventories as an independent validation (Figure 6).

L231: "Parameter data assimilation and model equifinality (Analysis III)" Rather than equifinality, this method appears to be about parameter optimization. Perhaps I am wrong.

We agree with the reviewer's comment. Analysis III is no longer included in the new version of the manuscript (no more parameter optimization was done in the new analysis).

L261-264: I can't find these numbers in the figures

Those numbers (re-computed as RMSD) were added in the new version of Figure 3 (now Figure 2 and supplementary Figure S2)

L265-271: These results are not clear. It is assumed you are describing Figure 3, but numbers comparing TLS and ED2 are not clear.

We now explicitly refer to the appropriate figure (now Figure 2).

L273-275: TLS and ground differences are important here and should not be in the supplement. Valuing the use of TLS in ED2.2 needs a rigorous comparison to ground-inventory data and initializations.

We expanded the discussion about the differences between TLS and ground-inventory and improved the figures comparing the TLS and the inventory tree size distribution

(Supplementary Figures S4-S5). Yet we decided to keep those in the supplementary file as it is not the main focus of this paper, and what this analysis reveals is essentially an almost perfect correlation between TLS and ground-inventory data.

L287: Start off making reference to the Figure you are about to explain

We made sure we properly start off making references to the appropriate figures in the new version of the manuscript.

L304: It is still unclear why you are doing this step, and why you are not including the new optimized results in the sensitivity analysis (Figure 5). Also it seems counter-intuitive that IC-Default Vcmax value of 47.3 compared to 17.5 would not largely affect GPP!

The V_{cmax} definitely influenced the modelled GPP but the effect of the parameter was compensated by the reduction in SLA resulting from the optimization, which eventually led to similar GPP. In the new analyses we replaced the optimizations by a global sensitivity analysis and compared the (sub-)ensembles with the (now) independent eddy-covariance data.

Figure 7: Confusing figure. What this tells me is that the TLS and FC initial conditions are not that close, especially with the results in SLA. How useful is this optimization, and why do this if you are not going to use it in the sensitivity (Figure 5), or in ultimately comparing the improvements of TLS vs ground-inventory initializations against carbon fluxes?

The optimization analysis was removed from the new version of the manuscript.

Figure 7 (bis) What does this mean?/What is this dot?

The dots were the data from the TRY database. Yet, this figure was removed from the new version.

L355: There is no Table 5

We are sorry that Table 5 was missing in the manuscript. It somehow disappeared in the submission process. Those data are now available in the new version (Table 6).

L359-365: It may be expected that better site level descriptions of BLeaf allometry will affect the LAI and GPP, and that better descriptions of Bdead will affect Woody biomass. Rather than sensitivity, what about overall improvement to LAI, GPP, Biomass??

Following yours and reviewer #2' suggestions, we now included this analysis in our study. Site-level leaf and woody biomass allometries improved the performance of the model for the simulated LAI, GPP, and AGB.

L383-384: Not sure you have done showing the effective improvement in model performance.

The new set of analyses allowed us to conclude about the effectiveness of using TLS data to constrain the model uncertainty and improve model performance. In the new version we show a strong reduction in uncertainty and a significant improvement of model performance thanks to the TLS data for multiple state variables independently observed (LAI, GPP).

L397-398: From your results, this 'equifinality' issue could be demonstrated without TLS, just by optimizing using bareground and ground-inventory initializations. Furthermore, the usefulness or added value of this equifinality/optimization excercise is still not clear.

The optimization exercise is no longer present in the new version of the manuscript and equifinality still emerges from the large ensemble runs.

L403-L404: You could use recurring forest inventories collected at your site as an additional calibration dataset

We did not use successive inventories from Wytham Woods in the first version of the manuscript. Since we could not get access to such information and we could not find any relevant data in the literature, our analysis only relies on a single forest inventory.

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 raised were fair and contributed to improving the overall quality of the study. The suggestions fall in line with the ones of the first reviewer and we agree the results were not presented in the clearest way.

As explained in the response to reviewer #1, we decided to reshape the manuscript around one central analysis, which replaced and extended the previous three. The concepts, data, and model remained essentially the same but were analysed and presented in a different manner. The new analysis assesses the global sensitivity of the model and evaluates its performance when the simulations are constrained or not by TLS data. More precisely, we ran large ensemble simulations under different model configurations (near bare-ground, inventory-initialised and TLS-informed), and partitioned

the overall uncertainty into its main components (model structure, and model parameters) for each configuration.

