We'd like to thank to referees for their comments as they helped us to greatly simplify the manuscript and improve its key messages.
This document contains a point-by-point response to the referees.

Referee #1

Major comments

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.
The description of benchmarks was moved to the background section (Section 2.3) so it comes earlier in the manuscript. Fig 4 was indeed overly complex and has been replaced by a different analysis presented in table 2. This change in focus also helped to better demonstrate which of the benchmarks can be used with ITRDB data and which benchmarks require data from an unbiased sampling design.

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 understand where this confusion comes from and largely simplified the study by only using the BACI data. Rather than using ITRDB, the BACI data were resampled to mimicked ITRDB-type of sampling bias. The data from the unbiased and biased sampling designs were then used to benchmark. Differences in conclusions based on the data from the biased and unbiased sampling designs suggest that the benchmark is sensitive to the sampling design. This comment was instrumental in the revisions as it helped to simplify the manuscript as a whole and at the same time better demonstrates which of the proposed benchmarks can be used with ITRDB data. Within all BACI sites, sites dominated by a single coniferous species (spruce) and a single deciduous species (beech) were analyzed (Section 3.2) in response to the referees' question.

The description of ORCHIDEE contains a lot of general details that I'm not sure are needed in a journal such as GMD e.g. capacity to be coupled/decoupled, variable grid size etc. On the other hand, I find the details of the model setup somewhat sparse, and these details are needed to understand the model results. A simple solution would be to just take the detailed description of the setup from the supplementary material and add it to the main text.
Details about the model setup were necessary for the previous manuscript to understand the 4 different configurations better. But, considering the referee #1's comment, the ITRDB test

cases were removed, and the new test uses only one configuration (called 'Ndyn' previously). The rest of the setup is described in Section 3.3.

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. 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?
These tests were a left-over from the initial work on simulating tree ring widths with ORCHIDEE. The referee's comment made us realize that they were no longer needed and that we could bring the same message with a much simpler simulation set-up. We removed the different model configurations from the manuscript which enables us to better focus on examining the proposed benchmarks (Section 3.3 and 3.4).

The paper opens with a relatively long discussion about the issues and biases in using tree ring data to benchmark models. The problem is then that the actual discussion section largely repeats the same arguments. It would be more interesting to see here a discussion about the generality of the benchmarks - can they be used at different sites? With different species? For other models?
The whole result section and half of the discussion were rewritten in line with the major revisions we made in the study design. This rewrite includes more details on the validity and capability of the proposed benchmarks when using ITRDB data.

Minor comments

L 83 Is the assumption that forests are unmanaged likely to be correct?
We got the management status of each BACI site. This was applied to each simulation (Table S2).

L 84 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?
The start and end year were matched to the length of the longest observation. This has been clarified in the revised manuscript (Added to L377).

L 80 Was there data on N deposition also used as forcing?
This information is added to L343-344.
(Note line numbering appears to break after 100)
We adjusted the page margin.

Figure 4 - I'm not sure what 'trend' refers to.
We added the paragraph describing benchmarks further in the discussion (L491-523) and modified Table 1.

The model value for ‘young’ does not appear to have error bars.
Because ‘young’ was not affected much by the leave-one-out approach, but this figure is not used anymore.

Why does the data set have 27 sites and the model is run at 10 sites only? Are all these values for coniferous sites only?
This was because the previous figure 4 compared BACI datasets with ITRDB simulation. This issue has been resolved by simulating only at BACI sites. This resulted in full consistency between the numbers of observations and simulations.

L 130 These benchmarks are discussed earlier but only explained here
This part has been moved to the background section (Section 2.3).

L 185 Are the results of the leave-one-out approach shown somewhere?
It was missed in the previous version but, since the result presentation has been changed, this comment was not applied.

L209 (I think, p 17) I don’t see why the dynamic leaf N in itself would cause a problem, as it is a realistic process. It is much more likely that having a more complex representation of N processes exposes an issue with other parts of the N cycle in the model (e.g soil)
Since the main result and presentation was changed to focus on the verification of the benchmarks, the content related to evaluating the model was removed.

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?
This comment in combination with several questions from referee #2, made us realize that the explanation for the proposed benchmarks and what they are targeting was too concise in the previous manuscript. A large effort was made to clarify these issues by simplifying the study design as well as by adding further explanation about benchmarks, especially in relation to the application for the model evaluation (L491-523).

And the European regional case study didn’t give a clear conclusion for the whole benchmarks.
Now the manuscript doesn’t use ITRDB for simulations, but focuses on verifying the proposed benchmarks by using bias-free datasets. Unfortunately, the datasets are collected mostly in Europe, so the simulations are still limited to Europe. However, we believe it is a better way to evaluate the proposed benchmarks than simulating ITRDB sites over different continents.

There are a few minor queries, especially for the four benchmarks.
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.
The model was manually adjusted against ITRDB sites (thus not against BACI sites) during model development. This information has been added in the revised manuscript Section 3.3.

Figure 4: more details about what is compared are needed. Is y-axis the mean of ring width?
The y-axis was the ratio of the difference in benchmarks, however, Figure 4 has been removed and replaced by a different verification approach.

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
We did not fully understand this comment. It was unclear how randomizing the trend can improve data processing. The trend in tree-ring width or diameter contains the information the model is expected to reproduce and the flat slope in Fig. 5 (d) implies the model simulated the trend relatively well. We interpreted this comment as a sign that we did not well explain this in the initial manuscript and used the revision to provide more explanation on these issues (L493-499, L519-523, and Table 1).

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?
We added the underlying reasoning in the revised manuscript (L264-267)

Figure 7: Some logical reason why only the first few decades (30ish) years are chosen for this benchmarking is needed. The comparison was limited within the first few decades of the time series for young trees comparison. 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).
We changed the example site in Figure 5 and 6, finl052 to brit021.

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?

Extreme even benchmark was built to evaluate plant growth sensitivity to year-to-year climate which requires more reliable climate data than the other benchmarks. Two of the other benchmarks quantifies long-term growth. We agree with the referee #2's opinion about the time frame of extreme benchmark, however, we think the complementary role of four different benchmarks can relieve the issue. As described above, the revised discussion and table 1 were improved to better clarify these issues.

The extreme value was extracted from the average of the observation, without any size related detrending. Would the size-related growth have any impact on the quantile statistic?

Only trees that passed the most dynamic phases of the size-related growth trend were considered in this analysis. The referee is right in asking this question as we forgot to provide this information in the previous manuscript. The revised manuscript explicitly mentions this selection criterion on L279.

Panel (d): "mm" in y-axis title should be "Normalized".

This comment has been applied to Table 2.

Panel (f): how different years' value were matched if only the quantile was applied for both observation and simulation? 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.

This was because the model overestimate overall tree-ring widths, however, the detailed description about the model evaluation since the test was removed.

Again, I am wondering whether there is a modelling tuning process to adjust the simulated ring width closer to the observation. 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. We added more information in Table 1 to distinguish the long-term and short-term benchmark.

Table 1: Wider space between each row of the table could enhance the readability.

The cell margin was increased for better readability.

Using the International Tree-Ring Data Bank (ITRDB) records as century-long benchmarks for global land-surface models

Jina Jeong[1], Jonathan Barichivich[2,3], Philippe Peylin[2], Vanessa Haverd[4], Matthew J. McGrath[2], Nicolas Vuichard[2], Michael N. Evans[5], Flurin Babst[6,7,8] and Sebastiaan Luyssaert[1]

[1] Department of Ecological Sciences, VU University, 1081HV Amsterdam, the Netherlands.
[2] Laboratoire des Sciences du Climat et de l'Environnement, IPSL, CNRS/CEA/UVSQ, 91191 Gif sur Yvette, France.
[3] Instituto de Conservación Biodiversidad y Territorio, Universidad Austral de Chile, 5090000 Valdivia, Chile.
[4] CSIRO Oceans and Atmosphere, Canberra, 2601, Australia.
[5] Department of Geology & ESSIC, University of Maryland, MD 20742-4211, USA.
[6] Dendro Sciences Group, Swiss Federal Research Institute WSL, Zürcherstrasse 111, CH-8903 Birmensdorf
[7] School of Natural Resources and the Environment, University of Arizona, Tucson, USA
[8] Laboratory of Tree-Ring Research, University of Arizona, Tucson, USA

Running head: Tree-ring records as century-long benchmarks

Correspondence to: Jina Jeong (j.jeong@vu.nl)

Key words: forest growth, tree-ring width, diameter growth, climate sensitivity, size-dependent growth, climate change

Abstract

The search for a long-term benchmark for land-surface models (LSM) has brought tree-ring data to the attention of the land-surface community as they record growth well before human-induced environmental changes became important. The most comprehensive archive of publicly shared tree-ring data is the International Tree-ring Data Bank (ITRDB). Many records in the ITRDB have, however, have been collected ~~almost exclusively~~ with a view on maximizing an environmental target signal (e.g. climate), which has resulted in a biased representation of the productivity of forested sites and landscapes and thus limits its use as a data source for benchmarking. The aim of this study is to examine when and how ~~propose advances in land-surface modelling and data processing to enable the land-surface community to re-use the~~ ITRDB data can, despite its~~as a much-needed century-long benchmark. Given that tree-ring width is largely explained by phenology, tree size, and climate sensitivity, LSMs that intend to use it as a benchmark should at least simulate tree phenology, size-dependent growth, differently-sized trees within a stand, and responses to changes in temperature, precipitation and atmospheric CO_2 concentrations. Yet, even if LSMs were capable of accurately simulating tree-ring width,~~ sampling biases, ~~in the ITRDB need to~~ be used as century-long benchmarks~~accounted~~ for LSMs. Combining~~. This study proposes two solutions: exploiting the observation that the variation due to size-related growth by far exceeds the variation due to environmental changes; and simulating a size-structured population of trees. Combining the proposed~~ advances in modelling and data processing resulted in four complementary benchmarks - reflecting different usage of the information contained in the ITRDB - each described by two performance metrics rooted in statistics. Each of the four proposed benchmarks was verified by calculating it twice: (1) based on an independent European tree-ring network ~~that quantify the performance~~ of biomass plots that were sampled in a locally representative way and are thus not biased by big-tree tree selection, and (2) following sub-sampling of this European biomass network by only considering the 15% biggest trees. This study showed that the ITRDB data can be used with 95% confidence to benchmark annual radial growth during extreme climate years. In about 70% of the test cases, using ITRDB data would result in the same conclusions as using the European biomass network when the model is benchmarked against the annual radial growth of mature trees or the size-related trend in annual radial growth. Care should be taken when using the ITRDB data to benchmark the annual radial growth of young trees, as only 50% of the test cases were consistent with the results from the European biomass network.~~benchmark.~~ Although the proposed benchmarks are unlikely to be

exact~~precise~~, they may advance the field of land surface modelling by providing a much-needed large-scale constraint on changes in the simulated maximum tree diameter and annual growth increment for the transition from pre-industrial to present-day environmental conditions over the past century. Hence, the proposed benchmarks open up new ways of harnessing~~exploring~~ the ITRDB archive, but at the same time show the need for~~stimulate~~ the dendrochronological community to refine its sampling protocols to produce new and spatially unbiased tree-ring networks~~, and help the modelling community to move beyond the short-term benchmarking of LSM.~~

1. Introduction

Earth system models integrate numerical models of atmospheric circulation, ocean dynamics and biogeochemistry, sea ice dynamics, and biophysical and biogeochemical processes at the land-surface. Climate projections made by Earth system models have been the corner-stone of the last five Assessment Reports of the Intergovernmental Panel on Climate Change (IPCC, 2013) and as such have made a tremendous impact on global environmental policy (Paris Agreements, 2015)~~ (Hecht and Tirpak, 1995).~~ The credibility of projections of the future climate from any Earth system model hinges on the ability of each of its~~the~~ four submodels~~model components of an Earth system model~~ to accurately reproduce the past (McGuffie, 2005). Although long-term changes that date~~dates~~ back to pre-industrial conditions (Luo et al., 2012) have been documented for vegetation distribution through pollen based reconstructions (Cao et al., 2019), land-surface models (LSMs) currently lack a long-term benchmark for forest ecosystem functioning. The absence of long-term benchmarks is thought to contribute substantial uncertainty to simulated future global carbon stocks in soil and vegetation (Friedlingstein et al., 2006; Friedlingstein et al., 2014) and as such to climate projections (Fig. 1a).

Tree-ring records provide annual information on historical tree growth and physiology in relation to environmental conditions, including the era before human activities started to affect the atmospheric CO_2 concentration (Fritts, 2012; Hemming et al., 2001). Even though trees grown in the absence of a clear annual rhythm of vegetative and dormant seasons~~climatic stress~~

may not develop distinct tree -rings, as ~~has been~~ observed for many~~several~~ species from the humid tropics, ~~hence,~~ tree-ring records have been proposed as a large-scale and long-term benchmark for the land surface component of Earth system models (Fig. 1b) (Babst et al., 2014a, 2014b, 2017, 2018; Zuidema et al., 2018).

Until now, tree-ring records have often been collected ~~almost exclusively~~ to reconstruct past climate and hydrological variability from sites where trees grow near the colder or drier fringes of their distribution (Briffa et al., 2004; D'Arrigo et al., 2008). The most comprehensive archive of publicly shared tree-ring data is the International Tree-ring Data Bank (ITRDB), with more than 4,000 locations from 226 species across most forested biomes (Grissino-Mayer and Fritts, 1997; Zhao et al., 2019). However, a shortage of site metadata and the prevailing geographical, species and tree selection sampling biases resulting from targeting climate-sensitive trees has limited the use of the ITRDB archive to infer long-term changes in forest growth (Bowman et al., 2013; Briffa and Melvin, 2011; Klesse et al., 2018; Zhao et al., 2019). Compared~~These issues may, likewise, limit the information content of the ITRDB records compared~~ to tree- ring records that were collected for the purpose of benchmarking LSMs, such as the European tree-ring network of biomass plots (hereafter called "European biomass network"; Klesse et al., 2018) that is available through the database of the BACI project (BACI, 2020), the aforementioned issues may limit the information content of the ITRDB records.~~land-surface models (http://www.baci-h2020.eu/).~~ This loss in information content should, however, be balanced against the associated benefits ~~of re-using data~~ in terms of time gain and resource savings when re-using the large ITRDB dataset.

When tree rings are to be used as benchmarks for LSMs~~land surface models~~, the models will need the skill to mechanistically simulate tree-ring width (TRW). In the past decades, the major physiological and ecological processes that are responsible for annual tree-ring growth became sufficiently well-understood to be formalized in mathematical models with different levels of details. ~~and complexities.~~ The first TRW models (Wilson and Howard, 1968) described processes at the cell level: cell division, cell enlargement, and cell wall thickening. Later, the carbon and water balance of trees was added (Fritts et al., 1999) as well as climatic influences on cambial activity (Vaganov et al., 2006). These models were capable of reproducing short-term radial growth at the tree level. Further developments introduced a notion of turgor and hormone regulation for cell

growth (Drew et al., 2010; Hölttä et al., 2006; Leuzinger et al., 2013; De Schepper and Steppe, 2010; Steppe et al., 2006)~~(Drew et al., 2010; Hölttä et al., 2006; De Schepper and Steppe, 2010; Steppe et al., 2006)~~.

At the same time, the spatial scale of models simulating wood formation based on cell dynamics was extended to the stand level by simplifying process representation. In one such model, photosynthate availability, air temperature and soil water content were used to constrain wood cell growth and successfully reproduced observations (Deleuze and Houllier, 1998; Hayat et al., 2017; Wilkinson et al., 2015). Further simplifications were proposed by simulating the radial growth of trees based solely on carbon allocation (Deleuze et al., 2004; Merganičová et al., 2019) rather than cell dynamics, the latter being computationally too expensive for large scale vegetation models (Li et al., 2014; Misson et al., 2004; Sato et al., 2007). Hence, a variety of approaches is~~are~~ now available to describe TRW growth in forest models, dynamic vegetation models and LSMs~~land-surface models~~, but to the best of our knowledge there is yet no land-~~-~~surface component of any~~an~~ Earth system model with such capability.

This study articulates an improved conceptual framework for benchmarking simulated radial growth against ITRDB tree-ring data, addressing limitations in the models, the data and the methods to compare models and data. The aims are to: (1) use current understanding of tree-ring growth to derive the minimal requirements for benchmarking LSMs~~land-surface models~~ against tree-ring records archived in the ITRDB; (2) review potential issues of using the ITRDB to benchmark LSMs~~land-surface models~~; (3) propose solutions for a meaningful comparison of LSMs~~land-surface models~~ against ITRDB records; and (4) verify~~demonstrate~~ the proposed solutions~~methodological framework~~ by benchmarking a LSM~~land-surface model across ten European Scots pine (*Pinus sylvestris* L.) forests~~ using a dataset that is not prone to sampling biases related to palaeoclimatological research. ~~tree-ring width chronologies archived in the ITRDB.~~

2. Background: model requirements~~,~~ ~~and~~ data limitations and benchmarks

2.1. Minimal requirements for land-surface models to mechanistically simulate TRW

The linear aggregate conceptual model of tree growth ~~model~~ (Cook and Kairiukstis, 1990) considers that the observed TRW at year *t* (in mm) consists of five additive growth contributions (Fig. 2):

~~(i)~~ Size-dependent growth is the dominant signal in raw tree-ring measurements~~records~~ (Cook et al., 1995). Conceptually it can be understood by considering an almost constant volume of wood due to a more or less constant primary production (Hirata et al., 2007) being added to the trunks year after year (Nash, 2011). The annual diameter increment of the trees will decrease as the trunk grows wider because a given wood volume has to be distributed over an increasing surface area as both the circumference and height of the stem are increasing. In reality, however, self-thinning reduces stand density and competition for resources, implying that the remaining trees can increase their crown volume and thus increase their primary production (Oliver and Larson, 1996) which largely compensates for the size-dependent decrease in TRW and contributes to the observed almost constant TRW of tall trees. Several of the common allocation schemes used in LSMs~~land surface models~~ account for size-dependent growth and stand self-thinning (Franklin et al., 2012; Wolf et al., 2011).

(i)

(ii) Climate-dependent growth reflects the sensitivity of tree growth to radiation, temperature, and water availability (Fritts, 2012) and is accounted for~~well developed~~ in LSMs~~land surface models~~, as it represents the core purpose of this type of ~~model. Land surface~~ models. LSMs often rely on the Farquhar model for the radiation and temperature dependency of photosynthesis (Farquhar, 1989), the McCree - de Wit - Penning de Vries - Thornley approach for the temperature dependence of respiration (Amthor, 2000), and account for a decoupling of photosynthesis and growth by the use of a labile carbon pool (Friend et al., 2019; Naudts et al., 2015; Zaehle and Friend, 2010). Plant water availability is accounted for through either simple transfer functions or more recently by accounting for the hydraulic architecture of the simulated trees (Bonan et al., 2014; Naudts et al., 2015).

(iii) Endogenous disturbances refer to within-stand resource competition and are being increasingly simulated in LSMs~~land surface models~~ albeit often by empirical approaches (Haverd et al., 2013; Moorcroft et al., 2001; Naudts et al.,

2015). From a benchmarking point of view, simulating individuals of different size or cohorts within a single forest is essential to reproduce the sampling biases present in the ITRDB (see section 2.2 and 2.3 below). Chronic exogenous disturbances such as increasing atmospheric CO_2 concentration (LaMarche et al., 1984) and N-deposition (Magnani et al., 2007) are ~~also~~ well-developed as they are among the main purposes of using LSMs.~~land-surface models.~~ The effect of CO_2 fertilization on photosynthesis is accounted for in the photosynthetic submodel ~~(see above)~~ whereas nitrogen dynamics are accounted for through static or dynamic stoichiometric approaches (Vuichard et al., 2019; Zaehle and Friend, 2010) .

