Reviewer #1

Thank you for the revision of this study. This version of the manuscript has seen much thought and improvements. Particularly, the combination of various uncertainties to GPP, LAI, AGB, PAR variances are informative and are presented well in the pie charts. Essentially the study investigates the contribution of process, parameterisation and initialization errors on carbon flux predictions mainly, using a combination of TLS and non-TLS specifications. This is worth pursuing and much of the information is already there, but the study could be better structured to best show the power of TLS in constraining ED2.2 simulations at Wytham Woods.

We thank the reviewer for both the positive evaluation of our in-depth revisions and more generally of the quality of our work. We were glad to read that the reviewer had seen much improvement between the two versions of the manuscript. We also take this opportunity to thank the reviewers from the previous round whose suggestions led to this revised version and greatly improved the quality of this work.

Essentially, the error reductions not directly associated with TLS can be presented separately first and then the different TLS components can be presented more clearly. Currently the information is there but it gets lost in the complexity of results displayed and presented.

There are indeed two ways to present the results of this study. The first separates the error reductions associated or not with TLS, and presents them sequentially (the option suggested by the reviewer). The second mixes them and discusses the sources of error reductions all together (the option we had chosen for the previous version of the manuscript). For our revised version of the manuscript we decided to go for an intermediary option. We would indeed like to keep the illustrations (e.g. Figures 3 and 4) and the results section structure as is but edited the presentation of the results, e.g. to present the reduction of the uncertainty caused by prescribing the initial conditions and constraining parameters with TRY data on the one hand and quantify the added value of using TLS on the other, as suggested by the reviewer.

The reasons why we kept the illustrations and most of the results section structure as in the previous version are multiple. First, we would like to point out that the current way we present the results (a grid of subplots/pie charts with the different initial condition and TRY constraint configurations as in Figures 3 and 4) encompasses the suggestion of the reviewer and even exceeds it. At the end of the day, the error reduction not directly associated with TLS is the change of the variance (and of its components) going from the "NBG-without TRY constraints" to the "Census-with TRY constraints" configuration. While the added value of TLS to the model configuration with the lowest uncertainty can be grasped from the change of uncertainty (and of its components) from the "Census-with TRY constraints" to the "TLS-with TRY constraints" configuration. So comparing specific pairs of pie charts or subplots in Figures 3-4.

Secondly, presenting the outputs from all model simulations together allows the reader to compare (i) the model performance and uncertainty (e.g. Figure 3) and (ii) the model variance and its components (e.g. Figure 4) directly between these specific pairs of configurations as well as others, which might be relevant in some cases. In this study, we had indeed the chance to have close-in-time inventory data, TLS-derived allometries and size distribution, species-level trait data from TRY and eddy-covariance data (or rather we selected the site for that reason). Yet, field inventories and/or trait data are not always available to reduce the initialization and the parameter errors. Therefore, presenting together the benefits of adding different data sources enables the comparison between the specific configuration that an interested reader could experience for his/her site as well as the potential added value of TLS for that site.

Finally, in our opinion it is important both from a theoretical and a practical point of view to weigh the relative and absolute benefits of different data sources to reduce the model uncertainty and increase model performance. Theoretically first, identifying the most important sources of model uncertainty provides insights and a better understanding of the functioning of such a nonlinear model. Our presentation also allows one to directly compare the benefits of TLS with respect to other data sources and hence its comparative value compared to other observation types, which might be important for field campaigns that are time- and resource-limited. Presenting the error reduction first with inventory and

trait data and then from that configuration to trait + inventory + TLS would make our study less systematic and comprehensive.

