We would like to thank the reviewers for their time and effort in reviewing our manuscript. The review comments pointed us to parts of the manuscript that could be simplified and parts that need additional clarification. We are confident that we can address the vast majority of the review comments in a revised manuscript. Addressing the referee comments will require simulations at new and more sites, preparing new figures, and analyzing new results and will thus result in major revisions (see below).

Major discussion

Referee #1

Having read the whole paper, it remains unclear to me if the method presented here actually works and in fact, how can we tell if it does work. I suspect that the answers to these questions lie in figure 4, but this figure is very hard to understand. First of all, it comes before the figures illustrating how each of the benchmarks works, so it is not clear to the reader what the different quantities in the figure are. It would help if figures 5-8 came first. Secondly, I am unsure about the comparison between the model and the BACI data here. As far as I understand, the model is run at different sites than the ones in the BACI data. Why is this and are the sites similar in terms of their climate and stand characteristics? Also, why does this figure only show coniferous sites? Is this an issue with the available BACI data? Would the results look similar for broadleaf deciduous sites?

We agree with the major comments of referee #1 concerning the complexity of the manuscript. Our initial idea was to use the BACI data to show that the assumptions required to use the ITRD data in model evaluation are acceptable. We then applied the model on the ITRDB data as a case study. The referee comment made us realize that the ITRDB simulations are not necessary to demonstrate the validity of the proposed benchmarks and we therefore propose to limit the analysis strictly to the BACI data. In a revised manuscript the validity of the proposed benchmarks could be tested by using 1) the BACI data and ORCHIDEE simulations for all trees; and 2) the BACI data and ORCHIDEE simulations for the biggest trees. Such an analysis could demonstrate that ITRDB data which typically contain the biggest trees in a stand contain useful information about the entire stand and can therefore be used for model development and evaluation. It is clear that this change in model experiment will result in many changes in the manuscript, likely including a more profound verification and discussion of the proposed benchmarks. Because the study will only use the BACI data, we hope we will be able to present simulations for both deciduous and conifer stands without making the results and discussion overly complex. We think this new and simplified approach will better illustrate the strengths and weaknesses of the proposed benchmarks and will better demonstrate which of the proposed benchmarks can be used with ITRDB data and which should not.

I'm not sure I understand why the four different models are needed, if the specifically stated purpose of this paper is not to evaluate the model. It adds an extra level of complexity that makes an already long and complicated paper even longer. If the application of the four model versions is insightful, it would help if this was discussed somewhere.

The four model configurations are a leftover of the initial test but the referee is right in questioning their need for the purpose of this study. A revised manuscript will only report the configuration labelled "Ndyn" which is now the ORCHIDEE default.

Referee #2

My biggest question or I hope to read from this paper is why and how this new approach works. For example, the data-based evidence is needed for why the size-related trend in diameter increment should be unique enough to be used as a character to distinguish different sites with different past century's climates. Why the diameter history, which contains not only the current year's growth signal but also carries previous years bias (possibly), was used to evaluate whether model performs well in diameter increment pattern in both young and mature trees? And the European regional case study didn't give a clear conclusion for the whole benchmarks.

We read this comment as an inquiry about the foundation of dendrochronology and its value for benchmarking models. The foundation of dendrochronology rests on the observation that at the site-level trends in tree-growth contain valuable information about the ontogenetic growth during establish and endogenous competition from canopy closure. If the tree-ring record is long enough a single stand may have experienced different environmental conditions. From a modelling point of view the first challenge is to simulate the response in tree growth to these environmental changes. A second challenge that could help to achieve the first is to simulate with a single model and a single set of parameters, the growth from sites which experienced different environmental changes. From this point of view, changes in diameter growth due to environmental changes can be used to benchmark models (but it requires that size related growth trends are accounted for). Simulating growth trends should therefore be prioritized over matching the endpoints in diameter.

We think the questions of referee #2 are the result of a too concise introduction in the manuscript about the prerequisite of tree-ring research. The revised manuscript will elaborate more on this context. Also, we think that the new approach that will be used to address the concern of referee #1 will help to clarify the first question of referee #2: why and how the proposed benchmarks (don't) work.

Figure 5: The exhibited slope estimation at Panel (d) looks not that convincing. The flat slope is heavily influenced by the big continuous underestimation for the young growth. And there is an obvious downward trend since the tree getting bigger. The slope estimation could make more sense (or be more robust) if data (difference) could be randomly arranged, not by age; or if it is not showing the consistent longer-term difference in either the positive or negative way for a certain period.