Model runs were initialised from near bare-ground (NBG), field inventory (Census) or TLS-derived size distribution (TLS). For the model structure, we tested the impact of the crown representation (finite or infinite), the choice of radiative transfer model, simulating more than a single PFT, and the trait plasticity (see Table 2 in the manuscript). Additionally, we increased 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 (see Table 4, 13 parameters in total). These global sensitivity and variance decomposition analyses allowed 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 used independent datasets that were previously used for model calibration (mainly eddy-covariance fluxes and LAI) as validation datasets and compared how the ensembles reproduce those. Most of those data were included in Table 5 (now Table 6) and we are sorry that it disappeared during the submission process. We are very confident that such a reshape of both the study and the manuscript significantly improved the quality and clarity of the protocol and now 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 was very useful to correct all minor points. However, as the manuscript changed significantly, some of the comments were no longer relevant. For that reason, below we only answered the comments that were still relevant for the new version of the manuscript and the new analysis.

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).

We agree that we mainly tested the sensitivity of the model simulations to TLS-informed structure and functions in the previous version of the manuscript. Yet, we think that the new version of the model analysis allowed us to be more conclusive about the reliability of the TLS-informed simulations, and the increased performance when integrating TLS data. Therefore we would prefer to keep this 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) but we no longer use this term in the new 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.

We agree that 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 described above allowed us to reach such a conclusion (TLS configuration with TRY constraints) and we now repeatedly state it in the paper, e.g. in the abstract.

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.

We thank the reviewer for pointing out these very relevant references that were unknown to us. We now included these references 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.

Unfortunately, there is not much that we can add on top of what was already described in the original publication reporting those data. Therefore, we now explicitly refer the reader to the Thomas et al. (2011) paper for more 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 now serve to inform the parameter distribution before the global sensitivity analysis. Model performance was now 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 now use those observations as validation datasets.

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 the manuscript 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. 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. In the new manuscript, we

overcame the issue by incorporating the number of modelled PFTs in the model structure uncertainty and doing so we now quantified the uncertainty associated with the simplification of simulating a single PFT in the new analysis. The tree species mapping to the model PFT was achieved using the allometric relationships derived from TLS, as well as previous classifications from the literature, as suggested by the reviewer. The best mapping was indeed to classify Acer as Late-successional Hardwood trees and all other species as Mid-successional Hardwood trees.

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) .

We now included 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.

We agree with the reviewer that “parameter data assimilation” or “Bayesian calibration” was more appropriate. This analysis no longer appears in the revised version of the manuscript and Figure 2 was removed.

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 rephrased this sentence:

“To assess the relative importance of TLS for the model initialisation, we compared the tree size distributions obtained from the field inventory and from the TLS data and computed the absolute and relative differences between both DBH distributions (ground-truthing of TLS).”

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). So if the model was perfect, the size distribution as simulated after 100 years should resemble the actual observation. However, 100 years are sufficient for the model to grow virtual trees larger than the largest trees observed in Wytham Woods.

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.

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 does not use this factorial design anymore so we got rid of those issues while keeping the main results of the

sensitivity analysis (relative contribution of the allometries to the overall model uncertainty).

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".

We agree with the reviewer's comment. Since the analysis was reshaped, this section title was removed from the main text.

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

Following the reviewer's comment, the new analysis redefined those configurations.

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 generally agree with this comment and the proposed new global analysis addressed most of the reviewer's issues. Note that we now included more parameters than just SLA and $V_{c,max}$ for the parameter uncertainty based on previous ED2 studies and used previously defined priors and posteriors based on available trait data (see Table 14 with 13

parameters). Furthermore, the model performance with and without TLS data was 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 added a discussion in the study limitations section about the uncertainties associated with the way GPP is modelled from the raw eddy-covariance data:

"In addition, we know that GPP is not directly observed but rather a derived (modelled) quantity, at least as opposed to the net ecosystem exchange of carbon and the latent heat flux of water. Unfortunately, we could not access water flux raw data nor were they reported in publications that we knew of. GPP uncertainties were also not quantified in the original publication of Thomas et al. (2011). While NEP values were reported, validating the model simulations with those values would have biassed our analyses as we could not constrain respiration parameters with data."