(iv) Although abrupt disturbances such as fires, pests and storms are increasingly being simulated by LSMs~~land-surface models~~ (Chen et al., 2018; Yue et al., 2014) these functionalities are at present of limited use for benchmarking against TRW data. Abrupt disturbances are often simulated as stand-replacing disturbances and will, therefore, not be reflected in the simulated TRWs. Furthermore, the timing of such events largely depends on the simulated diagnostics, for example, fuel wood build-up, insect population dynamics, and soil moisture, which could strongly deviate from the observed timing in decadal to century long simulation periods.

(v) The final term in the aggregate tree~~-~~growth model constitutes all processes and interactions between processes not previously accounted for in the LSM~~land-surface model~~, and will make up the model error.

This aggregate tree~~-~~growth model provides the conceptual basis for tree-ring standardization and climate signal extraction methods used in dendrochronology (Briffa and Melvin, 2011; Cook and Kairiukstis, 1990), which rely on the assumptions that the sampled trees capture the relevant common growth variability of the stand and that the contribution of each major driver can be statistically identified as either signal or noise. Note that alternative approaches have been proposed to attribute TRW to its major drivers (Stine, 2019). In practice~~growth variability of the stand and that the contribution of each major driver can be statistically identified as either signal or noise. If tree-rings are formed~~, observed TRW records cannot always~~, however,~~ be fully decomposed in the absence of metadata because several drivers might~~do~~ not leave a unique fingerprint in growth.~~the tree-ring record.~~ However, size effect and climate sensitivity have a much larger contribution to TRW than the

other processes (Hughes et al., 2011). ~~Nevertheless, alternative approaches have been proposed to attribute TRW to its major drivers (Stine, 2019).~~

In addition to accurate process representation, the model will need to be driven by historical climate, atmospheric CO_2 concentrations and N-deposition. In general, commonly-used century-long climate reanalyses such as NCEP (Kalnay et al., 1996), 20CR (Compo et al., 2011), and CERA-20C (Laloyaux et al., 2018) are based on the assimilation of instrumental observations in climate simulations and are thus independent from climate estimates derived from tree rings ~~tree-ring-based observations~~ or other proxy data. Nevertheless, the accuracy of the reanalyses decreases proportional to data availability, particularly in remote areas with a low density and temporal depth of meteorological stations. Given~~, and given~~ that local climate effects may have contributed to the TRW, it might be desirable to bias-correct~~align~~ the reanalysis with present day site-specific climate observations where they exist (Ols et al., 2018). When LSMs are forced by actual climate observations, reproducing the observed climate sensitivity in tree rings would both facilitate and add credibility to the land-surface simulation – if forcing, LSM and TRW models are all realistic and unbiased.

Given the above, LSMs~~land-surface models~~ that intend to use TRWs as a benchmark should at the minimum simulate: (1) dynamic plant phenology, (2) size-dependent growth, (3) differently-sized trees within a stand, and (4) responses to chronic exogenous environmental changes (Fig. 2). Whereas responses to chronic exogenous environmental changes are the reason LSMs –exist and are therefore to some extent accounted for by all current LSMs, size-dependent growth and size differentiation within a stand are at present only accounted for in few LSMs~~land-surface models~~, for example, CLM (ED) (Fisher et al., 2015), ORCHIDEE (Naudts et al., 2015), and LPJ-GUESS (Smith, 2001). The~~Revision 5698 of the~~ ORCHIDEE model (revision 5698) meets the aforementioned minimum requirements and therefore will be used in this study.

2.2. Challenges of using ITRDB data as a long-term benchmark

A typical record in the ITRDB consists of TRW measurements~~measurement~~ of increment cores from tens of individual trees from the same site and species. Each record may have ~~a~~ different starting and~~date,~~ ending dates,~~date~~ and thus length (Fig. 3 a and b). If a core reaches the centre of the trunk (i.e., pith),~~,~~ annual tree diameter can be reconstructed (Bakker, 2005). Even then diameter reconstruction may come with some uncertainty because trunks are not perfectly round. If the core does not contain the centre of the trunk, which is often the case for large trees, rings near the pith will be missed adding uncertainty to the diameter and age reconstruction (Briffa and Melvin, 2011). In this case, diameter increment can still be reconstructed (by subtracting the measured TRW) if trunk diameter at the time of sampling is known, but this metadata is rarely recorded in dendroclimatic collections and it is not stored in the ITRDB.

Despite of its known biases,~~ in~~ the ITRDB~~, it~~ can still be used to extract information useful for LSMs. The predominant sampling design in the ITRDB targets the presumably oldest trees, which should give the longest time series and are therefore ~~be~~ most useful to reconstruct the climate variability prior to instrumental records. The ITRDB is thus likely to overrepresent large trees (Brienen et al., 2012; Nehrbass-Ahles et al., 2014) relative to the population demographics at the time of sampling. This big-tree selection bias makes the ITRDB unsuitable to upscale growth of individual trees to larger spatial domains, i.e., stand, forest or the region (Babst et al., 2014a; Nehrbass-Ahles et al., 2014) but does not affect the value of the ITRDB archive for documenting individual tree growth as long as tree size and dominance effects are explicitly considered. Although the model-data comparison cannot -without additional data- correct for the big-tree selection bias in the ITRDB, models that simulate multiple tree diameter classes may accommodate this bias by comparing the largest simulated diameter class with the observed ITRDB tree-ring records (Fig. 3a).

Another bias related to the ITRDB sampling design comes from the fact that the growth rate of trees within a cohort differs between individuals (Melvin, 2004; Zuidema et al., 2011) resulting in slowly and fast-growing trees within the same cohort (Fig. 3b). Slow-growing trees tend to live longer than fast-growing trees in the same cohort (Mencuccini et al., 2005; Schulman, 1954). Records of TRW are thus likely to underestimate the mean tree growth of a stand in long-passed centuries as fast-growing trees would have died off before the samples were taken (Brienen et al., 2012). Another

~~The other~~ challenge of using ITRDB data is rooted in the difference between the observed and simulated forest structure. Tree-ring datasets are composed by cores of individuals from different cohorts (Fig. 3b). Comparing these data against simulations requires the model to be individual~~-~~-based or to align TRW records by age (Fig. 3a).

Given the above, only part of the information contained in TRW records can be used for benchmarking if their sampling protocol is poorly described or not rigorously enforced. A model-data comparison cannot correct for these biases but we propose to enhance the consistency between modelled and observed tree-rings for a stand under study ~~can be improved~~ by making use of virtual trees. Virtual trees will, however, require~~As the output of LSMs must be coupled to realistic models for TRW,~~ careful post-processing of the ITRDB data ~~may be required~~ to become suitable benchmarks for LSMs. Later in this study we propose four different ~~land surface models. The proposed~~ benchmarks based on the ITRDB data, ~~, thus, use~~ three of which make use ~~different definitions~~ of a virtual tree. Nevertheless, ~~trees~~ each of these benchmarks ~~which~~ addresses a different aspect or TRW and therefore uses a different definition for its virtual tree~~aspects~~:

(i) The~~(1) the~~ average virtual tree of a stand aligned by tree age is calculated as the time series for the average ring width after aligning the age of the individual trees (Fig. 3a). Age-aligned TRWs~~tree-ring widths~~ are widely used to calculate a statistic known as the mean regional curve of the sampled stand (Briffa and Melvin, 2011). This assumes that common drivers regardless of time, exceed the signal from local and individual differences in tree growth (see subsection ~~4.~~2.3 (i)).~~));~~

~~(i)~~ The~~ (2) the~~ average virtual tree of a stand aligned by calendar year is calculated by ordering individual tree-ring series~~ records~~ by calendar year (Fig. 3b) and for each year the average observed diameter is calculated. Alignment by calendar year thus reflects the real temporal evolution of the stand. This virtual tree can be used to cope with the challenge from difference in forest structure between the simulation and the observation by compiling a representative and comparable tree with the simulated tree (see subsection ~~4.~~2.3 (ii)).~~));~~

(ii) The~~(3) the~~ largest virtual tree of a stand is calculated after aligning individual trees by their age (Fig. 3c). The recommendation to remove the age trend from tree-ring records (Cook et al., 1995) confirms the assumption underling the alignment by age, i.e., that size dependent age exceeds the growth trends due to long-term environmental changes. Subsequently, the age-aligned TRWs can be used to compile a virtual fast-growing tree which has the maximum observed diameter of all trees for a given tree age. The virtual fast-growing tree thus gives a better idea of the true mean tree growth in old stands. (see subsection ~~4.~~2.3 (iii)).

The proposed data-model comparison thus largely relies on the concept of virtual trees to account for known sampling biases of the ITRDB as well as for the model definition of a forest stand. The validity of the concept of a virtual tree is evaluated in section 4.1. Except for LSMs~~land surface models~~ with an individual tree-based stand definition (Sato et al., 2007), benchmarking other models will have to consider the use of virtual trees as well. The~~Hence, the~~ proposed definitions and uses of virtual trees are partly customized~~specific~~ to ORCHIDEE r5698.

2.3. Benchmarks for comparing observed and simulated tree-ring widths

If a LSM explicitly accounts for the main factors contributing to TRW, i.e., size effects and climate sensitivity (Hughes et al., 2011), meaningful benchmarking against specific aspects of the observations becomes feasible in spite of the aforementioned biases in the ITRDB. Our technical framework considers four complementary aspects of the observations: (*i*) the size-related trend in tree-rings; (*ii*) diameter increment of mature trees; (*iii*) diameter increment of young trees; and (*iv*) extreme growth events. Each of these aspects formed the basis of a benchmark (Table 1):

(i) Size related diameter growth. The size-related trend in diameter increment can be assessed by calculating the average virtual tree for a stand aligned by tree age (Fig. 4 a, b) and subtracting its TRWs from the simulated TRWs of the largest diameter class (Fig. 4c). Subsequently, a linear regression is used to quantify the temporal trend in the residuals (Fig. 4d). If the simulations and observations have similar size-related trends, the temporal trend in the residuals will be close to zero. Furthermore, the root mean square error (RMSE) between the simulations and observations is calculated and normalized by

the length of time series used to calculate the difference in observed and simulated growth trends. A skilled model is expected to simultaneously show no trend in the residuals and a low RMSE across many sites.

(ii) Diameter *increment of mature trees*. In LSMs that account for within-stand competition, larger trees will consistently grow faster than smaller trees due to the way competition is formalized (Bellassen et al., 2010; Haverd et al., 2013). In reality, growing conditions can suddenly become favourable for trees that have previously been suppressed, resulting in fluctuating growth rates. This discrepancy between simulated and observed competition can be accounted for in the benchmark by using the observations to compile a virtual tree of the stand aligned by calendar year, taking the average tree diameter of all samples to construct the virtual tree (Fig. 5 a and b). Under the assumption that the observed trees are representative of the biggest trees from a given site, the virtual tree can be compared with the biggest diameter class from the model. Given that for the last decades both the quick and slowly growing trees are still alive and could have been sampled, only the growth in recent decades of the virtual tree are compared to the simulations (Fig. 5 c and d). The RMSE and trend of the residuals between the virtual tree and the largest diameter class simulated are calculated (Fig. 5d). A skilled model is expected to simultaneously show no trend in the residuals and a low RMSE across many sites.

(iii) Diameter *increment of young trees*. As mentioned above, the size-related trend in diameter increment can be assessed by calculating the largest virtual tree of the stand. The maximum age of a virtual tree equals the shortest observed individual TRW record for the stand, as it represents the age intersection between the TRW records for all individuals in the stand. The largest virtual tree is thus clearly biased towards higher observed diameters, compensating for the loss of observed high diameters in field sampling due to the fact that the old fast-growing trees died well before sampling took place (Fig. 6 a and b). The first three decades of growth of the virtual tree are then compared to the simulated growth of the largest diameter class (Fig. 6 c and d) by calculating the RMSE and trend of the residuals (Fig. 6d). The 30-year threshold is somewhat arbitrary but reflects the observation that most of the selected time series show fast changes in tree growth at the first 30 years. When benchmarking against other TRW data, this threshold could be adjusted to better fit the observed growth dynamics for other tree species and/or other regions. A skilled model is expected to simultaneously show no trend in the residuals and a low

RMSE across many sites. By using different approaches to evaluate the growth of young (this benchmark) and mature trees (the previous benchmark) the comparison accounts for the observation that the drivers of ring growth change as the trees grow taller (Cook, 1985).

(iv) Extreme growth events. Even a perfect LSM cannot be expected to reproduce all year-to-year variation due to uncertainties in forcing data, such as the reconstructed climate and N-deposition drivers. Nevertheless, well-constrained reanalysis-based climate reconstructions can be expected to contain extreme events, and hence a skilled model driven by well-constrained reconstructions should reproduce the statistics of the most extreme events. In this benchmark, extreme growth is defined as the first and last quartiles in TRW ordered by calendar year (i.e., not aligning establishment years). Since year-to-year variation of the simulation is more reliable after 1951 because the climate reconstructions used to drive the data rely more on observations rather than on a climate reanalysis as is the case for the years before 1950, this benchmark only uses TRW data that represent tree growth after 1951. As all selected stands were already 50 years or older in 1950 and TRW were thus past their juvenile dynamic growth phases, detrending was not required. Subsequently, individual tree records from the same site were averaged to obtain a single time series per site (Fig. 7a). The model skill to reproduce the absolute ring-width amplitude regardless of timing was tested by comparing the observed and simulated 25^{th} and 75^{th} percentiles of TRWs for the largest diameter class, which is the diameter class showing the strongest climate sensitivity (Fig. 7 e and f). Since other benchmarks test for model's capability to simulate absolute TRW, these benchmarks focus on the difference between high and low growth years. The mean TRW of the simulations or observations was subtracted to remove the effect of differences in, respectively simulated or observed TRWs. Additionally, model skill for reproducing the timing of individual extreme growth events was tested by comparing the simulated TRW for the exact years during which extreme growth was observed (Rammig et al., 2015; Fig.7 a-d). The amplitude and value of TRWs can affect the calculation, which is not the aim of the test, thus, TRWs were normalized by the standard deviation of the selected trees. For both the amplitude and timing of growth extremes, the similarity between simulations and observations was calculated as the RMSE with the error being the distance from the 1:1 line (Fig. 7 c-f). A skilled model is expected to simultaneously show low RMSE for both the amplitude and timing of extreme years across many sites.

3. Materials and Methods

3.1. The land-surface model ORCHIDEE

ORCHIDEE (Ducoudré et al., 1993; Krinner et al., 2005) is the land-surface model of the IPSL (Institute Pierre Simon Laplace) Earth system model (Dufrêne et al., 2005). Hence, by design, it can be coupled to an atmospheric general global circulation model or become a component in a fully coupled Earth system model. In a coupled setup, the atmospheric conditions affect the land-surface and the land-surface, in turn, affects the atmospheric conditions. However, when a study focuses on changes in the land-surface rather than on the interactions with climate, it can also be run as a stand-alone land-surface model. In both configurations the model receives as input atmospheric conditions such as precipitation, air temperature, air humidity, winds,~~and~~ incoming solar radiation, and CO_2; this combination of inputs is known as the climate forcing. Both configurations can cover any area ranging from global to regional domains and even down to a single grid point for the stand-alone case.

Although ORCHIDEE does not enforce a spatial or temporal resolution, the model does use a predefined spatial grid and equidistant time steps. The spatial resolution is an implicit user setting that is determined by the resolution of the climate forcing. Although the temporal resolution is not fixed, the processes were formalized at given time stepss~~tep~~: half-hourly (i.e. photosynthesis and energy budget), daily (i.e. net primary production), and annually (i.e. vegetation dynamics). Hence, meaningful simulations have a temporal resolution between 1 minute and 1 hour for the energy balance, water balance, and photosynthesis calculations.

ORCHIDEE builds on the concept of meta-classes to describe vegetation distribution. By default, it distinguishes 13 meta-classes (one for bare soil, eight for forests, two for grasslands, and two for croplands). Each meta-class can be subdivided into an unlimited number of plant functional types (PFTs). When simulations make use of species-specific parameters and age classes, several PFTs belonging to a single meta-class will be defined. Biogeochemical and biophysical variables are

calculated for each PFT or groups of PFTs (e.g. all tree PFTs in a pixel drawn from the same description of soil hydrology, known as a soil water column).

ORCHIDEE is not an individual-based model but instead it currently represents forest stand complexity and stand dynamics with diameter and age classes. Each class contains a number of individuals that represent the mean state of the class. Therefore, each diameter class contains a single modelled tree that is replicated multiple times and distributed at random throughout the PFT area. At the start of a simulation, each PFT contains a user-defined number of stem diameter classes. This number is held constant throughout the simulation, whereas the diameter boundaries of the classes are adjusted to accommodate for temporal evolution in the stand structure. By using flexible class boundaries with a fixed number of diameter classes, different forest structures can be simulated. An even-aged forest, for example, is simulated with a small diameter range between the smallest and largest classes. All classes will then effectively belong to the same stratum. An uneven-aged forest is simulated by applying a large range between the diameter classes. Different diameter classes will therefore effectively represent different strata. The limitations of this approach become apparent when the TRW~~tree-ring width~~ data and simulation are compared by calendar year as the model does not track individual trees. Although the dimensions of each model tree itself are~~is~~ well-defined, the amount of radiation it receives (and therefore the amount of carbon produced) is determined by the statistical distribution of all model trees in that grid cell.

Vegetation structure is then used for the calculation of the biophysical and biogeochemical processes of the model such as photosynthesis, plant hydraulic stress, and radiative transfer model. The r5698 version of ORCHIDEE, which is the version used in this study, combines the dynamic nitrogen cycle of ORCHIDEE r4999 (Vuichard et al., 2019; Zaehle and Friend, 2010) and the explicit canopy representation of ORCHIDEE r4262 (Chen et al., 2016; Naudts et al., 2015; Ryder et al., 2016). It is one of the branches of the ORCHIDEE model and it was further developed from Naudts et al. (2015) and Vuichard et al. (2019) (Text S1), parameterized, and tested to simulate TRW series~~tree ring widths~~, in order to meet the aforementioned minimum requirements of simulating the carbon, water, energy, and nitrogen cycle, while accounting for size-dependent allocation for three diameter classes within a forest stand.

In this study we use a ~~climate~~ data product for the climate forcing from a merged and homogenized gridded dataset developed for modelling ~~purpose~~purposes over the 20th century, i.e., CRU-NCEP (Viovy, 2016), the gridded nitrogen deposition product from CCMI (Eyring et al., 2013), and a gridded nitrogen fertilization product for N_2O (Lu and Tian, 2017) such that observed ~~such that observed tree-ring widths~~ TRWs for the past century can be used to evaluate the skill of the ~~land-surface model~~ LSM ORCHIDEE r5698 to simulate radial tree growth. A detailed overview of earlier developments (Krinner et al., 2005; Naudts et al., 2015; Vuichard et al., 2019) that resulted in the emerging capability of ORCHIDEE r5698 to match the aggregate tree growth model (Fig. 2) is given in the supplementary material (Text S1).