The suggestion from referee #2, which is randomizing the residuals of the tree-ring trend, is creative but we don't understand how it could overcome over interpretation. The current analysis shows no trend for the ordered residuals with as the referee noted years of underestimation followed by years of overestimation. If we would randomized such residuals we would most likely find no trend in the residuals. If we misunderstood the proposal of the referee, we are open to adjust the statistics following new instructions. It should be noted that for this study, the trend in tree-ring width contains the information we are seeking to use as it

quantifies the change of tree growth with time. The information gets its importance from anthropogenic-driven long-term environmental changes, for instance, an increase in CO2 and nitrogen deposition. Randomization would break the growth trend over time and would hide the information we are looking for. The lack of a slope in Fig. 5 (d) was mostly driven by the trend for the established stages and implies that the model mimicked the trend relatively well. Note the benchmark combines the slope (Y-axis) with an RMSE (x-axis). The absence of a slope with a high RMSE suggests that there are is substantial over and underestimation of the trend. A good model is expected to result in a zero slope and a low RMSE. In the revised manuscript we try to better explain why the benchmark targets the trend in growth.

Figure 8: It looks like the extreme event benchmark is the most climate-sensitivity related benchmark. However, the period is limited for the most recent years when the most reliable observed climate data is available, which is not consistent with the other three benchmarks. This somehow downsized the importance of this new benchmarking method. (Because the longer-term benchmark is one of the major breakthroughs.) Does this mean the other three benchmarks are not that sensitive to the quality of the climate data, especially to the climate variations?

We largely agree with the referee but this comment made us realize that we need to improve the flow and presentation of the manuscript. We are indeed looking for a long-term benchmark because those are rare. Tree ring records go back far enough in time so we selected these records. We then looked which known issues with the ITRDB data should be taken into account such that these data could still be used (some of the known ITRDB issues can only be addressed by adjusting the model. This process resulted in four possible benchmarks). Two out of the four proposed benchmarks are long-term. We agree with the referee that the time horizon of the extreme event benchmark is not long enough to qualify as long term. We would not conclude this downsizes the importance of the new method. Rather our study proposes four different benchmarks that could be used to evaluate different aspects of simulating historical growth. In a revised manuscript we will try to better clarify this reasoning and we will improve the current table 1 such that it systematically describes the characteristics of each benchmark and discusses the implications for model evaluation.

Specific comments

Specific comments, for example, line numbers, the structure of the discussion, the order of the figures, figure caption and tables will all be addressed in the revised manuscript. The remaining specific comments are discussed below

Referee #1

From Fig. 9 it looks as if for some benchmarks the differences between sites are bigger than the differences between model versions - is this caused by climate, stand age, stand density?

The large variation across sites is indeed expected to come from climate, stand density but also nitrogen and water availability. This variation is one of the main reasons why for large-scale models multi-site parameter optimization is preferred rather above tuning the parameters against a single site. ITRDB could provide multi-site observations for the same PFT but from

locations with rather different climates. Given that only a single model version will be shown in the revised manuscript this comment will not be addressed in the revised manuscript.

Is the assumption that forests are unmanaged likely to be correct?

We are trying to select site that are 150 years or older. Given the age and the location of the sampled forest, we think that such sites are unmanaged or experienced little management. Even in heavy managed regions as Europe such unmanaged forest fragments often less than a 1ha are abundant especially in the mountain regions. Although the revised manuscript will focus on BACI data for which the management status of the sites is better documented than for the ITRDB we will add this assumption in revised manuscript as the purpose of the manuscript is to propose benchmarks that could be used with the ITRDB data.

How was the start year set to match observations? Is this based on inferred tree age from the tree ring data or is there more information on forest age?

ITRDB doesn't have metadata for forest age so start and end year was matched to the length of the longest observation. The same applies to the older stands in the BACI data set. We will add this assumption in the revised manuscript.

Was there data on N deposition also used as forcing?

ORCHIDEE r5698 does use a global N deposition map based on the ongoing CMIP6 efforts. We will add a short description of these data in the revised manuscript.

Are the results of the leave-one-out approach shown somewhere?

The error-bar in Fig. 9 is the outcome of the leave-one-out approach. This was not reported in the manuscript. We will add it in the revised manuscript.

I don't see why the dynamic leaf N in itself would cause a problem, as it is a realistic process.