Yet, we agree with the reviewer that the suggested way of presenting the results also has its advantages. Therefore, in the new version of the manuscript, we added more direct description of the error reduction from the "NBG-without TRY constraints" to the "Census-with TRY constraints" configuration (i.e. the error reduction not associated with TLS), from the "Census-with TRY constraints" to the "TLS-with TRY constraints" configuration (i.e. the added value of TLS with respect to the configuration with the lowest uncertainty), as well as from the "NBG-without TRY constraints" to the "TLS-with TRY constraints" configuration (i.e. the added value of TLS with respect to the configuration with the highest uncertainty). We added/adapted the following paragraphs in the results section:

"When both parameters were constrained and realistic initial conditions were prescribed to the model (i.e. going from the NBG-without TRY constraints to the Census-with TRY constraints configuration), the variability of the simulated GPP experienced a three-fold decrease. Similarly, the variability of LAI (supplementary Figure S6-7) and AGB (supplementary Figure S8) was drastically reduced, with a four-fold and and a two-fold decrease respectively."

Given the similarities of the tree size distributions derived from the inventory and TLS (see results section 3.1), prescribing initial conditions had a similar impact on the variability of the outputs for the TLS and for the Census configurations. Combined with the constraints on allometries, it led to a reduction of the ensemble standard deviation for GPP in June to 3.78 µmol m^{-2} s^{-1} for the TLS configuration without TRY constraints. As for the Census configuration, constraining SLA and $V_{c,max}$ with TRY data had a larger impact on the model uncertainty: ensemble standard deviation of GPP in June for the TLS configuration with TRY constraints decreased to 1.54 µmol m^{-2} s^{-1}."

"All in all, the predicted variability of the ecosystem LAI and GPP was the lowest for the TLS configuration with TRY constraints: 3.79 ± 0.50 m^2 m^{-2} for the ensemble mean (± one standard deviation) of the ecosystem LAI (Supplementary Figure S6), 9.86 ± 2.89 µmol m^{-2} s^{-1} for the ensemble mean (± one standard deviation) of the ecosystem GPP (Figure 3), both

during leaf-on conditions, which compared well with independent observations (Table 6). The confidence interval of the simulated ecosystem GPP in June for the TLS configuration with TRY constraints was significantly reduced (11.8 - 17.6 µmol m^{-2} s^{-1}) and much closer to the confidence interval of the observations (11.5 - 14.6 µmol m^{-2} s^{-1}).

Reducing parameter and process errors is crucial for TBM studies. In this study the purpose of doing this is to investigate component errors, yes, but also to present a model configuration that has the lowest uncertainty before implementing TLS related constraints.

This is probably where we disagree with the reviewer. Why would we necessarily need to examine the added value of TLS as compared to the configuration with the lowest uncertainty? In other words, what should we only consider the benefits of TLS-related constraints on top of all other data sources, and not rather also in replacement of a part or of the totality of these other data sources? Modellers can sometimes find themselves with no data sources (no inventories, no field data) to reduce the model errors. Our presentation provides an estimate of the benefits of TLS scanning in such situations, together with the benefits of incrementally adding TLS data to trait data, field inventories and both. We also added in the results section:

"*Incrementally adding the TLS-related information to the Census-with TRY constraints configuration had a positive, yet more limited effect on the reduction of the model variability of GPP: ensemble standard deviation of GPP in June was reduced by 30% between the Census and TLS configurations with TRY constraints. Constraining allometries with TLS had a more significant impact on LAI (supplementary Figures S6-S7) and AGB (supplementary Figure S8), with a three-fold decrease of the ensemble standard deviation from the Census-with TRY constraints to the TLS-with TRY constraints configurations.*"

and

"*In total, the variability of the simulated GPP experienced a four-fold decrease when parameters were constrained, realistic initial conditions were prescribed, and TLS data were used to constrain the allometries (i.e. going from the NBG-without TRY constraints to the TLS-with TRY constraints configuration).*""

Therefore, perhaps it is better to present the process errors first (crown size / RTM / PFT) for the NBG and census simulations. Second you would present the parameter errors (from Table 4). Here you can present the TRY parameters including all other parameters from Table 4. What will come out of this is a) the contribution of these parameter and process errors to ED2.2 carbon flux variance, and b) The model configuration that more closely matches GPP observations (and LAI/AGB/PAR).