~~**3.2. Simulation set-up for the data-model comparison**~~

~~We selected 10 forest sites from the ITRDB for comparison with simulations, based on the following criteria: (1) located in Europe; (2) composed of *Pinus sylvestris* L.; (3) between 100 to 150 years old; and (4) ranging from Spain to Finland and therefore thought to largely cover the climatic conditions encountered across the species range of *P. sylvestris* within Europe. The location of the selected forests is detailed in Table S2. ORCHIDEE r5698 was run for 10 individual pixels, each containing one of the selected sites. An observed time series of atmospheric CO_2 concentrations was used (Keeling et al., 1996) and all forest were considered to be unmanaged. Every simulation started from a 300-year long spinup required to bring the simulation to equilibrium with respect to the slow carbon and nitrogen pools in the soil. The start year and the length of each simulation was set to match the site observations. A more detailed description of the test case and the ORCHIDEE model is given in the Supplementary Information (Text S2).~~

~~The model run was repeated four times for every site to obtain simulated tree ring widths for four different model configurations. The first configuration is the most basic configuration in this test (hence its label 'basic'): sapling recruitment is not accounted for, the nitrogen cycle is open and the parameter quantifying resource competition within a stand (f_power, for more details see Text S1 Eq. 15 and 16) was fixed at 2. The second configuration (labelled 'power') was a copy of the first but used a modified expression for resource competition (Eq. 16 in Supplementary Information) was used ('power'). The~~

~~third configuration built on the second but also accounted for recruitment ('recru'). Finally, the fourth configuration used a closed and dynamic nitrogen cycle, recruitment, and the modified within-stand competition f_power ('Ndyn').~~

~~The configuration with an open nitrogen cycle prescribed the leaf carbon-to-nitrogen-ratios with the average leaf carbon-to-nitrogen-ratio obtained from the 'Ndyn' simulation following the method proposed by Vuichard et al. (2019). This ensured that the differences came from the C-N feedbacks rather than from differences in leaf nitrogen. In the absence of bias corrected climate forcing, the simulations cycled through the climate forcing from 1901 to 1910 for the years prior to 1901. From 1901 onwards, climate forcing matching the simulation years was used. The four configurations with increasing model functionality are summarized in Table 2.~~

~~3.9.~~3.2. ~~Reference data of~~ productivity-oriented ~~ecologically sampled~~ TRW sampling~~data~~

The European biomass network contains TRW samples from ~~A~~ "fixed-plot sampling". The ~~approach" TRW~~ database was established within ~~under the BACI project (http://www.baci-h2020.eu/) by archiving 48 datasets from~~ multiple research projects and made publicly available through the EU Horizon-2020 project BACI (BACI, 2020). It archives at present 48 datasets from a variety of research efforts in Eurasia~~Europe~~ (Klesse et al., 2018). All ~~To be retained and archived, *all*~~ trees larger than 5.6 cm in diameter at breast height had to be sampled in a 10 to 40 m radius plot, depending ~~of which the exact radius depended~~ on stand density, to be archived in the BACI database (Babst et al., 2014). The European biomass network~~BACI archive~~ is, therefore, considered to be free from the big-tree selection bias that has plagued the ITRDB, although other known biases (e.g. slow-grower survivorship bias; (Bowman et al., 2013)) may still be present. ~~which is present in the ITRDB.~~ The records from the European biomass network are thus suited ~~BACI were used~~ to evaluate the validity of using virtual trees constructed from ITRDB records to cope with~~combat~~ the aforementioned sampling biases.

3.3. Simulation set-up

We selected sites from the European biomass network based on the following criteria: (1) the site had to be dominated by a single species for enhanced compatibility with ORCHIDEE, which is monospecific by design; and (2) stand age should

exceed 50 years as a requirement to apply ~~For~~ all four proposed benchmarks (Section 2.3). The benchmarks were applied to a common evergreen and a common deciduous species. Hence, within the filtered sites, only sites dominated by *Picea Abies* or *Fagus Sylvatica* were retained, resulting in 12 sites out of the total of 48 sites. CIM, a site dominated by *Fagus Sylvatica*, was removed from the selection (decreasing the final number of sites to 11) because only one tree out of 61 trees was aged over 100 years, resulting in a diameter distribution that is not at all compatible with the default diameter distribution of the model. The details of the selected sites are in Table S2.

For the simulations, the LSM ORCHIDEE r5698 was used. This model version accounts for the aforementioned minimum requirements for LSMs to mechanistically simulate TRW. ORCHIDEE r5698 was run for 11 individual pixels, each containing one of the selected sites. An observation-based~~reflecting a different usage of the information~~ time series of atmospheric CO_2 concentrations was used (Keeling et al., 1996) and forest management followed the reported management status of the site (Table S2). No formal model optimization took place but during model development, model parameters were manually adjusted to better match the TRW data of 10 ITRDB sites (aust112, cana106, chin037, finl055, fran4, id007, japa011, mo009, nepa003, spai055, and turk027). Every simulation started from a 300-year long spinup required to bring the simulation to equilibrium with respect to the slow carbon and nitrogen pools in the soil. The spinup was concluded with a clear cut such that the start year and the length of each simulation matched the observed stand age. The model configuration distinguished five diameter classes. The smallest diameter class contained 15% of the total number of trees, the intermediate diameter classes contained 21, 27, 21%, and the largest diameter class represented 15% of the total number of trees. A more detailed description of the ORCHIDEE model is given in the Supplementary Information (Text S2).

3.4. Verification of the benchmarks

The European biomass network data were used to verify, whether the big-tree selection bias that is present in the ITRDB data invalidates its~~in the ITRDB, the virtual trees were constructed by making~~ use for benchmarking LSMs. The verification used the data from the European biomass network in two different ways: 1) all trees in European biomass network data were used (hereafter called “all-tree data”) to calculate the four proposed benchmarks at the site level. The results of these benchmarks

were used as the reference in the verification, and 2) only big trees were sub-sampled from the data (hereafter called "big-tree data") and all four benchmarks were calculated against this sub-sample of data. Big trees were defined as ~~of only~~ the top 15% of the trees based on their diameter, and the~~. The~~ 15% threshold was taken to match the diameter distribution in ORCHIDEE, where by definition the largest diameter class contains 15% of the trees.

The verification required three additional steps (Fig. 8): 1) The metrics of each benchmark based on the big trees were optimized by simple arithmetic operations (see below) to obtain the best possible fit between the model and the observations. The best possible fit was quantified by the two metrics for each benchmark; 2) the same arithmetic optimizations were applied using the all-tree data for each of the four benchmarks and both metrics were calculated, and 3) the actual verification tested whether for a given metric and a given benchmark the arithmetic optimization 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-optimization value for that metric from its post-optimization 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 optimization improved for both datasets were counted ~~. Subsequently, the ratio of the diameter of the virtual trees over the diameter of the average stand diameter was calculated~~ to estimate the confidence in using ITRDB in benchmarking LSMs.

Arithmetic optimization differed between the different metrics used for the proposed benchmarks: 1) find a multiplier that minimises the metrics, i.e. RMSE or amplitude, when applying it to the simulated TRW, 2) find a modifier of the simulated growth trend that minimises the slope of the residuals, when subtracted from the observed TRW and 3) rearrange the years of the simulated outputs such that they match the order of observed extreme events~~error introduced by using a virtual tree. Only the 27 coniferous forests contained in the BACI archive were selected to enhance consistency within our test case~~.

4. Results

4.1. Verification of ~~Evaluating~~ the concept ~~of virtual trees~~

~~The bias introduced by constructing a virtual tree based on the diameters of the largest 15% of the trees from the representatively sampled BACI data sets is thought to be a proxy of the bias inherent to the ITRDB (Fig. 4). No statistically significant differences were found between the simulation-based and data-based virtual trees except for the virtual tree used in the benchmark for young trees (t-test, $p<0.01$). In the latter case the statistical difference seems to be caused the lack of variation in the simulation-based tree rather than a large difference between the means of the virtual trees (Fig. 4). This lack of variation for the simulation-based tree is expected as the method uses the maximum simulated diameter. Hence, the test supports the use of virtual trees constructed from the ITRDB as a benchmark for the largest diameter class of a forest simulated by ORCHIDEE.~~

~~4.2. Benchmarks for comparing observed and simulated tree-ring widths~~

~~If a land surface model explicitly accounts for the main factors contributing to TRW, i.e., size effects and climate sensitivity (Hughes et al., 2011), meaningful~~**behind** benchmarking ~~against specific aspects of the observations becomes feasible in spite of the aforementioned biases in the ITRDB. Our technical framework considers four complementary aspects of the observations: (*i*) the size-related trend in tree-rings; (*ii*) diameter increment of mature trees; (*iii*) diameter increment of young trees; and (*iv*) extreme growth events. Each of these aspects formed the basis of a benchmark (Table 1):~~

~~*(i)* The size-related trend in diameter increment can be assessed by calculating the average virtual tree for a stand aligned by tree age (Fig. 5 a, b) and subtracting its TRWs from the simulated TRWs of the largest diameter class (Fig. 5c). Subsequently, a linear regression is used to quantify the temporal trend in the residuals (Fig. 5d). If the simulations and observations have similar size-related trends, the temporal trend in the residuals will be close to zero. Furthermore, the root mean square error (RMSE) between the simulations and observations is calculated and normalized by the length of time series used to calculate the difference in observed and simulated growth trends. A skilled model is expected to simultaneously show no trend in the residuals and a low RMSE across many sites.~~

(ii)(i) Diameter increments of mature trees. In land-surface models that account for within-stand competition, larger trees will consistently grow faster than smaller trees due to the way competition is formalized (Bellassen et al., 2010; Haverd et al., 2013). In reality, growing conditions can suddenly become favourable for trees that have previously been suppressed, resulting in fluctuating growth rates. This discrepancy between simulated and observed competition can be accounted for in the benchmark by using the observations to compile a virtual tree of the stand aligned by calendar year, taking the average tree diameter of all samples to construct the virtual tree (Fig. 6 a and b). Under the assumption that the observed trees are the bigger trees from a given site, the virtual tree can be compared with the biggest diameter class from the model (see section 4.1). Given that for the last decades both the quick and slowly growing trees are still alive and could have been sampled, only the growth in recent decades of the virtual tree are compared to the simulations (Fig. 6 c and d). The RMSE and trend of the residuals between the virtual tree and the largest diameter class simulated are calculated (Fig. 6d). A skilled model is expected to simultaneously show no trend in the residuals and a low RMSE across many sites.

(iii)(i) Diameter increments of young trees. As mentioned above, the size-related trend in diameter increment can be assessed by calculating the largest virtual tree of the stand. The maximum age of a virtual tree equals the shortest observed individual TRW record for the stand, as it represents the age intersection between the TRW records for all individuals in the stand. The largest virtual tree is thus clearly biased towards higher observed diameters, compensating for the loss of observed high diameters in the sampling approach due to the fact that the old fast growing trees died well before sampling took place (Fig. 7 a and b). The first decades of growth of the virtual tree are then compared to the simulated growth of the largest diameter class (Fig. 7 c and d) by calculating the RMSE and trend of the residuals (Fig. 7d). A skilled model is expected to simultaneously show no trend in the residuals and a low RMSE across many sites. By using different approaches to evaluate the growth of young (this benchmark) and mature trees (the previous benchmark) the comparison accounts for the observation that the drivers of ring growth change when the trees grow taller (Cook, 1985).

(iv)(i) Extreme growth events. Even a perfect land-surface model cannot be expected to reproduce all year-to-year variation due to uncertainties in forcing data, such as the reconstructed climate and N-deposition drivers. Nevertheless, well-constrained

~~climate reconstructions can be expected to contain extreme events, and hence a skilled model driven by well constrained reconstructions should reproduce the statistics of the most extreme events. In this benchmark, extreme growth is defined as the first and last quartiles in TRW ordered by calendar year (i.e., not aligning establishment years) and averaged over the individual trees records (Fig. 8a). The model skill to reproduce the absolute ring-width amplitude regardless of timing was tested by comparing the observed and simulated 25th and 75th percentiles of TRWs for the largest diameter class, which is the diameter class showing the strongest climate sensitivity (Fig. 8 e and f). Additionally, model skill for reproducing the timing of individual extreme growth events was tested by comparing the simulated TRW for the exact years during which extreme growth was observed (Rammig et al., 2015; Fig. 8 a–d). For both the amplitude and timing of growth extremes, the similarity between simulations and observations was calculated as the RMSE with the error being the distance from the 1:1 line (Fig. 8 e–f). A skilled model is expected to simultaneously show low RMSE for both the amplitude and timing of extreme years across many sites.~~

ITRDB

~~4.3. Test case~~

The verification was applied at 11 sites. Given that each benchmark consists of two metrics, each of the four proposed benchmark generated 22 test cases. Across the four benchmarks 88 test cases were thus available. Despite its simplicity, the arithmetic optimization was found to be robust as it improved all metrics of the four proposed benchmarks at each of the eleven sites when benchmarking against the sub-sampled big-tree data. Applying the same optimizers to the all-tree data improved the match between the simulations and observations in 72~~80~~% of the test cases (634 out of the 88 test cases; Table 2). This overall number hides large differences between tree species and individual benchmarks. The verification appeared to be more successful for beech with an overall confidence level of 87% (28 out of 32 test cases) compared to spruce with a 64% (36 out of 56 test cases) confidence level. The performance differences between the individual benchmarks are detailed in the remainder of this section.

When benchmarking the size trend, big-tree data can be used with 72% (16 out of 22) confidence for benchmarking LSMs. Given the reasoning underlying the verification, this suggests that for 72% of the cases the conclusions would be similar

irrespective of whether ORCHIDEE is benchmarked against the big-tree data rather than the all-tree data. Some sites such as DEO and DVN showed marginal positive difference suggesting that simulations with ORCHIDEE r5698 using default parameters matched the observed size-related growth trend reasonably well, leaving limited room for improvements. One site, SCH showed a positive difference because it contained two slowly growing trees which lived roughly 40 years longer than the rest of trees but whose diameter was too small to be contained in the big-tree sample (Fig. S1). Except for this site, the size-related trend in tree growth can be derived from either the big-tree or the all-tree data.

For the mature trees benchmark, big-tree data can be used with 68% (15 out of the 22) confidence for benchmarking against LSMs. Two of the sites (HD2 and TIC) for which the all-tree data results in different benchmarking results from the big-tree data, have 36% to 44% of small trees in their size distribution, compared to an average 28% at the other nine sites. The proportion of small trees in the observation was estimated by counting trees in the smallest bin when trees are divided into 5 size classes similar to the model. On the other hand, ZOF had a bimodal size distribution, which has the biggest number of trees in the 1st and 4th bins (35% and 32% respectively). The default size distribution in ORCHIDEE has 15% of its trees in the smallest-sized class, and 21% in the 4th sized-class. At sites DEO and SOB, the average simulation matched well with the average diameter trend as shown by the calculated slope of residuals: 0.08 and 0.09 (Fig. S2 a). However, the growth rate for big trees was higher in the observations (0.95 and 0.50 for the slope of residuals) since the difference in big trees and small trees are bigger in the observations (Fig. S2 b). These results suggest that the mature trees benchmark is sensitive to the stand structure.

With 50% (11 out of 22) confidence in using the big-tree data in benchmarking LSMs, this benchmark appears to be the most demanding in terms of its data. At sites DEO, HD2, and SOB, inconsistencies between benchmarking the big-tree data and the all-tree data stemmed from: (1) the average simulations and observations being similar, with RMSE around 10 mm and; (2) the difference between big trees and small trees growths being larger in the observations (Fig. S3). The site labelled as SCH contained two extremely fast-growing young trees resulting in a very fast-growing virtual tree in the optimized model output (Fig. S4). For SOR the difficulties may have come from the model itself more specifically difficulties with the carbon

allocation (Fig. S5). These results suggest that a variety of issues decreases the confidence in using big-tree data for benchmarking.

For the extreme growth benchmark, big-tree data can be used with 95% (21 out of the 22) confidence for benchmarking LSMs. The observed consistency between benchmarking the big-tree data and the all-tree data suggests that extreme growth happens in the same years, irrespective of which dataset is being used. The site (DVN) showed the smallest RMSE for amplitude when it is calculated with the all-tree data (0.02) but the site has the biggest ratio of big-trees to all-trees for amplitudes compared to simulation (1.30, Fig. S6). In other words, if the simulation is adjusted to the big trees in the observation, since the difference between sub-sampled big-tree and all-tree is larger in the observations, the average simulation becomes bigger than the average observation as shown in Fig. S2 and S3. This result suggests that the extreme growth benchmark is the least demanding benchmark in terms of the sampling-design.

~~As the test case aimed at illustrating the four benchmarks rather than evaluating the ORCHIDEE model itself, the main objective of the test case is confirming that each benchmark indeed tests different aspects of the model's performance and that the proposed benchmarks can, therefore, be considered complementary to each other. The four benchmarks presented above were applied to a test case consisting of ten *Pinus sylvestris* L. sites across Europe. For each site the dataset consisted of 11 to 304 cores (Table S2); a leave-one-out approach was used to evaluate the sensitivity of the benchmark to individual tree records. For the simulations, the land surface model ORCHIDEE r5698 was used. This model versions accounts for the aforementioned minimum requirements for land surface models to mechanistically simulate TRW. The benchmarks were used on four configurations with increasing model functionality (Table 2). A more detailed description of the test case, the ORCHIDEE model and the different configurations is given in the Supplementary Information (Text S1 and S2).~~

~~The model reproduced the observed age-trends across the four configurations tested, i.e., the 'basic' set-up, the 'recru' set-up with an increasing number of individuals through recruitment (Eqs. 28 and 29 in Text S1), the 'power' set-up with a decreasing share of stand-level photosynthesis allocated to the largest size classes (Eq. 15 in Text S1), and the 'Ndyn' set-up where growth may be limited by additional nitrogen requirements (Eq. 32 in Text S1). Increasing the functionality of the model did~~

not help to reduce the RMSE of the overall age trend but did improve the match (i.e. the slope) between the simulated and observed age trends at 8 of 10 sites (Fig. 9a). For mature trees, the additional functionality acted as an offset, and hence growth rate estimates of the largest trees containing the biggest share of stand biomass improved at sites where diameter growth was underestimated (indicated by the trend in the residuals). Diameter growth rates were, however, further overestimated at sites where this was already the case (Fig. 9b). The RMSE of the diameter of mature trees (Fig. 9b, x-axis) were negatively correlated with their growth rate (Fig. 9b, y-axis) ($\rho = -0.5$, $p < 0.01$), suggesting that little new insight is to be expected from using both metrics from this benchmarks (Table 1) compared to simply selecting one.

Increasing model functionality had a much smaller effect on early stand developmental stages, as shown by the much smaller absolute changes in both RMSE and trends in the residuals for the young trees (Fig. 9c) compared to the old trees (Fig. 9b). Interestingly, considering feedbacks between the nitrogen and carbon cycles (configuration 'Ndyn') increased the RSME for forests located in regions with relatively low nitrogen deposition loads (Fig 9c, x-axis). When the dynamic nitrogen cycle is used, the carbon-to-nitrogen ratio in the leaves increases with increasing age whereas it remains constant for the open nitrogen cycle used in the other three configurations. This simulated age related dynamic for leaf nitrogen in combination with its initial value contributes to the overestimation of the TRW for young trees. Finally, adding functionality to the model did not help to enhance the model skill in terms of simulating extreme growths which is not a surprise because the added functionalities are not expected to improve the climate sensitivities of ORCHIDEE (Fig. 9d).

4. ~~5.~~ Discussion

The wealth of approaches available for modelling tree-ring growth ~~(see the introduction for a summary)~~ has been largely overlooked by the global land-surface community. Until ~~and until~~ now, benchmarking LSMs~~land-surface models~~ against ring-width records still relies mostly on interannual variation in the simulated net primary productivity as a proxy for TRW (Klesse et al., 2018; Kolus et al., 2019; Rammig et al., 2015; Zhang et al., 2018). Although such an indirect approach is valid to certain extent to benchmark the capability of LSMs~~land-surface models~~ to simulate interannual variability, the observations will need to be detrended to remove the size-related growth signal, adding considerable uncertainty to the benchmark (Bunde

et al., 2013; Cedro, 2016; Nicklen et al., 2019; Stine, 2019). Moving beyond the net primary production proxy by explicitly simulating stem radial growth and TRW enriches the benchmark since potentially confounding factors including climate responses, forest structure, age and size trends (Alexander et al., 2018; Nickless et al., 2011), as well as sampling biases (Babst et al., 2014a), can be better accounted for. Whereas previous studies had to rely on a single qualitative benchmark, i.e., interannual variability, our method shows that at least four benchmarks, each of them defined by two metrics, are available for models that meet the minimal requirements to simulate TRW determined by Cook's conceptual~~aggregate tree growth~~ model of aggregate tree growth (Fig. 2). Given that tree-ring width is largely explained by phenology, tree size, and climate sensitivity, LSMs that intend to use it as a benchmark should at least simulate tree phenology, size-dependent growth, differently-sized trees within a stand, and responses to changes in temperature, precipitation and atmospheric CO_2 concentrations.