We agree that dynamic leaf N itself is not a problem. Rather, because leaf N is dynamic in ORCHIDEE it could be overestimated especially if the optimal value, which is prescribed for each PFT. Future model use, evaluation and optimization may help us to establish whether this is indeed a very sensitive model parameter. We will rephrase this paragraph of the manuscript to avoid similar confusion from reading the revised manuscript.

Referee #2

I am curious about whether the simulated ring width has been tuned before the final model run by adjusting some of the parameters. Could the authors be clear about whether there is the tuning process? And if so, the way of using RMSE or difference between observation and simulation can be tricky. Because those "artificial" bias could potentially have a big influence on such RMES-based benchmarks by simply changing/tuning the level of growth.

During model development data from ten ITRDB sites (aust112, cana106, chin037, finl055, fran4, id007, japa011, mo009, nepa003, spai055, and turk027) were used to assess the impact of the developments and to search for the most sensitive parameters. No formal model tuning took place because that is the objective of a follow-up study. We will clarify this issue in the revised manuscript by adding this information. The purpose of this manuscript is to show which benchmarks could be used such that ITRDB data could be used as one of the data streams that

is routinely used in model parameterization. This paper does not focus on how well/poorly ORCHIDEE can simulate tree rings. We see this as a two-stage process: (1) can the data be used and how, and (2) can ORCHIDEE be used?

Figure 4: more details about what is compared are needed. Is y-axis the mean of ring width?

Fig 4 compares the benchmarks. In the revised manuscripts all figures showing results will change given the initial comment of referee 1. While preparing the revisions we will pay special attention to improve the caption of the substitute of Fig 4.

Figure 6: Details to explain how the "recent year" at Panel (d) was decided is needed? And would this "cut" of data scarify the length of data availability, considering this new methodology is targeting for "century-long" model-data comparison, and the mature tree is one of the more important benchmarks in the four?

The 50-year cut off is somehow arbitrary but most of series considered during this study show a slowly decreasing growth-trend after 50 years. Also, the 30-year cut off used in Figure 7, is arbitrary. Here 30 years instead of 50 years were chosen because for most time series considered the first 30 years are characterized by fast changes in tree growth. The benchmarks target time series of 150 years and longer because those trees experienced considerable environmental changes. If the first 50 years are cut, the time series still contains 100 years during which CO2, temperature, precipitation and nitrogen availability may have changed. We will add this reasoning in the revised manuscript.

Figure 7: It was mentioned because the old fast-growing trees died well before sampling took place. But actually, those "young" fast-growing trees lived through a much longer period shown in Panel (a).

For consistency reasons, the same site was used to illustrate each of the four benchmarks (Fig. 5-8). The referee is correct that this approach may confuse readers. The site that was selected represents a relatively even-aged forest. In the revised manuscript we will chose a good example to illustrate each benchmark. This implies that consistency will be traded for readability.

Would the size-related growth have any impact on the quantile statistic?

We do indeed expect that the most dynamic phases of size-related growth could have an impact on the quantile statistics. For that reason, we select sites that already passed the dynamic phase in the year we start the analysis. This means that for this specific benchmark, trees should be 50 years or older in 1950 (the year after which we expect the climate reconstructions to become more reliable). By doing so we expect that there is almost no age-related growth trend in the years that are considered in the benchmark (1950-present). We will try to better clarify this line of reasoning in the summary table proposed to address the previous comment (improved version of the current table 1).

Panel (f): Is there any explanation about why the model is always overestimating the growth for both the good and bad years. Is it because the original value of TRW (not the standardized one) was used.

We agree with the hypothesis of the referee. The target is the absolute value of tree-ring width, and this shows the model overestimates overall tree-ring widths for this site. We plan to add

directions on how each benchmark can be used for evaluating models. Because Fig. 8 is drawn to explain the extreme benchmark, we omitted explanations about model performance.

I understand Panel (e) and (f) is to test the ability to reproduce the amplitude of TRW, which has also been majorly targeted by the former three benchmarks. However, it might also logically make sense by simply using the normalized value if the above three benchmarks passed. Meanwhile, relative change can be more relative to climate sensitivity comparison, if the simulated growth was tuned.

In this study we use amplitude as the difference between the lowest and highest observed diameter increment. According to this definition the fourth proposed benchmark is the only benchmark that is considering the amplitude, all other benchmarks are considering the trend in diameter growth. In the early years this trend is caused by ontogeny in later years the climate sensitivity could become more pronounced in the tree ring records. We will try to better clarify each benchmark in an improved version of the current table 1.