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

They were sampled from the joint posterior distribution but this is no longer relevant in the new analysis that we ran.

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

We made sure all units are consistent in the new version ($V_{c,max}$ 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 were redefined and/or constrained by data, whenever those data exist (SLA and $V_{c,max}$).

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 tested the possible classifications and retained the one that minimised the Watanabe information criterion. It indeed led to classifying Acer as late hardwood and all other species as mid hardwood. In any case, the number of considered PFTs had a very limited impact on the model uncertainty.

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 made 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.

We thank the reviewer for the suggestion. We now provide the residual plots for the allometries in the new version of the supplement (with one and several PFTs), see Supplementary Figure S3.

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.

Those results were moved to be the first thing reported in the results. However, we kept the figures in the supplement because as they essentially show a perfect correlation between TLS and the ground-inventory data, the figure is not especially information-rich.

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 were 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 were able to show that slightly different NBG initializations could provide different resulting tree size distributions (new Figure 6, keeping in mind that that is drawn from the "vegetated" simulations only and many others showed zero vegetation after 100 years). 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 completely redefined the configurations.

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 does 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 is now discussed in the revision:

"...more woody biomass does not translate into exacerbated light interception."

L329: Unfortunately, there is no Table 5.

We are sorry that this table somehow disappeared during the submission process. It was included in the new version as Table 6.

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 made clear which configurations perform better for every single independent observation (eddy covariance fluxes, leaf area).

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 now investigate in the revised manuscript. The figures and tables changed significantly but we tried to keep them digestible, as suggested.

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 were able to demonstrate the benefits of TLS to more adequately represent the modelled ecosystem. To do so we compared the performance of the different tested configurations against independent validation datasets (see previous comments). By partitioning the overall uncertainty, we also quantified the relative

benefits of prescribing the initial conditions, the model structure and the allometric equations (see Figure 4 in the manuscript).

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

This problem was 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 now clearly emerges from the new analysis (TLS + TRY constraints), see Table 6 and Figure 7 as well as the corresponding pieces of text.

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 now discriminate between equifinal model versions in the revised manuscript (see previous comments).

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)

Unfortunately we could not access the raw data of previous tree inventories. As for the time lag between the flux data observations and the inventories/TLS, we are aware of this limitation that we discuss in the paper (first paragraph of the study limitations subsection). In a nutshell, we assume that in 8 years time, the forest composition and tree

size distribution should not have changed dramatically enough to alter the main conclusions of this study that address the sources of uncertainty of the ED2.2 model more than the link between functional ecosystem composition and land fluxes.

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.

We now solved this issue by (i) testing more than one PFTs and (ii) using more informative priors for some of the traits (see comments above)

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 tried to make sure the conclusion was tailored towards our findings, by completely revising it.

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 points should be hopefully addressed by the new conclusions emerging from the global sensitivity analysis.

Response to the Executive Editor comment

Dear authors,

We have checked your manuscript, and unfortunately, at the moment, it does not comply with our 'Code and Data Policy'. Currently, you archive the scripts that you use in Github. However, as we state in our policy and Github on its website, it is not a suitable repository for long-term archival.

Therefore, please, move your code to one of the suitable repositories that we list before the end of the Discussions period and make the necessary changes in the manuscript in potential reviewed versions. Be aware that failing to comply with these rules will prevent your manuscript from being considered for publication. https://www.geoscientific-model-development.net/policies/code_and_data_policy.html#item3 Also, you have included the link to Github of the ED-2.2 model, however, you must cite the corresponding Zenodo repository, as again Github is not a secure repository. The Zenodo repository for ED-2.2 is: https://doi.org/10.5281/zenodo.3365659. Please, remember using the corresponding DOI to cite it in the text.

Best regards,

Juan A. Añel, Geosc. Mod. Dev. Executive Editor

Dear Juan A. Añel,

We would like to apologise for not complying with the code and data policy of GMD. The new version of the manuscript includes the links and DOIs of the Zenodo repositories for both the ED2 model and the scripts and data that are necessary to repeat the analyses.

On behalf of all co-authors,

Félicien Meunier