Irrespective of the model approach, the largest archive of tree-ring records that is freely available to the land-surface community, i.e., the ITRDB, is prone to~~notorious for its~~ sample biases (Klesse et al., 2018; Zhao et al., 2019). Although it may be difficult to correct the data for these biases, we propose two solutions for comparing LSM output to biased observations. Simulating a size-structured population of trees enables comparing the observations relative to a benchmark for a tall simulated tree, which compensates for the tendency of dendroclimatic sampling~~samplers~~ to select the oldest trees in a stand (which often turn out to be the larger trees). Although the ITRDB does not contain the site metadata that would be required to make this comparison exact, i.e., the diameter and true age distribution of the sampled stand, it protects against comparing extreme samples to mean simulations. The second solution relies on the observation that the variation due to size-related growth by far exceeds the variation due to environmental changes and helps to constrain the survivor bias which is derived from the growth of young fast-growing trees that died a long time ago and are therefore absent from records made from present day sampling of old growth forests (Brienen et al., 2017). The benchmarks proposed here provide a tool to start using TRWs as a much-needed large-scale constraint on the maximum tree diameter and annual growth for the transition from pre-industrial to present-day environmental conditions.

Combining these two solutions with targeted processes resulted in four benchmarks, each of them defined by two complementary metrics (Table 1):

(i)). The size-trend benchmark targets the long-term trend in TRW. This trend contains information about ontogenetic growth during establishment and endogenous competition from canopy closure (Cook and Kairiukstis, 1990). Although this trend is removed in many dendrochronological studies to amplify the climate signal contained in TRW (Briffa and Melvin, 2011), we suggest to test the skill of the model in reproducing it because it is important to constrain biomass production. Benchmarking a suitable LSM against observed size-related trends in TRW may help to develop, evaluate or parameterize allometric relationships and changes in stand density.

(ii) The mature trees benchmark tests the capability of the model in simulating annual growth of mature forest. Since thisease demonstrated that these benchmarks aligns the observations by calendar year, it may reflect the effects of long-term environmental changes if there were any and the record is long enough as to experience them (Hess et al., 2018; Panthi et al., 2020). As a skilled LSM is expected to reproduce plant responses to the long-term environmental changes, this benchmark could be used to develop, evaluate and parameterize the processes that simulate endogenous disturbances and plant responses to, e.g., increasing atmospheric CO_2 concentrations, atmospheric N deposition and warming.

(iii) Tree growth during stand establishment can be tested with the young trees benchmark. The growth of establishing trees differs from that of mature canopy trees, and this difference has been accounted for by using separate benchmarks for young and mature stands. This benchmark could be used to develop, evaluate or parameterize allometric growth of young trees as well as tree mortality prior to canopy closure.

(iv) The extreme growth benchmark tests the occurrence and range of extreme growth events. Previously, interannual variability in TRW has been used to evaluate the climate sensitivity of LSMs. Inter-annual variability has a limitation because we cannot expect the model to simulate the timing of either endogenous nor every exogenous disturbance such as fire, pest and disease outbreak or dead of big trees leading to sudden growth releases in adjacent trees. Following this reasoning, it was

decided to benchmark only 25% and 75% extreme growth. This benchmark could be used to develop, evaluate or parameterize the plant water stress and the temperature dependency of plant growth in the model.

The metrics of the first three benchmarks are RSME and slope. RMSE examines if the model reproduces the absolute values of TRWs. However, even though a model might reproduce well the value of TRWs, it is still expected to simulate the long-term trend in TRW that comes from climate changes or endogenous competition. This latter aspect is quantified by the slope-metric. For the large-scale models such as a LSM and for sites with little high-quality site information, correctly simulating growth trends should be prioritized over matching the endpoints in tree diameter.

The results show that the four proposed benchmarks are ~~largely~~ independent in terms of their information content and that combining them would result ~~thus resulted~~ in a rich and rather refined description of some of the remaining model deficiencies. The conclusion is supported by the verification (Table 2) where, except site GIU, each site showed different weaknesses and strengths. ~~(Fig. 9).~~ The novel benchmarks proposed here thus provide new targets for evaluating LSMs' ~~land-surface models'~~ performance as each of the eight metrics could be used in the objective function of any data assimilation technique (Peylin et al., 2016) to rigorously account for the information contained in TRW records .

If the ITRDB is to be used, the number of benchmarks that can be used with confidence becomes more limited. The verification results (Table 2) show that if ORCHIDEE is benchmarked against data with a big-tree bias as the ITRDB, there is 70% confidence that the benchmark will come to the same conclusions as whether a data free from the big-tree selection bias would have been used. Higher chances for beech (87%) compared to spruce (64%) suggest that the validity of the assumptions underlying the use of ITRDB data, partly depend on tree species. In this study, the variety in species was too limited to generalize this results in terms of plant functional types. Across species, benchmarking against extreme events, mature trees, and the size-related trend appeared to be the least demanding in terms of the biases present in the data used for benchmarking. Benchmarking against young trees will benefit most from using data free from the big-tree selection bias.

The validity of benchmarking TRW of young trees against ITRDB data is questionable for sites were the default diameter distribution of ORCHIDEE poorly describes the observed diameter distribution (Fig S2 to S4). This finding limits the use of the ITRDB data for the young tree benchmarks. Because the true diameter distribution is not contained in this database, it is neither possible to select only ITRDB sites for which the actual diameter distribution matches the ORCHIDEE's distribution nor to adjust the diameter distribution in ORCHIDEE to the observed distribution. Matching observed and simulated distributions appears to be essential when benchmarking the growth of young trees. The same finding suggests, however, that forest inventory data for which the diameter distribution is known but only a few big trees were cored would be a reliable data source for benchmarking LSM.

Benchmarking, even against ITRDB data is useful, irrespective of the proposed benchmark when the simulated TRWs differ substantially from the observed TRWs. For example, at the GIU site, where the simulated TRW is much smaller than the observed (Fig. S7 a), the simulation could be improved for all metrics of all four benchmarks despite the difference in simulated and observed stand structure (Fig. S7 b). However, as shown above, inconsistencies between benchmarking the big-tree data and the all-tree data start to appear when the simulated TRWs are better approaching the observations (Fig. S 2-4, and 6). This implies that: 1) ITRDB can be used as a first approximation to benchmark the growth of young and mature trees in LSMs, 2) as the model improves, the need for unbiased datasets will increase as biases in stand structure and growth rates could hamper the use of especially these benchmarks.

Tree-ring records thus should complement well-established but short-term benchmarks for LSMs~~land-surface models~~ (Randerson et al., 2009), such as forest inventory data (Bellassen et al., 2010; Naudts et al., 2015), FLUXNET sites (Blyth et al., 2010; Williams et al., 2009), Free Air CO_2 Enrichment experiments (De Kauwe et al., 2013) and satellite observations of vegetation activity (Chen et al., 2011; Demarty et al., 2007). The value of tree-ring records can be further enriched by: (i) developing new and unbiased networks to complement the ITRDB, such as the European biomass network, (ii) adding their stable isotope ratios as benchmarks (Levesque et al., 2019; Barichivich et al. in prep)~~,~~) and (iii) combining their use with high-frequency but short-term eddy covariance measurements (Pappas et al., 2020; Teets et al., 2018)~~(Curtis et al., 2002)~~,

experimental data from plant growth under pre-industrial CO_2 concentrations (Temme et al., 2015), and proxies of atmospheric composition (Campbell et al., 2017).

7. Code and data availability

In line with GMD requirements, the model code has been archived and made accessible: https://doi.org/10.14768/20200228001.1. The scripts required for reproducing the figures, the ORCHIDEE simulations and intermediate results are available at https://github.com/j-jeong/J.Jeong_GMD_2020. BACI dataset is freely available in online: http://www.baci-h202i0.eu/ but requires registration by email.

8. Competing interest

The authors declare that they have no conflict of interest.

9. Author~~author~~ contribution

Proposed benchmarks are the outcome of discussions between JJ, JB, PP, VH, and SL. JJ ran the model, analysed the output and prepared figures. FB collected the~~and shared~~ BACI data. JJ and SL wrote the ~~first version of the~~ manuscript, all authors contributed to revising and editing the different versions of the~~final~~ manuscript.

10. Acknowledgements

JJ, PP, MJM and SL were funded by the VERIFY project under the European Union's Horizon 2020 research and innovation programme under grant agreement No. 776810. JB was supported by the Centre National de la Recherche Scientifique (CNRS) of France through the program ""Make Our Planet Great Again". VH acknowledges support from the Earth Systems and Climate Change Hub, funded by the Australian Government's National Environmental Science Program. SL would like to thank Antonio Lara (Universidad Austral de Chile) for early discussions on the topic. MNE was supported by NSF/AGS1903626 and the University of Maryland, and acknowledges insights arising from work with the PAGES/Data Assimilation and Proxy System Modeling Working Group. F.B. acknowledges funding from the project "Inside out"

(#POIR.04.04.00-00-5F85/18-00) funded by the HOMING programme of the Foundation for Polish Science co-financed by the European Union under the European Regional Development Fund. Stefan Klesse co-developed the European biomass network and provided management information.

745 **Table**

Table 1. Characteristics of the proposed~~four~~ benchmarks~~ making use of ITRDB records.~~ These benchmarks were designed to better constrain physiological and ecological processes in land-surface models. Given their intended~~the~~ use with~~of the~~ ITRDB data, the benchmarks had to propose solutions for well-known issues of the ITRDB (Table S2).

Benchmark	Metrics	Targeted process understanding	Solutions for meaningful model comparison with ITRDB	Figure
Size dependent growth	· RMSE ~~(Fig. 9a X-axis)~~ · Slope of the residuals ~~(Fig. 9a Y-axis)~~	· Long-term size~~Size~~-related growth · Within-stand competition	· Select the biggest tree of the simulation · Construct an average virtual tree aligned by tree age	Fig. 4~~5~~
Diameter increment of mature trees	· RMSE ~~(Fig. 9b X-axis)~~ · Slope of the residuals ~~(Fig. 9b Y-axis)~~	· Long-term tree~~Tree~~ growth after establishment ~~· Temporal shift in drivers~~ · Within-stand competition	· Select the biggest tree of the simulation · Construct an average virtual tree aligned by calendar year	Fig. 5~~6~~
Diameter increment of young trees	· RMSE ~~(Fig. 9c X-axis)~~ · Slope of the residuals ~~(Fig. 9c Y-axis)~~	· Short-term (i.e. 30-year) tree growth during establishment~~Tree growth~~ ~~· Temporal shift in drivers~~ · Size-related growth	· Select the biggest tree of the simulation · Construct a fast-growing virtual tree	Fig. 6~~7~~

Extreme growth	· Extreme events ~~(Fig. 9d X-axis)~~ · Amplitude ~~(Fig. 9d Y-axis)~~	· Yearly climate~~Climate~~ sensitivity	· Select the biggest tree of the simulation · Define extreme growth using 25% smallest and 75% largest observations	Fig. 7~~8~~

Table 2. Verification of the benchmarks and their metrics. Each cell represents the result from a single site. The values show the difference for each metric before and after optimization. Bold cells show the cases where the optimization for the all-tree data was inconsistent with the optimisation result of the big-tree data.

~~**Table 2**. Description of the processes included in the four configurations used in the test case~~

Benchmark~~Configuration~~	Size dependent growth~~Nitrogen dynamics~~		Diameter increment of mature trees~~Recruitment~~		Diameter increment of young trees~~Refined within-stand competition~~		Extreme growth~~Carbon, water and energy cycles~~	
Metrics~~Ndyn~~	**RMSE (mm)~~+~~**	**Slope of residuals (mm/yr) ~~+~~**	**RMSE (mm)~~+~~**	**Slope of residuals (mm/yr) ~~+~~**	**RMSE (mm)**	**Slope of residuals (mm/yr)**	**Amplitude (mm)**	**Extreme growth (scaled)**
Picea Abies~~Recur~~	DEO (-0.005)~~-~~	**DEO (0.000)~~+~~**	DEO (-75.97)~~+~~	**DEO (0.78)~~+~~**	**DEO (8.11)**	**DEO (1.47)**	DEO (-0.04)	DEO (-0.60)
	DVN (-0.182)~~-~~	**DVN (0.000)~~-~~**	DVN (-161.69) ~~+~~	DVN (-0.70)~~+~~	DVN (-11.68)	DVN (-0.43)	**DVN (0.02)**	DVN (-0.77)
	GIU (-0.600)~~-~~	GIU (-0.007)~~-~~	GIU (-131.25)~~-~~	GIU (-0.68)~~+~~	GIU (-39.70)	GIU (-1.30)	GIU (-0.32)	GIU (-0.96)
	HD2 (0.009)	HD2 (-0.002)	**HD2 (15.60)**	**HD2 (0.53)**	**HD2 (1.95)**	**HD2 (0.56)**	HD2 (-0.04)	HD2 (-0.97)
	SCH (0.029)	SCH (-0.004)	SCH (-182.46)	SCH (-0.96)	**SCH (57.56)**	**SCH (5.49)**	SCH (-0.15)	SCH (-1.29)
	SOB	**SOB**	SOB	**SOB**	**SOB**	**SOB**	SOB	SOB

	(-0.008)	**(0.001)**	(-20.22)	**(0.50)**	**(9.32)**	**(1.29)**	(-0.08)	(-1.54)
	TIC (-0.151)	TIC (-0.001)	**TIC (24.47)**	**TIC (1.52)**	TIC (-10.33)	TIC (-0.27)	TIC (-0.19)	TIC (-1.63)
Fagus sylvatica	CAN (-0.046)	CAN (-0.003)	CAN (-74.03)	CAN (-1.27)	CAN (-5.81)	**CAN (0.16)**	CAN (-0.07)	CAN (-1.17)
	SOR (0.007)	SOR (-0.004)	SOR (-116.26)	SOR (-1.62)	**SOR (2.69)**	SOR (-1.13)	SOR (-0.04)	SOR (-1.00)
	TER (-0.06)	TER (-0.000)	TER (-3.73)	TER (-0.08)	TER (-15.93)	**TER (0.26)**	TER (-0.07)	TER (-0.99)
	ZOF (-0.183)	ZOF (-0.000)	ZOF (-42.72)	**ZOF (0.02)**	ZOF (-11.98)	ZOF (-0.05)	ZOF (-0.17)	ZOF (-1.11)

Figures

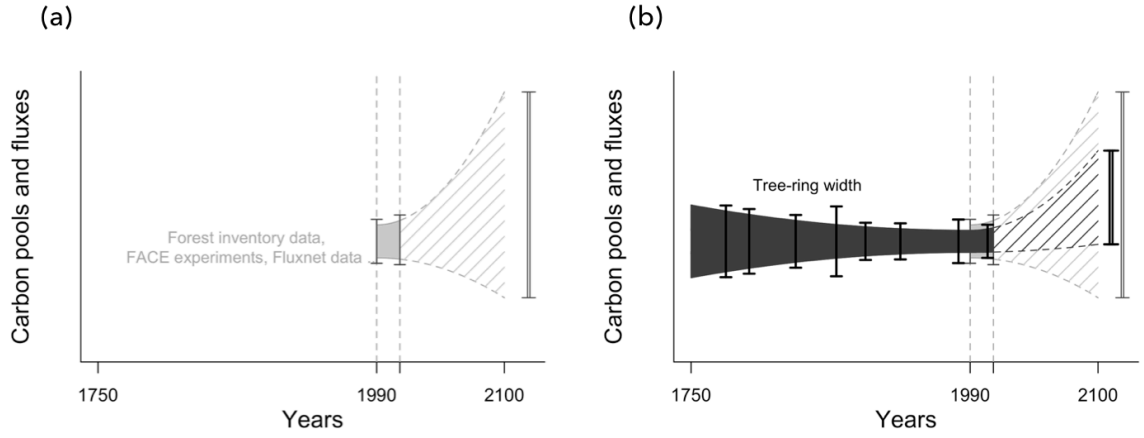


Figure 1. Conceptual illustration of the expected reduction in model uncertainty following the use of tree-ring width records to benchmark land-surface models. Note that the anticipated uncertainty reduction assumes that a large part of the model uncertainty comes from the model formulation and its parameters rather than from the initial conditions and drivers. (a) Observational constraints (grey vertical bars) from short-term benchmarks such as forest inventory data, FACE experiments, and FLUXNET data, have been used to parameterize and evaluate the response of ecosystems to environmental changes (light-grey coloured area). When used in projecting the present-day to future carbon pools and fluxes, uncertainty in ecosystem response to climate change is propagated through the model resulting in unacceptably large uncertainties (light-grey hatched area). (b) Tree-ring records going back to pre-industrial times (black vertical bars) are expected to better constrain the response of ecosystems to environmental changes (dark-grey coloured area) which should result in smaller uncertainties when used to project future ecosystem responses (dark-grey hatched area).

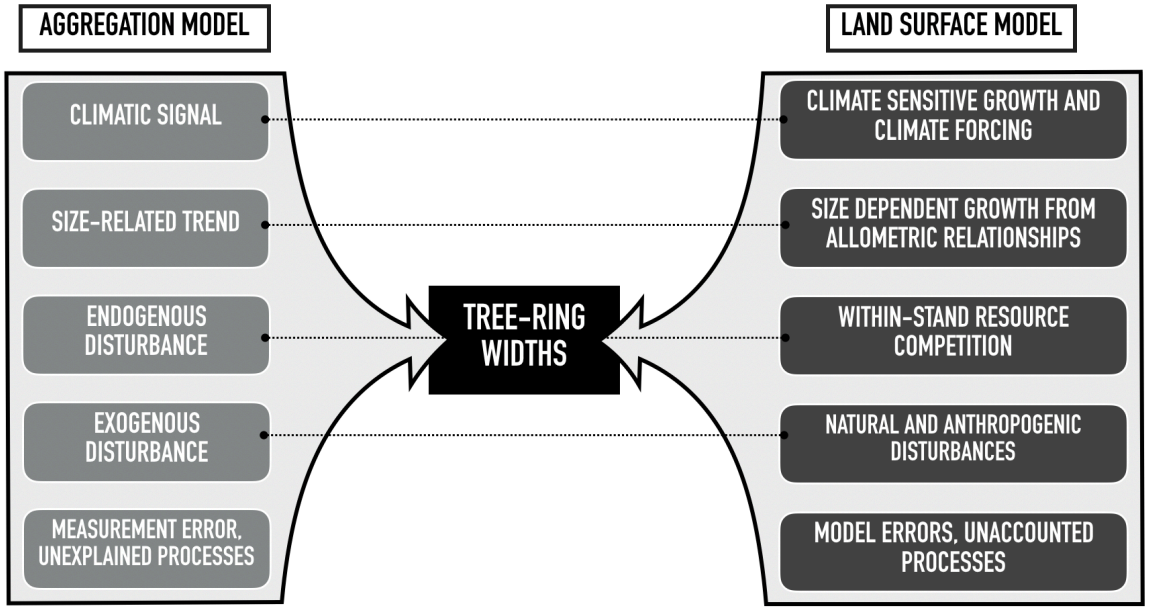
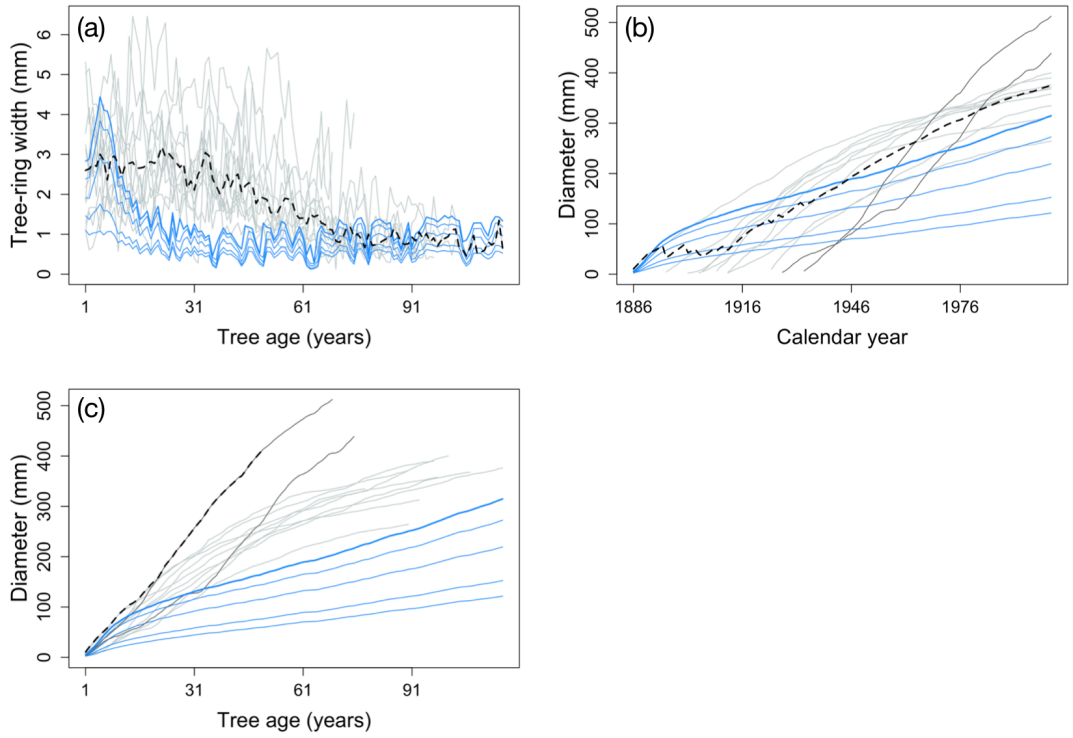


Figure 2. Main drivers of the linear aggregate conceptual~~tree growth~~ model of tree-ring growth and the equivalent processes in land-surface models. The dotted lines connect the related components. Note that both the aggregation and the land-surface model come with errors, uncertainties and unaccounted processes which are not explicitly modelled.