Once you have achieved this, you will delve into the TLS benefits to constrain structure and function in ED2.2, as stated in the title and in your study aims. The TLS benefits are a) Initial ecosystem structure – INITIALIZATION ERROR; b) Allometry (bleaf, tree height, AGB, crown area) – PARAMETERISATION ERROR components; and c) SLA, reflectance, clumping – PARAMETERISATION ERROR components. You can present each of these components combined, or separately – or as part of the pie chart format you created. Currently allometry (dark green) and initialization error (inferred in circle size) are in Figure 4 (S7-S9). It would be good to separate these from the process errors and other errors not directly linked to TLS improvements. This way, the improvements from TLS can be clearly visible.

For the first suggestion, we refer to our previous two responses. Yet, we would like to point out that we modified the order of the sentences to follow reviewer' suggestion to first go from error reduction using IC and trait data, to error reduction from IC + trait data to IC + trait + TLS.

Regarding the TLS benefits, as we describe in more detail below, there are only two (initial ecosystem structure and allometry) - not three - that can currently be considered in this study since theoretical, technological, and technical challenges still exist to derive parameters like clumping and reflectance from TLS data.

The pie charts we present allow one to grasp the benefits of using TLS to reduce both initialization and the parameterisation errors from different situations (when field inventories are available or not, when trait data exist or not). We disagree with the reviewer that we should separate the benefits of TLS from other errors not directly linked to TLS improvements for the reasons mentioned above (e.g. to easily compare the relative

benefits of TLS vs other data sources starting from multiple situations for multiple outputs and multiple error types).

In terms of the third component above (SLA, clumping and reflectance (?) from Table 4), I do not think this has been considered yet. It may be worth thinking about quantifying these parameters using TLS at Wytham Woods and them determining their potential to constrain ED2.2. If not possible for this study, you could cover the error reduction you might expect from TLS. All in all, this separation of error components will very clearly show the improvements in model performance TLS data can have.

In recent years, TLS has been indeed used to extract more and more information about canopy and plant structural traits, next to the allometries. Yet, those studies remain as of today exploratory and the methods supporting them are either under development or still suffer from important drawbacks. Therefore, we cannot apply them directly to our TLS dataset to derive additional traits like SLA, clumping, leaf angle distribution or reflectance for our study site. It might be possible in the future though. Therefore we added (L461-466, see new manuscript version with no track change):

"In the future, TLS could inform vegetation models even more. The TLS community is indeed actively working on the derivation of additional tree- or stand-scale parameters from lidar raw data and 3D point clouds. Those parameters include leaf angle distributions (Boni Vicari et al. 2019), clumping (Zhao et al. 2012), and reflectance (Calders et al. 2017), which have been shown to significantly contribute to the overall model uncertainty (Meunier et al. 2021; Shiklomanov et al. 2020; Viskari et al. 2019). Yet, theoretical, technological, and technical challenges specific to each parameter still need to be raised before one can constrain these sensitive traits with TLS in a study similar to this one."

Also: Why not present GPP evaluations for the whole 2 year period?

There are two main reasons why we do not present the evaluation against GPP data for the full 2 years period. First, the year-to-year variability in the dataset is extremely low (see the original publication) and therefore so is the added value of using the years separately rather than aggregated. Second, we could not access local met drivers to force our model simulation and used gridded meteorological forcings instead (CRU-NCEP). Doing

so, the model cannot capture small timescale variations or day-to-day variability. Yet, it can (and does) reproduce the observed seasonal cycle and that is what we decided present here. We added in the discussion:

"In addition, in the absence of locally observed meteorological drivers, we had to force the model simulations with regional datasets that cannot serve the purpose of capturing the day-to-day variability or the diel cycle, which forced us to only compare the modelled and observed seasonal GPP cycle. "

Are you sure the calculations of NEP are correct in table 6? The seem much lower than the difference between GPP and Ecosystem respiration.

The confusion probably comes from the fact that the GPP presented in Table 6 is for the leaf-only period why ecosystem respiration and NEP are for all year round, as indicated by the superscript (1). We re-designed Table 6 and added all flux variables for both the leaf-on period and all year round to make it more clear.

