Dear editor,

We would like to thank you and the referee(s) for the constructive comments and much appreciate your help and patience. Despite the considerable experience of the author team, we seem to have been struggling to clearly articulate the thinking behind this study. We have revised the manuscript according to your and the referee's comments and suggestions (below is a point-by-point response). It is our hope that this study may help to end an era in which the use of the ITRDB data for model benchmarking was often discarded based on rather superficial reasoning. In this study, we aim to show that the ITRDB data contain information that can be used to benchmark land surface models.

Kind regards,

Jina Jeong on behalf of the author team

Referee

Why the optimisation is needed

The verification approach is explained in its own section and in that section the optimization is explained (L394-399 and L408-417). This explanation is supported by Fig 9. In a previous version of the text we used the term "optimization" in the context of the verification. We realize that this may have added to the confusion because "parameter optimization" (aka "tuning") in widely used in the modelling community. The approach we applied in the verification does not touch (or tune) any of the model parameters but instead modifies the outcome (in a transparent way), hence, we now use the term "modifiers". The text now reads: "The verification required three additional steps (Fig. 9): 1) The simulated TRW from the largest diameter class were transformed by modifiers (see below) to minimize the two metrics of each benchmark ; 2) the same modifiers were then applied to all simulated diameter classes both metrics for all four benchmarks were calculated using the all-tree data, and 3) the actual verification tested whether for a given metric and a given benchmark the modifier improved for the big-tree sample as well as the all-tree data. Improvement of a specific metric of a benchmark was quantified by subtracting the pre-modified value for that metric from its post-modified value for the all-tree data. A negative value thus indicated an improvement. If this was the case, the benchmarks of the big-tree and all-tree data were said to be consistent, implying that using this benchmark in combination with the ITRDB data would reveal the same model shortcomings as benchmarking ORCHIDEE against TRW data from all-tree networks. Across the 11 sites and for each of the four proposed benchmarks, sites where the test improved for both datasets were counted to estimate the confidence in using ITRDB in benchmarking LSMs."

If the agreement between the two datasets was tested for statistically in any way.

This question comes with a nuanced answer. Each individual benchmark is based on well-established statistics (RMSE and regressions) and the modifier is embedded in well-established statistics (RMSE). The change in RMSE and regression as shown in Table 2 is not subject to a statistical test. Note that the objective of comparing the datasets is not to establish whether there are significant differences but to establish whether the direction of the conclusions (e.g., an over or underestimation) based on ITRDB would similar to the direction of the conclusions based on the biomass network. If the direction of the conclusions is the same, their significance has little to no meaning in this context.

I am wondering if it is possible to show plots of the bias and the slope (like in figures 4-7, last panel) for a few sites,

Fig. 10 was added to the main texts and shows exactly the details of the benchmarks for four (out of 11) individual sites. We show 2 deciduous and 2 conifer sites and selected these sites randomly (first sites after sorting them in alphabetical order).

Editor

---

it remains unclear in how far the new benchmark methodology is applicable. I agree that more information and a clear presentation is required to show this. For example, the optimizations and model parameter adjustments are challenging to interpret in the context of applying and benchmarking LSMs. E.g. L375 states that ORCHIDEE parameters were adjusted to ITRDB sites while L426 states that also the default parameterization was used.

This text was revised to avoid confusion. As a summary; five model parameters where manually tuned against 10 ITRDB sites when developing the model functionality to simulate tree ring width (which was well before this study started). These parameters are either global (apply to all forests irrespective of their PFT and location) or PFT-specific (irrespective of their location). Since the start of this study no additional parameters were tuned, all the results are thus based on the default parameter settings.

Table 2 shows an improvement or deterioration of benchmark metrics if optimization parameters are applied - with mixed results and for few sites -, but from the discussion of these results it remains largely unclear what this really implies for being able to use ITRDB for the proposed purpose [continued]

The section explaining the verification test was edited for clarity. We refrained from using the term “optimization” because the latest comments made us realize that this caused confusion (L394-399 and L408-425). L384 (L375 in the previous manuscript) was modified for consistency in the model parameter description. We elaborate on the implication around lines L546-550 and L556-567.

[continued] including information on how representative the experiments in this study are for ITRDB

Because of the natural variation in forests and site conditions, it is near to impossible to show applicability based on a proof of concept (which is given in this study). Nevertheless, we added Fig. S1 in which we show which domain in climate space was tested in this study. We compare the climate space of the study domain (i.e. the European biomass network) with the climate space of the ITRDB and the climate space of forest biomes. This new figure is referred to in the text around lines 67, 358, and 566.

Section 4.1: The counts and corresponding percentages of improved metrics with optimization appear not to agree with the bold boxes in Table 2. Please check and/or clarify.

Thanks for noticing. This has now been corrected in L433 and 435. Also, bolding in Table 2 was reversed to highlight the cases in which the verification scheme was consistent between the two datasets.

Discussion: Subheadings may aide in better following the presentation.

Done.

Please check captions of supplementary figures, i.e. Fig. S2-S4 show diameter.

We have checked the captions and confirm that they are correct. These graphs show indeed diameter instead of tree-ring width. Young tree and mature tree benchmarks use diameter growth (which equals tree ring width).

Regarding the clarity of the manuscript, i.e. the modelling, data processing and optimizations, it may be helpful to include a diagram of the study design and workflows

We added a figure that depicts the workflow and shows the dependencies between the different parts of the background and the method section (see Fig. 2).

I recommend updating the code repository, especially if this may aide in reproducing your experiments during further review.

The repository was cleaned and updated.