(a)
Tree-ring width (mm)
0
1
2
3
4
5
6
1
31
61
91
Tree age (years)
(b)
Diameter (mm)
0
100
200
300
400
500
1886
1916
1946
1976
Calendar year
(c)
Diameter (mm)
0
100
200
300
400
500
1
31
61
91
Tree age (years)

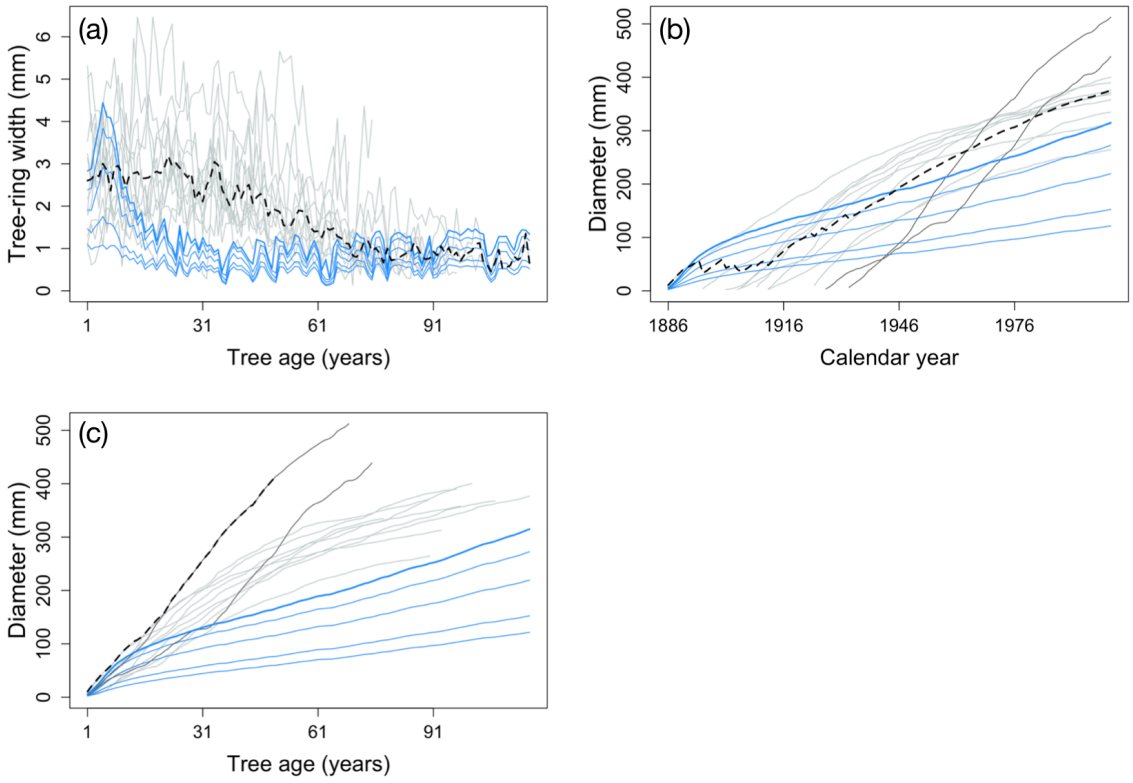


Figure 3. Solutions and virtual trees~~tree compilations~~ to account for challenges to use ITRDB datasets when evaluating land-surface models against ITRDB tree-ring data. (a) Data-model comparison may overcome the big-tree selection bias by comparing~~compare~~ only the simulated biggest diameter class (bold blue line) for evaluation rather than all diameter classes (thin blue lines), with the compiled virtual tree (black dotted line). Grey lines represent individual trees from observations. (b) The observed tree-ring records are a mixture of relatively slowly-growing trees (light-grey lines) and fast-growing trees (dark-grey lines). Fast-growing trees don't attain the same age as slowly-growing trees because they tend to die earlier. Without further consideration this would lead to underestimating tree growth at the time of stand establishment and thus result in a flawed test when compared against simulated tree growth (blue line) as the virtual tree (black dotted line) is much smaller during stand~~at the~~ establishment. (c) Aligning observations by the age of individual trees better reconstructs tree growth during stand establishment, facilitating data-model comparison. Note the change in the label of the X-axis between panels (b) and (c). Observations taken from a French oak forest archived as germ214 (NOAA, 2020c)~~(NOAA, 2020b)~~ (Table S2).

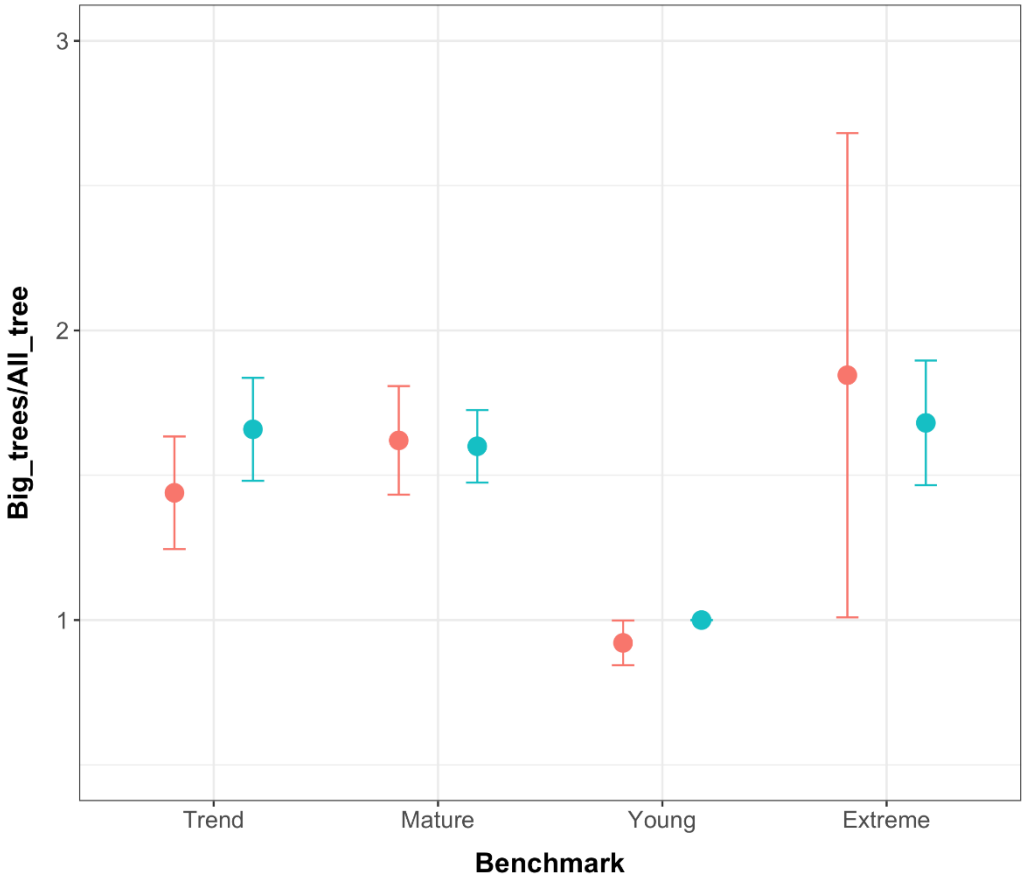


Figure 4. Comparison of bias for each of the benchmarks introduced by sampling the 15% biggest trees rather than the average tree (red; observational based) and the bias introduced by reporting the largest simulated diameter class rather than the average tree (blue; model based). Error bars are the standard deviation of the 27 coniferous sites for the BACI data set (red) or the 40 model configurations (4 model different set-ups for 10 sites each).

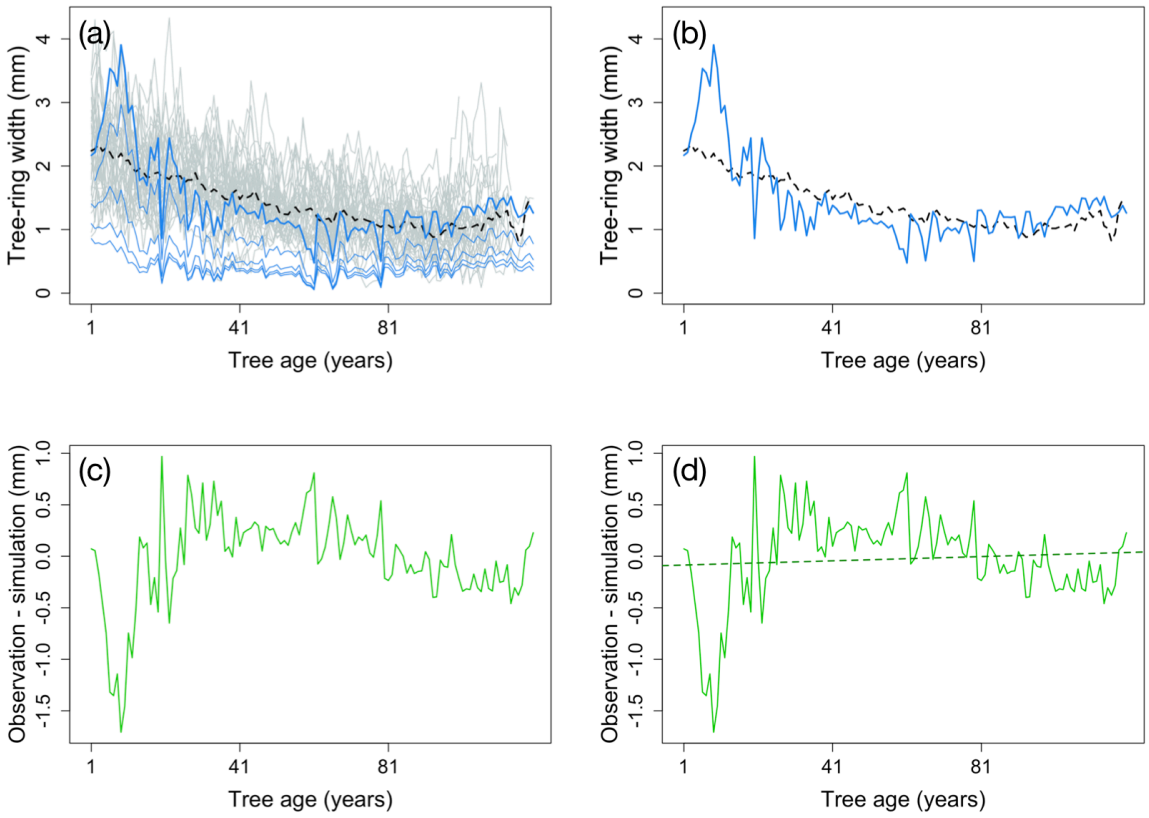
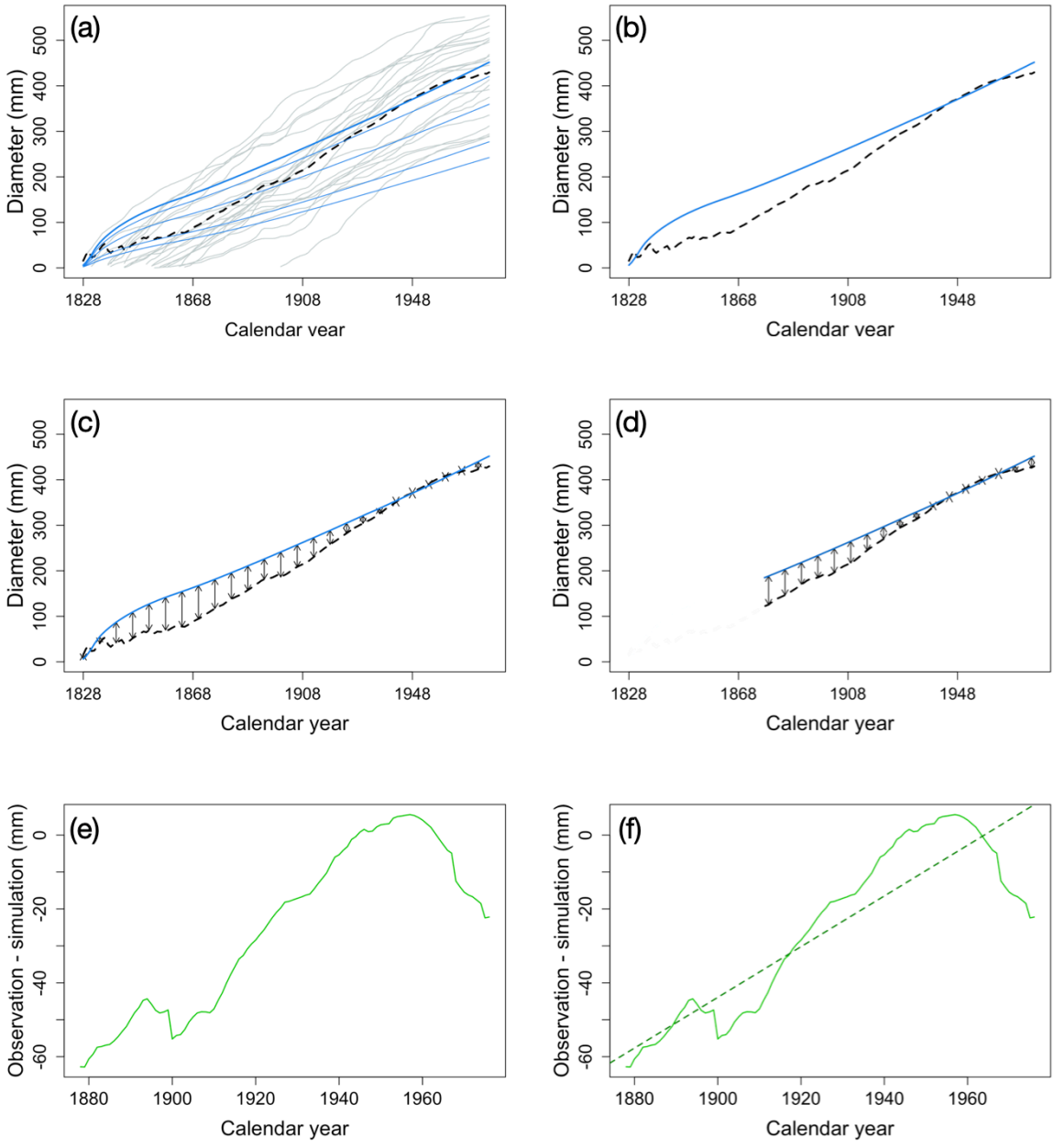


Figure 4. Example~~5. Illustration~~ of the major steps for~~in~~ calculating the metrics of the benchmark for the size-related trend in diameter increment. The size-related trend in diameter increment can be assessed by calculating a time series for the average ring width after aligning the age of the individual trees (a, b). Observations are shown as grey lines and simulation as blue lines. The biggest class is presented by the bold line. The black dotted lines represent the virtual tree based on the observations. The TRWs of this virtual tree are then subtracted from the simulated TRWs of the largest diameter class. (c~~).~~ Subsequently, a linear regression is used to quantify the temporal trend in the residuals (d). The green line denotes the model residuals and the green dotted line is the linear regression of the model residuals. Furthermore, the root mean square error (RMSE) between the simulations and observations is calculated (b~~not shown~~) and normalized by the length of time series to calculate the difference in observed and simulated growth trends. Observations and simulation are from site finl052 (NOAA, 2020b). For this example, calculated RMSE is 0.39 (mm), and the slope of residuals is -0.002 (mm/yr). ~~(NOAA, 2020a) (Table S2).~~


(a)
(b)
(c)
(d)
(e)
(f)
Diameter (mm)
Observation - simulation (mm)
Calendar year
1828
1868
1908
1948
1880
1900
1920
1940
1960
0
100
200
300
400
500
-20
-40
-60