Reviewer #2

Revised submission by Meunier et al. addresses the previous reviewer concerns to a great extent. Re-organizing and re-focusing the paper made it clearer, results are now more intelligible and conclusive. I believe it will be a valuable reference for both the TLS and the modeling (especially ED2) community. I recommend its publication with the following minor comments (page and line numbers refer to the revised un-tracked manuscript):

We thank the reviewer for the very positive feedback and once more for the suggestions from the previous round that significantly improved the quality of the manuscript.

p4.l81: "would be" could be removed?

We removed "would be"

p5.l5 / p16.l14: This is probably a matter of taste but to me, fusing data and models have a more formal statistical meaning where multiple data sources are formally combined in a data assimilation algorithm or a framework that encapsulates DA and analyses around it towards generating a synthesized data product. Whereas here, the authors are merely informing individual parts of the model independently. Therefore, I would prefer terms like "to inform" or "to constrain" (like authors also say in many parts of the text, e.g. p4.l86) instead of "to fuse". In the end, I leave it to the authors but I wanted to point it out.

We replaced the three instances of "fuse/fusion" in the previous version of the manuscript with other terms like "to inform" or "to constrain".

p7.l52: Could you state what this default value for the temperature coefficient Q10 is already here?

Added. The new text reads (L155):

"(...) using the model default value for the temperature coefficient Q10 of 2.4"

p8.L83: Could you state what this dominant soil type is already here?

Added. The sentence now reads (L185-186):

"Soil texture was set according to the dominant soil type (clay), (...)"

p9.l99: Could you be a little more specific here, was the trait meta-analysis run only with TRY data or including TRY data? In other words, was there any other additional data informing this analysis besides TRY?

The analysis was performed with data from TRY only. We now emphasise this in the new version of the manuscript L203:

"The meta-analysis was informed by TRY data only"

p10. l39 Could you elaborate on how the uncertainty of data was accounted for in this Bayesian analysis?

We fitted the allometric models using all the available data and the 'brms' package, which generates the posterior distribution of the allometric parameters. To account for the uncertainty of the data, we used a bootstrapping method. We added a description of such a method (L241-244):

"More specifically, we fitted the parameters of the four allometries of ED2.2 using a Bayesian approach and the 'brms' package of R (Bürkner 2017). To account for the uncertainty of the data, we repeated the same analysis multiple times (N = 100) using random sampling with replacement and aggregating the resulting allometric parameter posterior distributions."

Table 6 caption: Would it be a bit more accurate to say "states and fluxes" as one wouldn't really call GPP/NEP/resp states.

We agree with the reviewer. We replaced "state variables" with "states and fluxes" (see new caption of Table 6).

p14.l38: Was the meta-analysis run with random effects turned off? See Raczka et al., meta-analysis posteriors can be too narrow when that is the case, please acknowledge this here or in the discussion.

Indeed the meta-analysis was run with random effects turned off. We added it in the analysis description (L202) and now discuss this additional limitation (L501-503):

"Third, the trait meta-analysis was run with random effects turned off, which can generate too narrow parameter posterior distributions (Raczka et al. 2018), and hence underestimate the contribution of TRY-constrained parameters (see e.g. Figure 4). A similar analysis including random effects should be repeated to evaluate such an underestimation."

p16.l01 Please correct the typo: use "satisfactorily" or "to satisfactory levels"

Corrected.

p.16.l03-06: Looking at Table 2, in general, I see results and discussion on many of the configurations listed here, except the trait plasticity. This could be a topic of interest for the readers, could the authors report more on it in the manuscript?

We agree with the reviewer and now added the following description of the results (L371-L373):

"Process uncertainty was dominated by the type of crown model (5%) and the radiative transfer model (4%). Trait plasticity only contributed marginally to the overall variance (< 1% on average)."

p17.l17 Please correct the typo: delete "is"

Corrected.