805

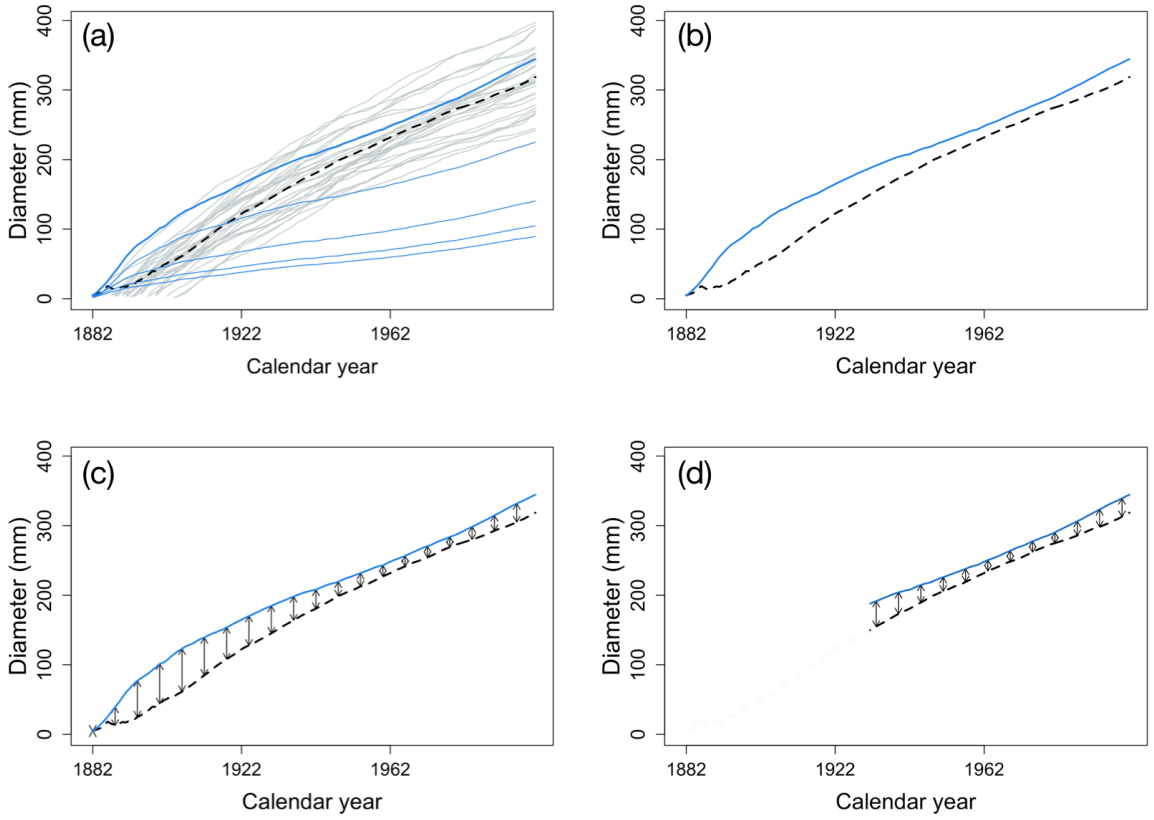
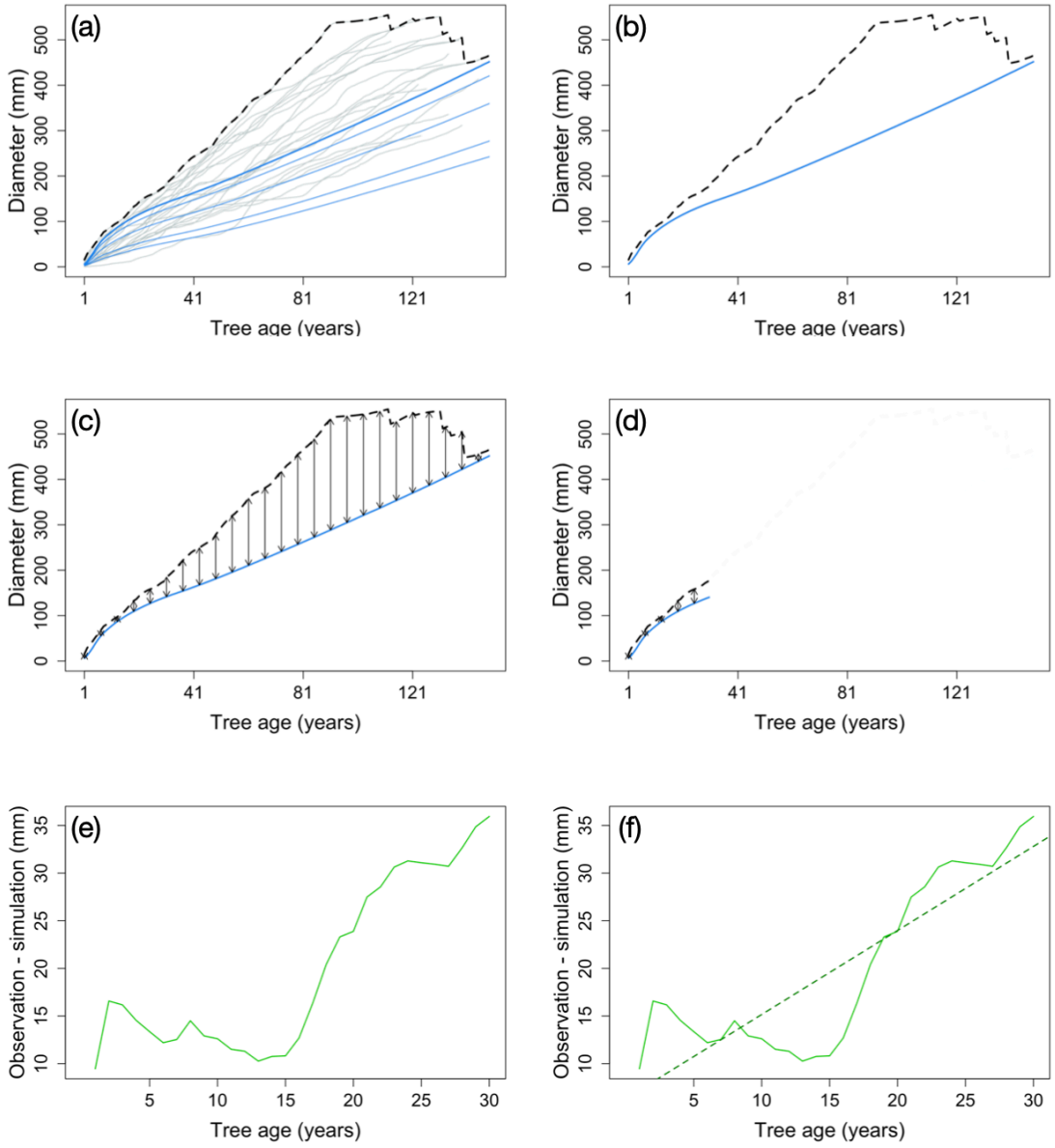


Figure 5. Example~~6. Illustration~~ of the major steps in calculating the metrics of the benchmark for the diameter increment in mature trees. Individual tree records are ordered by calendar year and for each year the average observed diameter is calculated (a). Observations are shown as grey lines and simulation as blue lines. The biggest class is presented by the bold line. Black dotted lines represent the yearly average of observations. Note that X-axis in Fig. 5~~6~~ is different from Fig. 4~~5~~. Under the assumption that the observed trees are the biggest trees from a given site, the virtual tree is compared with the biggest diameter class from the model (b) and the RMSE is calculated (c). Given that for the most recent decades both the fast and slow growing trees are still alive and could have been sampled, only the recent decades of the virtual tree growth are compared to the simulations. The RMSE (grey arrows) and trend ~~(not shown)~~ of the residuals between the virtual tree and the largest diameter class simulated are calculated (e, f). The x-axes of e, f zooms in on the selected period. The green line denotes the residuals and the green dotted line is the linear regression of the model residuals.~~d).~~ Observations and simulation are from site brit021~~fnl052~~ (NOAA, 2020a). In this case, RMSE and the slope of residuals were calculated as 33.65 (mm) and 0.68 (mm/yr), respectively.~~(NOAA, 2020a) (Table S2).~~


(a)
(b)
(c)
(d)
(e)
(f)
Diameter (mm)
Tree age (years)
Observation - simulation (mm)
0
100
200
300
400
500
10
15
20
25
30
35
1
41
81
121
5
10
15
20
25
30

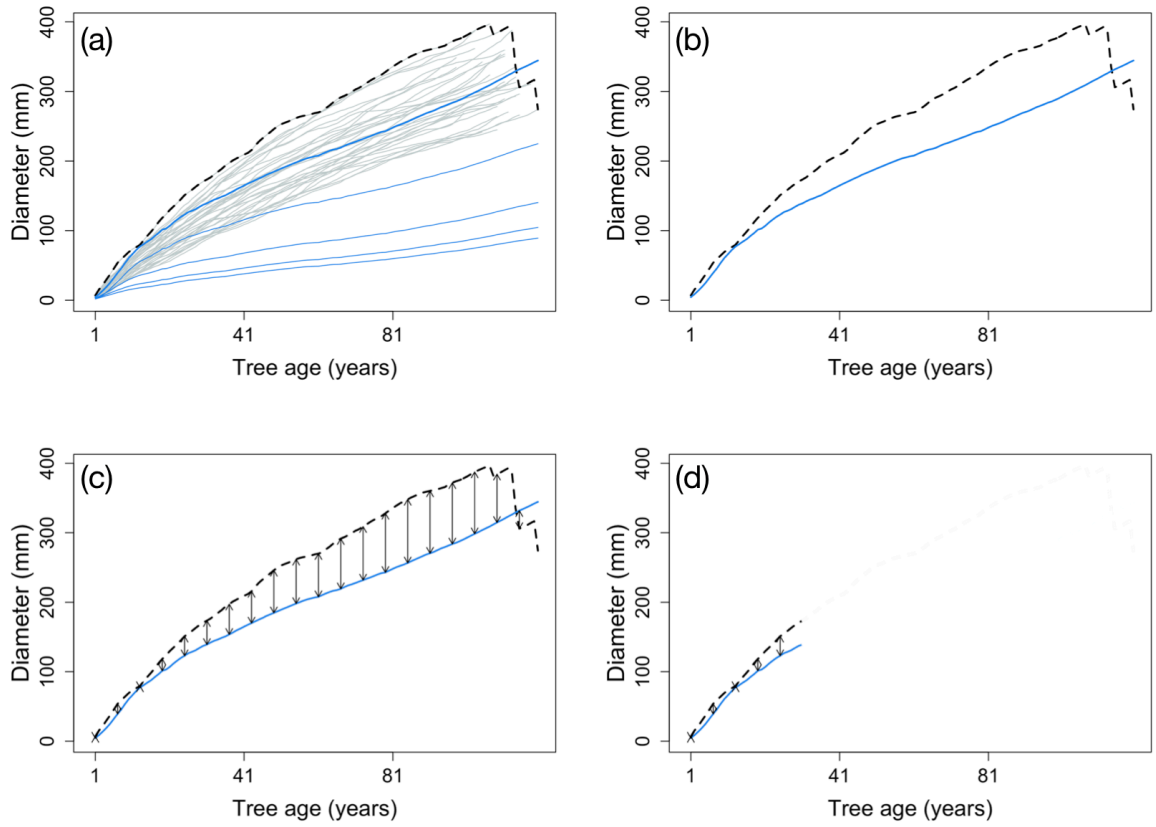
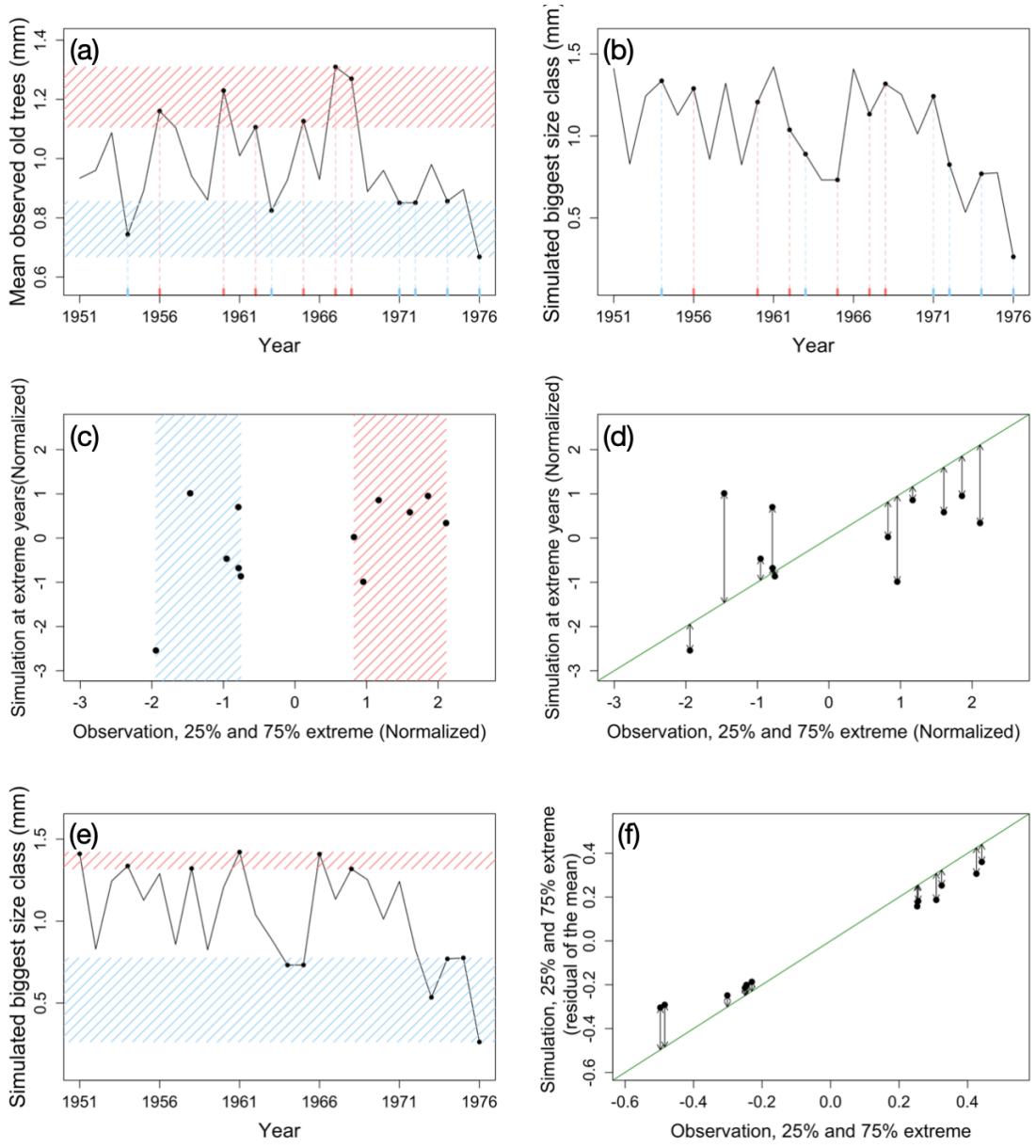


Figure 6. Example~~7. Illustration~~ of the major steps in calculating the metrics of the benchmark for diameter increment in young trees. After aligning the TRW records of the individual trees by their age, a virtual tree is constructed by taking the maximum observed diameter of all trees for each year (a). Observations are shown as grey lines and simulation as blue lines. The biggest class is presented by the bold line. Black dotted lines represent the yearly maximum of the observations. The growth of the virtual tree is then compared to the simulated growth of the largest diameter class (b) by calculating the RMSE (c) and trend of the residuals (e, f). The x-axes of e, f zooms in on the selected period, and the green line denotes the model residuals and the green dotted line is the linear regression of the model residuals.~~not shown).~~ These calculations are limited to the first decades of the time series (d) to compensate for the bias caused by the fact that the old fast-growing trees died well before sampling took place. By using different approaches to evaluate the growth of young (this benchmark) and mature trees (the previous benchmark) the comparison accounts for the observation that the drivers of ring growth change when the trees grow taller (Cook, 1985). Observations and simulation are from site brit021~~fnl052~~ (NOAA, 2020a). The calculated RMSE was 21.86 (mm) and the slope of residuals was 0.88 (mm/yr) for this example.~~(NOAA, 2020a) (Table S2).~~


(a)
Mean observed old trees (mm)
0.6
0.8
1.0
1.2
1.4
1951
1956
1961
1966
1971
1976
Year
(b)
Simulated biggest size class (mm)
0.5
1.0
1.5
1951
1956
1961
1966
1971
1976
Year
(c)
Simulation at extreme years(Normalized)
-3
-2
-1
0
1
2
-3
-2
-1
0
1
2
Observation, 25% and 75% extreme (Normalized)
(d)
Simulation at extreme years (Normalized)
-3
-2
-1
0
1
2
-3
-2
-1
0
1
2
Observation, 25% and 75% extreme (Normalized)
(e)
Simulated biggest size class (mm)
0.5
1.0
1.5
1951
1956
1961
1966
1971
1976
Year
(f)
Simulation, 25% and 75% extreme (residual of the mean)
-0.6
-0.4
-0.2
0.0
0.2
0.4
-0.6
-0.4
-0.2
0.0
0.2
0.4
Observation, 25% and 75% extreme

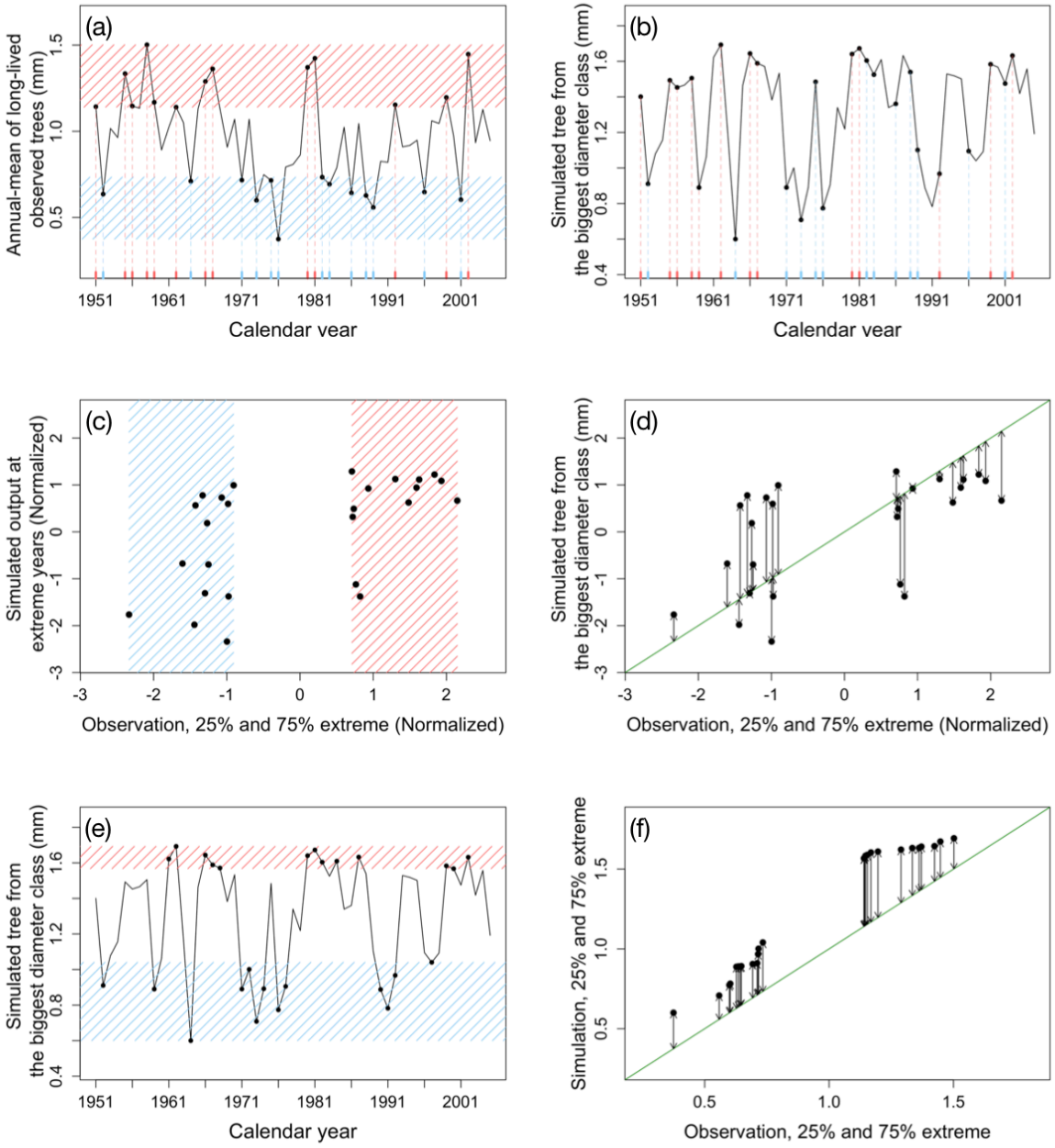


Figure 7. Example~~8. Illustration~~ of the major steps in calculating the metrics of the benchmark for extreme growth events. In this benchmark, extreme growth is defined as the first and last quartiles in TRW ordered by calendar year and averaged over the individual trees records (a). Red shaded area and ticks represent observations exceeding the 75 percentile and blue shaded area and ticks represents observation below the 25 percentile (a). The TRW simulated for the largest diameter class are then extracted for the years identified in a (b).

Both observations and simulations were normalized to remove the difference in the range of values between configurations. These normalized values correspond to the X and Y axis in (c) and (d) for observation and simulation, respectively. Subsequently, the similarity between simulations and observations was tested by calculating the distance from the 1:1 line (shown in green in d), which is equivalent to the RMSE for years with extreme growth (d). An additional metric is calculated in a similar way but by using both the 25% and 75% extreme values of the simulation and observation regardless of the year (e, f). This test identifies if the simulation can reproduce the amplitude of TRW. The observations and simulations were not normalized to~~To~~ assess the absolute amplitude. Possible~~, the observation and simulation were not normalized. To avoid possible~~ uncertainties from using reconstructed climate forcing, were avoided by limiting the calculations of both metrics ~~are limited~~ to the past five decades for which climate observations are available. Observations and simulation are from site spai006 (NOAA, 2020d). In this test case, RMSE for extreme years was 0.57 (mm) and RMSE for extreme growth was 0.03 (scaled).

~~(NOAA, 2020c) (Table S2).~~

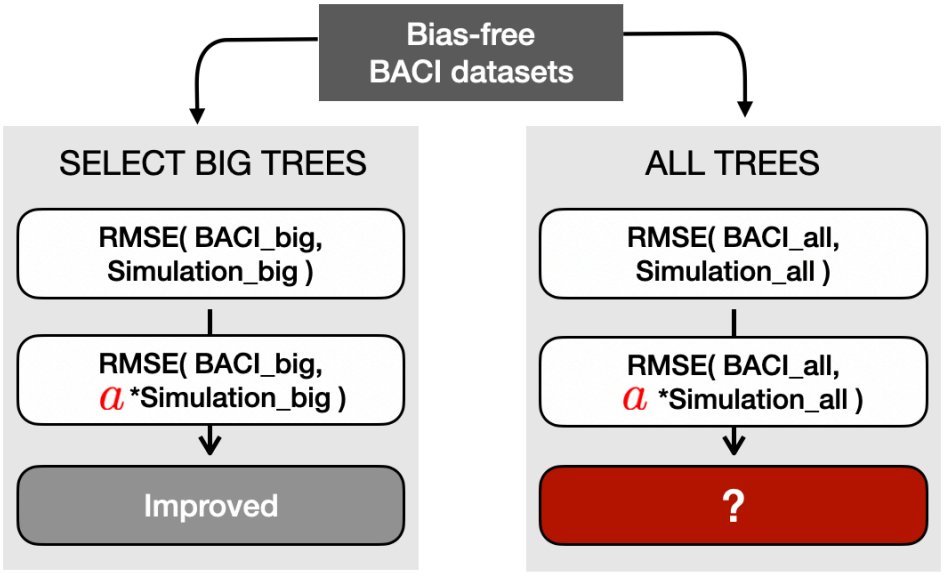
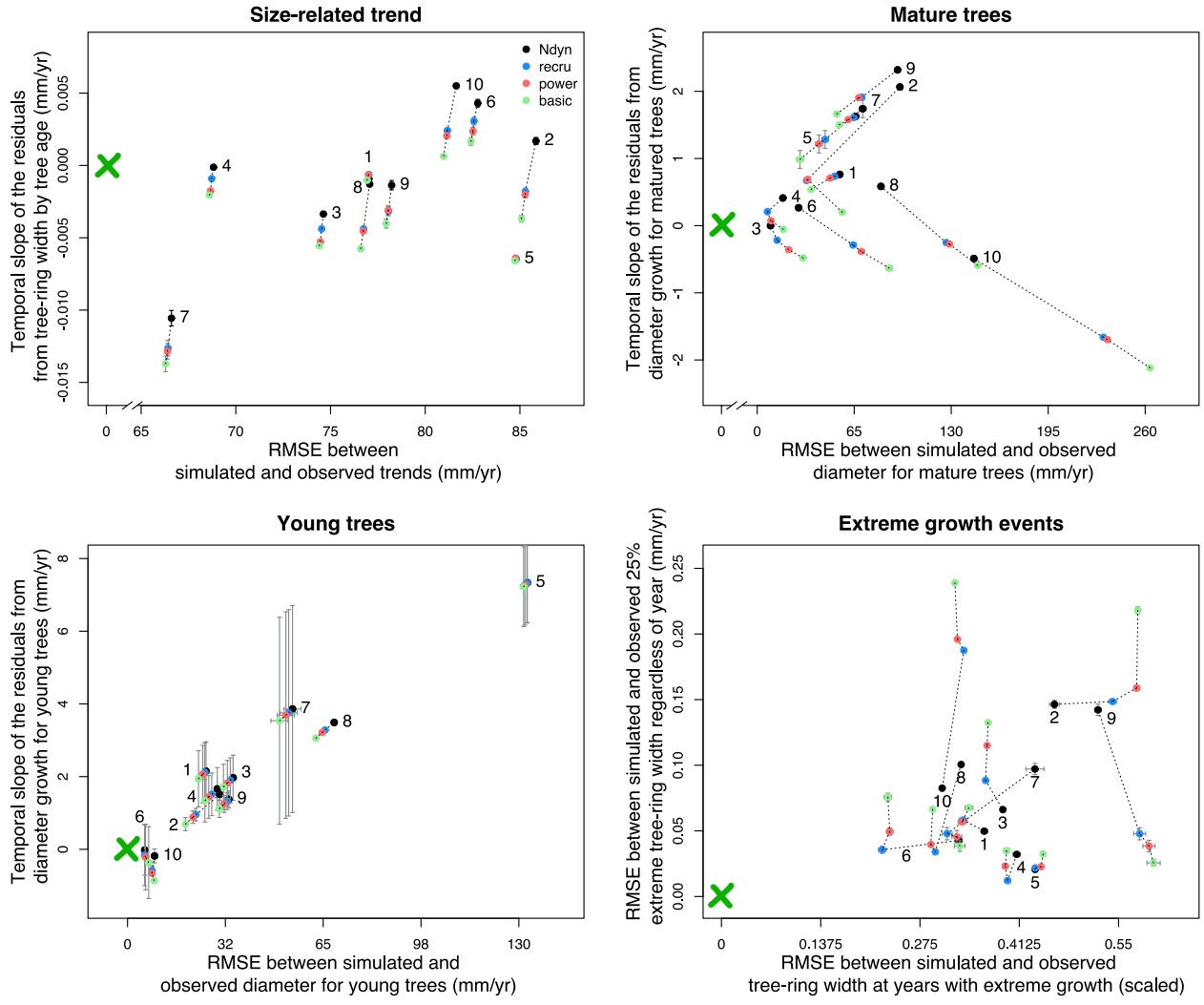


Figure 8. Schematic representation of the verification process for the RMSE-metric. Before the verification, two types of datasets were prepared: big-tree data (limited to the 15% biggest trees) and all-tree data. The arithmetic optimization proceeded by using big-tree data to find a multiplier for the simulated TRW that minimizes RMSE. The same multiplier was then applied to the all-tree data and the RMSE was calculated. Finally the decrease or increase in RMSE with the multiplier was compared to the RMSE obtained without the multiplier. The other two arithmetic optimization (Section 3.4) follow a similar approach.



~~**Figure 9. The proposed four benchmarks (a-d) applied to a European test case of 10 sites (Table S2 to link numbers to site names and locations). Each colour represents a configuration of the land surface model (Table 2 details the configurations), where black denotes configuration Ndyn, blue Recru, red Power and green the configuration labelled Basic. The green X marks indicate the desired outcomes. (a) Benchmark for the trend in tree ring width. The X axis shows the RMSE of the difference between simulated and observed trend (Fig. 5b) and the Y axis shows using the slope of the temporal trend in the residuals (Fig. 5d). (b) Benchmark for diameter increment of mature trees. The X axis shows the RMSE of the difference between simulation and averaged observation aligned by calendar year for matured trees (Fig. 6d), and the Y axis shows the slope of the temporal trend in the residuals. (c) Benchmark for diameter increment of young trees. The X axis shows the differences between simulation and averaged observation**~~

~~aligned by age of trees for young trees (Fig. 7d), and the Y-axis shows the slope of the temporal trend in the residuals. (d) Benchmark for climate sensitivity. The X-axis shows the RMSE of the difference between the observed and simulated tree ring widths for years in which the observed tree ring width was extreme. The Y-axis shows the RMSE of the difference between the observed and simulated extreme tree ring widths irrespective of the year they occur.~~

References

Alexander, M. R., Rollinson, C. R., Babst, F., Trouet, V. and Moore, D. J. P.: Relative influences of multiple sources of uncertainty on cumulative and incremental tree-ring-derived aboveground biomass estimates, Trees, 32(1), 265–276, doi:10.1007/s00468-017-1629-0, 2018.

Amthor, J. S.: The McCree–de Wit–Penning de Vries–Thornley Respiration Paradigms: 30 Years Later, Ann. Bot., 86(1), 1–20, doi:10.1006/anbo.2000.1175, 2000.

Babst, F., Alexander, M. R., Szejner, P., Bouriaud, O., Klesse, S., Roden, J., Ciais, P., Poulter, B., Frank, D., Moore, D. J. and Trouet, V.: A tree-ring perspective on the terrestrial carbon cycle, Oecologia, 176(2), 307–322, doi:10.1007/s00442-014-3031-6, 2014a.

Babst, F., Bouriaud, O., Alexander, R., Trouet, V. and Frank, D.: Toward consistent measurements of carbon accumulation: A multi-site assessment of biomass and basal area increment across Europe, Dendrochronologia, 32(2), 153–161, doi:https://doi.org/10.1016/j.dendro.2014.01.002, 2014b.

Babst, F., Poulter, B., Bodesheim, P., Mahecha, M. D. and Frank, D. C.: Improved tree-ring archives will support earth-system science, Nat Ecol Evol, 1(2), 8, doi:10.1038/s41559-016-0008, 2017.

Babst, F., Bodesheim, P., Charney, N., Friend, A. D., Girardin, M. P., Klesse, S., Moore, D. J. P., Seftigen, K., Björklund, J. and Bouriaud, O.: When tree rings go global: challenges and opportunities for retro-and prospective insight, Quat. Sci. Rev., 197, 1–20, 2018.

Bakker, J. D.: A new, proportional method for reconstructing historical tree diameters, Can. J. For. Res., 35(10), 2515–2520, doi:10.1139/x05-136, 2005.

Bellassen, V., Le Maire, G., Dhôte, J. F., Ciais, P. and Viovy, N.: Modelling forest management within a global vegetation model—Part 1: Model structure and general behaviour, Ecol. Modell., 221(20), 2458–2474, doi:10.1016/j.ecolmodel.2010.07.008, 2010.

Blyth, E., Gash, J., Lloyd, A., Pryor, M., Weedon, G. P. and Shuttleworth, J.: Evaluating the JULES Land Surface Model Energy Fluxes Using FLUXNET Data, J. Hydrometeorol., 11(2), 509–519, doi:10.1175/2009JHM1183.1, 2010.

Bonan, G. B., Williams, M., Fisher, R. A. and Oleson, K. W.: Modeling stomatal conductance in the earth system: linking leaf water-use efficiency and water transport along the soil–plant–atmosphere continuum, Geosci. Model Dev., 7(5), 2193–2222, doi:10.5194/gmd-7-2193-2014, 2014.

Bowman, D. M. J. S., Brienen, R. J. W., Gloor, E., Phillips, O. L. and Prior, L. D.: Detecting trends in tree growth: not so simple, Trends Plant Sci., 18(1), 11–17, doi:10.1016/j.tplants.2012.08.005, 2013.

Brienen, R. J. W., Gloor, E. and Zuidema, P. A.: Detecting evidence for CO 2 fertilization from tree ring studies: The potential role of sampling biases, Global Biogeochem. Cycles, 26(1), n/a-n/a, doi:10.1029/2011GB004143, 2012.

Briffa, K. R. and Melvin, T. M.: A Closer Look at Regional Curve Standardization of Tree-Ring Records: Justification of the Need, a Warning of Some Pitfalls, and Suggested Improvements in Its Application, in Dendroclimatology, pp. 113–145, Springer., 2011.

Briffa, K. R., Osborn, T. J. and Schweingruber, F. H.: Large-scale temperature inferences from tree rings: a review, Glob. Planet. Change, 40(1), 11–26, doi:https://doi.org/10.1016/S0921-8181(03)00095-X, 2004.

Bunde, A., Büntgen, U., Ludescher, J., Luterbacher, J. and von Storch, H.: Is there memory in precipitation?, Nat. Clim. Chang., 3(3), 174–175, doi:10.1038/nclimate1830, 2013.

Campbell, J. E., Berry, J. A., Seibt, U., Smith, S. J., Montzka, S. A., Launois, T., Belviso, S., Bopp, L. and Laine, M.: Large historical growth in global terrestrial gross primary production, Nature, 544(7648), 84–87, doi:10.1038/nature22030, 2017.

Cao, X., Tian, F., Li, F., Gaillard, M.-J., Rudaya, N., Xu, Q. and Herzschuh, U.: Pollen-based quantitative land-cover reconstruction for northern Asia covering the last 40 ka cal BP, Clim. Past, 15(4), 1503–1536, doi:10.5194/cp-15-1503-2019, 2019.

Cedro, A.: Growth-climate relationships of wild service trees on the easternmost range boundary in Poland, STR16/04, 24, doi:10.2312/GFZ.b103-16042, 2016.

Chen, Y.-Y., Gardiner, B., Pasztor, F., Blennow, K., Ryder, J., Valade, A., Naudts, K., Otto, J., McGrath, M. J., Planque, C. and Luyssaert, S.: Simulating damage for wind storms in the land surface model ORCHIDEE-CAN (revision 4262), Geosci. Model Dev., 11(2), 771–791, doi:10.5194/gmd-11-771-2018, 2018.

Chen, Y., Yang, K., He, J., Qin, J., Shi, J., Du, J. and He, Q.: Improving land surface temperature modeling for dry land of

China, J. Geophys. Res. Atmos., 116(D20), doi:10.1029/2011JD015921, 2011.

Chen, Y., Ryder, J., Bastrikov, V., McGrath, M. J., Naudts, K., Otto, J., Ottlé, C., Peylin, P., Polcher, J., Valade, A., Black, A., Elbers, J. A., Moors, E., Foken, T., van Gorsel, E., Haverd, V., Heinesch, B., Tiedemann, F., Knohl, A., Launiainen, S., Loustau, D., Ogée, J., Vesala, T. and Luyssaert, S.: Evaluating the performance of the land surface model ORCHIDEE-CAN on water and energy flux estimation with a single- and a multi- layer energy budget scheme, Geosci. Model Dev. Discuss., 1–35, doi:10.5194/gmd-2016-26, 2016.

Compo, G. P., Whitaker, J. S., Sardeshmukh, P. D., Matsui, N., Allan, R. J., Yin, X., Gleason, B. E., Vose, R. S., Rutledge, G., Bessemoulin, P., Brönnimann, S., Brunet, M., Crouthamel, R. I., Grant, A. N., Groisman, P. Y., Jones, P. D., Kruk, M. C., Kruger, A. C., Marshall, G. J., Maugeri, M., Mok, H. Y., Nordli, Ø., Ross, T. F., Trigo, R. M., Wang, X. L., Woodruff, S. D. and Worley, S. J.: The Twentieth Century Reanalysis Project, Q. J. R. Meteorol. Soc., 137(654), 1–28, doi:10.1002/qj.776, 2011.

Cook, E. R.: A time series analysis approach to tree ring standardization, 1985.

Cook, E. R. and Kairiukstis, L. A.: Methods of Dendrochronology, edited by E. R. Cook and L. A. Kairiukstis, Springer Netherlands, Dordrecht., 1990.

Cook, E. R., Briffa, K. R., Meko, D. M., Graybill, D. A. and Funkhouser, G.: The "segment length curse" in long tree-ring chronology development for palaeoclimatic studies, The Holocene, 5(2), 229–237, doi:10.1177/095968369500500211, 1995.

Curtis, P. S., Hanson, P. J., Bolstad, P., Barford, C., Randolph, J. ., Schmid, H. . and Wilson, K. B.: Biometric and eddy-covariance based estimates of annual carbon storage in five eastern North American deciduous forests, Agric. For. Meteorol., 113(1–4), 3–19, doi:10.1016/S0168-1923(02)00099-0, 2002.

D'Arrigo, R., Wilson, R., Liepert, B. and Cherubini, P.: On the 'Divergence Problem' in Northern Forests: A review of the tree-ring evidence and possible causes, Glob. Planet. Change, 60(3), 289–305, doi:https://doi.org/10.1016/j.gloplacha.2007.03.004, 2008.

Deleuze, C. and Houllier, F.: Simple process-based xylem growth model for describing wood microdensitometric profiles, J. Theor. Biol., 193(1), 99–113, doi:10.1006/jtbi.1998.0689, 1998.

Deleuze, C., Pain, O., Dhôte, J.-F. and Hervé, J.-C.: A flexible radial increment model for individual trees in pure even-aged

stands, Ann. For. Sci., 61(4), 327–335, doi:10.1051/forest:2004026, 2004.

Demarty, J., Chevallier, F., Friend, A. D., Viovy, N., Piao, S. and Ciais, P.: Assimilation of global MODIS leaf area index retrievals within a terrestrial biosphere model, Geophys. Res. Lett., 34(15), doi:10.1029/2007GL030014, 2007.

Drew, D. M., Downes, G. M. and Battaglia, M.: CAMBIUM, a process-based model of daily xylem development in Eucalyptus, J. Theor. Biol., 264(2), 395–406, doi:10.1016/j.jtbi.2010.02.013, 2010.

Ducoudré, N. I., Laval, K. and Perrier, A.: SECHIBA, a New Set of Parameterizations of the Hydrologic Exchanges at the Land-Atmosphere Interface within the LMD Atmospheric General Circulation Model, J. Clim., 6(2), 248–273, doi:10.1175/1520-0442(1993)006<0248:SANSOP>2.0.CO;2, 1993.

Dufrêne, E., Davi, H., François, C., Le Maire, G., Le Dantec, V. and Granier, A.: Modelling carbon and water cycles in a beech forest. Part I: Model description and uncertainty analysis on modelled NEE, Ecol. Modell., 185(2–4), 407–436, doi:10.1016/j.ecolmodel.2005.01.004, 2005.

Farquhar, G. D.: Models of Integrated Photosynthesis of Cells and Leaves, Philos. Trans. R. Soc. B Biol. Sci., 323(1216), 357–367, doi:10.1098/rstb.1989.0016, 1989.

Fisher, R. A., Muszala, S., Verteinstein, M., Lawrence, P., Xu, C., McDowell, N. G., Knox, R. G., Koven, C., Holm, J., Rogers, B. M., Spessa, A., Lawrence, D. and Bonan, G.: Taking off the training wheels: the properties of a dynamic vegetation model without climate envelopes, CLM4.5(ED), Geosci. Model Dev., 8(11), 3593–3619, doi:10.5194/gmd-8-3593-2015, 2015.

Franklin, O., Johansson, J., Dewar, R. C., Dieckmann, U., McMurtrie, R. E., Brännström, Å. and Dybzinski, R.: Modeling carbon allocation in trees: a search for principles, Tree Physiol., 32(6), 648–666, doi:10.1093/treephys/tpr138, 2012.

Friedlingstein, P., Cox, P., Betts, R., Bopp, L., Von Bloh, W., Brovkin, V., Cadule, P., Doney, S., Eby, M., Fung, I., Bala, G., John, J., Jones, C., Joos, F., Kato, T., Kawamiya, M., Knorr, W., Lindsay, K., Matthews, H. D., Raddatz, T., Rayner, P., Reick, C., Roeckner, E., Schnitzler, K. G., Schnur, R., Strassmann, K., Weaver, A. J., Yoshikawa, C. and Zeng, N.: Climate-carbon cycle feedback analysis: Results from the (CMIP)-M-4 model intercomparison, J. Clim., 19(14), 3337–3353, doi:10.1175/jcli3800.1, 2006.

Friedlingstein, P., Meinshausen, M., Arora, V. K., Jones, C. D., Anav, A., Liddicoat, S. K. and Knutti, R.: Uncertainties in CMIP5 climate projections due to carbon cycle feedbacks, J. Clim., 27(2), 511–526, doi:10.1175/JCLI-D-12-00579.1, 2014.

Friend, A. D., Eckes-Shephard, A. H., Fonti, P., Rademacher, T. T., Rathgeber, C. B. K., Richardson, A. D. and Turton, R. H.: On the need to consider wood formation processes in global vegetation models and a suggested approach, Ann. For. Sci., 76(2), doi:10.1007/s13595-019-0819-x, 2019.

Fritts, H. C.: Tree rings and climate, Elsevier., 2012.

Fritts, H. C., Shashkin, A. and Downes, G. M.: A simulation model of conifer ring growth and cell structure, Tree-ring Anal. Biol. Methodol. Environ. Asp. CABI Publ. Wallingford, UK, (January 1999), 3–32, 1999.

Grissino-Mayer, H. D. and Fritts, H. C.: The International Tree-Ring Data Bank: an enhanced global database serving the global scientific community, The Holocene, 7(2), 235–238, doi:10.1177/095968369700700212, 1997.

Haverd, V., Smith, B., Cook, G. D., Briggs, P. R., Nieradzik, L., Roxburgh, S. H., Liedloff, A., Meyer, C. P. and Canadell, J. G.: A stand-alone tree demography and landscape structure module for Earth system models, Geophys. Res. Lett., 40(19), 5234–5239, doi:10.1002/grl.50972, 2013.

Hayat, A., Hacket-Pain, A. J., Pretzsch, H., Rademacher, T. T. and Friend, A. D.: Modeling Tree Growth Taking into Account Carbon Source and Sink Limitations, Front. Plant Sci., 8, doi:10.3389/fpls.2017.00182, 2017.

Hecht, A. D. and Tirpak, D.: Framework agreement on climate change: a scientific and policy history, Clim. Change, 29(4), 371–402, doi:10.1007/BF01092424, 1995.

Hemming, D., Fritts, H., Leavitt, S. W., Wright, W., Long, A. and Shashkin, A.: Modelling tree-ring δ13C, Dendrochronologia, 19(1), 23–38, 2001.

Hirata, R., Hirano, T., Saigusa, N., Fujinuma, Y., Inukai, K., Kitamori, Y., Takahashi, Y. and Yamamoto, S.: Seasonal and interannual variations in carbon dioxide exchange of a temperate larch forest, Agric. For. Meteorol., 147(3), 110–124, doi:https://doi.org/10.1016/j.agrformet.2007.07.005, 2007.

Hölttä, T., Vesala, T., Sevanto, S., Perämäki, M. and Nikinmaa, E.: Modeling xylem and phloem water flows in trees according to cohesion theory and Münch hypothesis, Trees, 20(1), 67–78, doi:10.1007/s00468-005-0014-6, 2006.

Hughes, M. K., Swetnam, T. W. and Diaz, H. F.: Dendroclimatology, edited by M. K. Hughes, T. W. Swetnam, and H. F. Diaz, Springer Netherlands, Dordrecht., 2011.

IPCC: Annex VI: Expert Reviewers of the IPCC WGI Fifth Assessment Report, in Climate Change 2013: The Physical Science

Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, edited by T. F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, and P. M. Midgley, pp. 1497–1522, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA., 2013.

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., Zhu, Y., Leetmaa, A., Reynolds, R., Chelliah, M., Ebisuzaki, W., Higgins, W., Janowiak, J., Mo, K. C., Ropelewski, C., Wang, J., Jenne, R. and Joseph, D.: The NCEP/NCAR 40-Year Reanalysis Project, Bull. Am. Meteorol. Soc., 77(3), 437–471, doi:10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2, 1996.

De Kauwe, M. G., Medlyn, B. E., Zaehle, S., Walker, A. P., Dietze, M. C., Hickler, T., Jain, A. K., Luo, Y., Parton, W. J., Prentice, I. C., Smith, B., Thornton, P. E., Wang, S., Wang, Y.-P., Wårlind, D., Weng, E., Crous, K. Y., Ellsworth, D. S., Hanson, P. J., Seok Kim, H.-, Warren, J. M., Oren, R. and Norby, R. J.: Forest water use and water use efficiency at elevated CO2: a model-data intercomparison at two contrasting temperate forest FACE sites, Glob. Chang. Biol., 19(6), 1759–1779, doi:10.1111/gcb.12164, 2013.

Keeling, C. D., Chin, J. F. S. and Whorf, T. P.: Increased activity of northern vegetation inferred from atmospheric CO2 measurements, Nature, 382(6587), 146–149, doi:10.1038/382146a0, 1996.

Klesse, S., Babst, F., Lienert, S., Spahni, R., Joos, F., Bouriaud, O., Carrer, M., Di Filippo, A., Poulter, B., Trotsiuk, V., Wilson, R. and Frank, D. C.: A Combined Tree Ring and Vegetation Model Assessment of European Forest Growth Sensitivity to Interannual Climate Variability, Global Biogeochem. Cycles, 32(8), 1226–1240, doi:10.1029/2017GB005856, 2018.

Kolus, H. R., Huntzinger, D. N., Schwalm, C. R., Fisher, J. B., McKay, N., Fang, Y., Michalak, A. M., Schaefer, K., Wei, Y., Poulter, B., Mao, J., Parazoo, N. C. and Shi, X.: Land carbon models underestimate the severity and duration of drought's impact on plant productivity, Sci. Rep., 9(1), 2758, doi:10.1038/s41598-019-39373-1, 2019.

Krinner, G., Viovy, N., de Noblet-Ducoudré, N., Ogée, J., Polcher, J., Friedlingstein, P., Ciais, P., Sitch, S. and Prentice, I. C.: A dynamic global vegetation model for studies of the coupled atmosphere-biosphere system, Global Biogeochem. Cycles, 19(1), doi:10.1029/2003GB002199, 2005.

Laloyaux, P., de Boisseson, E., Balmaseda, M., Bidlot, J.-R., Broennimann, S., Buizza, R., Dalhgren, P., Dee, D., Haimberger, L., Hersbach, H., Kosaka, Y., Martin, M., Poli, P., Rayner, N., Rustemeier, E. and Schepers, D.: CERA-20C: A Coupled

Reanalysis of the Twentieth Century, J. Adv. Model. Earth Syst., 10(5), 1172–1195, doi:10.1029/2018MS001273, 2018.

LaMarche, V. C., Graybill, D. A., Fritts, H. C. and ROSE, M. R.: Increasing Atmospheric Carbon Dioxide: Tree Ring Evidence for Growth Enhancement in Natural Vegetation, Science (80-.)., 225(4666), 1019–1021, doi:10.1126/science.225.4666.1019, 1984.

Levesque, M., Andreu-Hayles, L., Smith, W. K., Williams, A. P., Hobi, M. L., Allred, B. W. and Pederson, N.: Tree-ring isotopes capture interannual vegetation productivity dynamics at the biome scale, Nat. Commun., 10(1), 742, doi:10.1038/s41467-019-08634-y, 2019.

Li, G., Harrison, S. P., Prentice, I. C. and Falster, D.: Simulation of tree-ring widths with a model for primary production, carbon allocation, and growth, Biogeosciences, 11(23), 6711–6724, doi:10.5194/bg-11-6711-2014, 2014.

Luo, Y. Q., Randerson, J. T., Friedlingstein, P., Hibbard, K., Hoffman, F., Huntzinger, D., Jones, C. D., Koven, C., Lawrence, D., Li, D. J., Abramowitz, G., Bacour, C., Blyth, E., Carvalhais, N., Ciais, P., Dalmonech, D., Fisher, J. B., Fisher, R., Friedlingstein, P., Hibbard, K., Hoffman, F., Huntzinger, D., Jones, C. D., Koven, C., Lawrence, D., Li, D. J., Mahecha, M., Niu, S. L., Norby, R., Piao, S. L., Qi, X., Peylin, P., Prentice, I. C., Riley, W., Reichstein, M., Schwalm, C., Wang, Y. P., Xia, J. Y., Zaehle, S. and Zhou, X. H.: A framework for benchmarking land models, Biogeosciences, 9(10), 3857–3874, doi:10.5194/bg-9-3857-2012, 2012.

Magnani, F., Mencuccini, M., Borghetti, M., Berbigier, P., Berninger, F., Delzon, S., Grelle, A., Hari, P., Jarvis, P. G., Kolari, P., Kowalski, A. S., Lankreijer, H., Law, B. E., Lindroth, A., Loustau, D., Manca, G., Moncrieff, J. B., Rayment, M., Tedeschi, V., Valentini, R. and Grace, J.: The human footprint in the carbon cycle of temperate and boreal forests, Nature, 447(7146), 849–851, doi:10.1038/nature05847, 2007.

McGuffie A., K. and H.: Practical Climate Modelling, in A Climate Modelling Primer., 2005.

Melvin, T.: Historical growth rates and changing climatic sensitivity of boreal conifers, University of East Anglia., 2004.

Mencuccini, M., Martínez-Vilalta, J., Vanderklein, D., Hamid, H. A., Korakaki, E., Lee, S., Michiels, B., Martínez-Vilalta, J., Vanderklein, D., Hamid, H. A., Korakaki, E., Lee, S. and Michiels, B.: Size-mediated ageing reduces vigour in trees, Ecol. Lett., 8(11), 1183–1190, doi:10.1111/j.1461-0248.2005.00819.x, 2005.

Merganičová, K., Merganič, J., Lehtonen, A., Vacchiano, G., Sever, M. Z. O., Augustynczik, A. L. D., Grote, R., Kyselová,

I., Mäkelä, A., Yousefpour, R., Krejza, J., Collalti, A. and Reyer, C. P. O.: Forest carbon allocation modelling under climate change, Tree Physiol., doi:10.1093/treephys/tpz105, 2019.

Misson, L., Rathgeber, C. and Guiot, J.: Dendroecological analysis of climatic effects on Quercus petraea and Pinus halepensis radial growth using the process-based MAIDEN model, Can. J. For. Res., 34(4), 888–898, doi:10.1139/x03-253, 2004.

Moorcroft, P. R., Hurtt, G. C. and Pacala, S. W.: A Method for Scaling Vegetation Dynamics: The Ecosystem Demography Model (ED), Ecol. Monogr., 71(4), 557–585, doi:10.2307/3100036, 2001.

Nash, S. E.: Fundamentals of tree-ring research. James H. Speer., Geoarchaeology, 26(3), 453–455, doi:10.1002/gea.20357, 2011.

Naudts, K., Ryder, J., McGrath, M. J., Otto, J., Chen, Y., Valade, A., Bellasen, V., Berhongaray, G., Bönisch, G., Campioli, M., Ghattas, J., De Groote, T., Haverd, V., Kattge, J., MacBean, N., Maignan, F., Merilä, P., Penuelas, J., Peylin, P., Pinty, B., Pretzsch, H., Schulze, E. D., Solyga, D., Vuichard, N., Yan, Y. and Luyssaert, S.: A vertically discretised canopy description for ORCHIDEE (SVN r2290) and the modifications to the energy, water and carbon fluxes, Geosci. Model Dev., 8(7), 2035–2065, doi:10.5194/gmd-8-2035-2015, 2015.

Nehrbass-Ahles, C., Babst, F., Klesse, S., Nötzli, M., Bouriaud, O., Neukom, R., Dobbertin, M. and Frank, D.: The influence of sampling design on tree-ring-based quantification of forest growth, Glob. Chang. Biol., 20(9), 2867–2885, doi:10.1111/gcb.12599, 2014.

Nicklen, E. F., Roland, C. A., Csank, A. Z., Wilmking, M., Ruess, R. W. and Muldoon, L. A.: Stand basal area and solar radiation amplify white spruce climate sensitivity in interior Alaska: Evidence from carbon isotopes and tree rings, Glob. Chang. Biol., 25(3), 911–926, doi:10.1111/gcb.14511, 2019.

Nickless, A., Scholes, R. J. and Archibald, S.: A method for calculating the variance and confidence intervals for tree biomass estimates obtained from allometric equations, S. Afr. J. Sci., 107(5/6), 86–95, doi:10.4102/sajs.v107i5/6.356, 2011.

NOAA: finl052, NOAA/WDS for Paleoclimatology, https://www.ncdc.noaa.gov/paleo/study/3998, 2020a.

NOAA: germ214, NOAA/WDS for Paleoclimatology, https://www.ncdc.noaa.gov/paleo/study/16747, 2020b.

NOAA: spai006, NOAA/WDS for Paleoclimatology, https://www.ncdc.noaa.gov/paleo/study/4405, 2020c.

Oliver, C. D. and Larson, B. C.: Forest stand dynamics, Wiley New York., 1996.

Ols, C., Girardin, M. P., Hofgaard, A., Bergeron, Y. and Drobyshev, I.: Monitoring Climate Sensitivity Shifts in Tree-Rings of Eastern Boreal North America Using Model-Data Comparison, Ecosystems, 21(5), 1042–1057, doi:10.1007/s10021-017-0203-3, 2018.

Rammig, A., Wiedermann, M., Donges, J. F., Babst, F., Von Bloh, W., Frank, D., Thonicke, K. and Mahecha, M. D.: Coincidences of climate extremes and anomalous vegetation responses: comparing tree ring patterns to simulated productivity, Biogeosciences, 12(2), 373–385, doi:10.5194/bg-12-373-2015, 2015.

Randerson, J. T., Hoffman, F. M., Thorton, P. E., Mahowald, N. M., Lindsay, K., Lee, Y., Nevison, C. D., Doney, S. C., Bonan, G., Stöckli, R., Covey, C., Running, S. W. and Fung, I. Y.: Systematic assessment of terrestrial biogeochemistry in coupled climate-carbon models, Glob. Chang. Biol., 15(10), 2462–2484, doi:10.1111/j.1365-2486.2009.01912.x, 2009.

Ryder, J., Polcher, J., Peylin, P., Ottlé, C., Chen, Y., van Gorsel, E., Haverd, V., McGrath, M. J., Naudts, K., Otto, J., Valade, A. and Luyssaert, S.: A multi-layer land surface energy budget model for implicit coupling with global atmospheric simulations, Geosci. Model Dev., 9(1), 223–245, doi:10.5194/gmd-9-223-2016, 2016.

Sato, H., Itoh, A. and Kohyama, T.: SEIB–DGVM: A new Dynamic Global Vegetation Model using a spatially explicit individual-based approach, Ecol. Modell., 200(3–4), 279–307, doi:10.1016/j.ecolmodel.2006.09.006, 2007.

De Schepper, V. and Steppe, K.: Development and verification of a water and sugar transport model using measured stem diameter variations, J. Exp. Bot., 61(8), 2083–2099, doi:10.1093/jxb/erq018, 2010.

Schulman, E.: Longevity under Adversity in Conifersca, Science (80-.)., 119(3091), 396–399 [online] Available from: http://www.jstor.org/stable/1682970, 1954.

Smith, B.: LPJ-GUESS-an ecosystem modelling framework, Dep. Phys. Geogr. Ecosyst. Anal. INES, Sölvegatan, 12, 22362, 2001.

Steppe, K., De Pauw, D. J. W., Lemeur, R. and Vanrolleghem, P. A.: A mathematical model linking tree sap flow dynamics to daily stem diameter fluctuations and radial stem growth, Tree Physiol., 26(3), 257–273, doi:10.1093/treephys/26.3.257, 2006.

Stine, A. R.: Global demonstration of local Liebig's law behavior for tree-ring reconstructions of climate, Paleoceanogr. Paleoclimatology, 34, doi:https://doi.org/10.1029/2018PA003449, 2019.

Temme, A. A., Liu, J. C., Cornwell, W. K., Cornelissen, J. H. C. and Aerts, R.: Winners always win: growth of a wide range of plant species from low to future high CO2, Ecol. Evol., 5(21), 4949–4961, doi:10.1002/ece3.1687, 2015.

Vaganov, E. A., Hughes, M. K. and Shashkin, A. V: Growth dynamics of conifer tree rings: images of past and future environments, edited by M. K. Hughes, T. W. Swetnam, and H. F. Diaz, Springer, New York., 2006.

Viovy, N.: CRUNCEP data set, 2016.

Vuichard, N., Messina, P., Luyssaert, S., Guenet, B., Zaehle, S., Ghattas, J., Bastrikov, V. and Peylin, P.: Accounting for carbon and nitrogen interactions in the global terrestrial ecosystem model ORCHIDEE (trunk version, rev 4999): multi-scale evaluation of gross primary production, Geosci. Model Dev., 12(11), 4751–4779, doi:10.5194/gmd-12-4751-2019, 2019.

Wilkinson, S., Ogée, J. J., Domec, J.-C. C., Rayment, M. and Wingate, L.: Biophysical modelling of intra-ring variations in tracheid features and wood density of Pinus pinaster trees exposed to seasonal droughts, Tree Physiol., 35(3), 305–318, doi:10.1093/treephys/tpv010, 2015.

Williams, M., Richardson, A. D., Reichstein, M., Stoy, P. C., Peylin, P., Verbeeck, H., Carvalhais, N., Jung, M., Hollinger, D. Y., Kattge, J., Leuning, R., Luo, Y., Tomelleri, E., Trudinger, C. M. and Wang, Y.-P.: Improving land surface models with FLUXNET data, Biogeosciences, 6(7), 1341–1359, doi:10.5194/bg-6-1341-2009, 2009.

Wilson, B. F. and Howard, R. A.: A computer model for cambial activity, For. Sci., 14(1), 77–90, doi:10.1093/forestscience/14.1.77, 1968.

Wolf, A., Ciais, P., Bellassen, V., Delbart, N., Field, C. B. and Berry, J. A.: Forest biomass allometry in global land surface models, Global Biogeochem. Cycles, 25(3), doi:10.1029/2010GB003917, 2011.

Yue, C., Ciais, P., Cadule, P., Thonicke, K., Archibald, S., Poulter, B., Hao, W. M., Hantson, S., Mouillot, F., Friedlingstein, P., Maignan, F. and Viovy, N.: Modelling the role of fires in the terrestrial carbon balance by incorporating SPITFIRE into the global vegetation model ORCHIDEE – Part 1: simulating historical global burned area and fire regimes, Geosci. Model Dev., 7(6), 2747–2767, doi:10.5194/gmd-7-2747-2014, 2014.

Zaehle, S. and Friend, A. D.: Carbon and nitrogen cycle dynamics in the O-CN land surface model: 1. Model description, site-scale evaluation, and sensitivity to parameter estimates, Global Biogeochem. Cycles, 24(1), doi:10.1029/2009GB003521, 2010.

Zhang, Z., Babst, F., Bellassen, V., Frank, D., Launois, T., Tan, K., Ciais, P. and Poulter, B.: Converging Climate Sensitivities of European Forests Between Observed Radial Tree Growth and Vegetation Models, Ecosystems, 21(3), 410–425, doi:10.1007/s10021-017-0157-5, 2018.

Zhao, S., Pederson, N., D'Orangeville, L., HilleRisLambers, J., Boose, E., Penone, C., Bauer, B., Jiang, Y. and Manzanedo, R. D.: The International Tree-Ring Data Bank (ITRDB) revisited: Data availability and global ecological representativity, J. Biogeogr., 46(2), 355–368, doi:10.1111/jbi.13488, 2019.

Zuidema, P. A., Vlam, M. and Chien, P. D.: Ages and long-term growth patterns of four threatened Vietnamese tree species, Trees, 25(1 LB-Zuidema2011), 29–38, doi:10.1007/s00468-010-0473-2, 2011.

Zuidema, P. A., Poulter, B. and Frank, D. C.: A Wood Biology Agenda to Support Global Vegetation Modelling, Trends Plant Sci., 23(11), 1006–1015, doi:10.1016/j.tplants.2018.08.003, 2018.