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, Switzerland.
[7]School of Natural Resources and the Environment, University of Arizona, Tucson, USA.
[8]Laboratory of Tree-Ring Research, University of Arizona, Tucson, USA.

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

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~~ have recorded 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, however, have been collected with a view ~~on~~ of 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~~ has thus limited its use as a data source for ~~benchmarking~~model evaluation. The aim of this study is to ~~examine~~ propose and examine an improved conceptual framework of when and how ITRDB data can, despite ~~its~~ their sampling biases, be used as century-long ~~benchmarks for~~ data for evaluating LSMs. Combining advances in modelling and data processing resulted in four complementary benchmarks ~~, reflecting~~ for evaluating size related diameter growth, diameter increment of mature trees, diameter increment of young trees, and the response of tree growth to extreme events. These benchmarks reflect different usage of the information contained in the ITRDB~~, each~~ , and each of them is described by two performance metrics rooted in statistics ~~. Each of the~~ such as RMSE and linear regression. The four proposed benchmarks ~~was verified by calculating it twice~~were verified at 11 sites by calculating them: (1) based on an independent European tree-ring network ~~of biomass plots that were sampled in a locally representative way and are thus not biased by~~ for which sampling was designed to overcome the big-tree ~~tree selection~~ selection bias to better represent stand biomass, and (2) following sub-sampling of this European biomass network by only considering the 15% biggest trees to reproduce the big-tree selection bias present in the ITRDB. This study showed that ~~the ITRDB data can be used with~~ in about 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~~ during extreme climate years. The ITRDB data can be used with 70% confidence of 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. Although the proposed benchmarks are unlikely to be exact, 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 the ITRDB archive, but at the same time, illustrate ways in which tree-ring width observations may be collected and processed to provide long-term validation tests for land surface models.

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

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

1 Introduction

Earth system models integrate numerical ~~models~~ submodels 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~~ all Assessment Reports of the Intergovernmental Panel on Climate Change (**?**) and as such have made a tremendous impact on global environmental policy (**?**). The ~~50~~ credibility of projections of the future climate from any Earth system model in part relies on the ability of each of its four submodels to accurately reproduce the past (**?**). Although long-term changes that date back to pre-industrial conditions (**?**) have been documented for vegetation distribution through pollen based reconstructions (**?**), 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~~ substantially to uncertainties in simulated future global carbon stocks in soil and vegetation (**??**) and as such to climate projections (Fig. ~~1a~~S1a).

Tree-ring records provide annual information on historical tree growth and physiology in relation to environmental conditions, including ~~the era~~ during the time before human activities started to affect the atmospheric ~~CO2~~ CO_2 concentration (**??**). Even though trees grown in the absence of a clear annual rhythm of vegetative and dormant seasons may not develop distinct tree rings, as observed for many species from the humid tropics, 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~~S1b; See section 6 for more details) (**?????**).

Until now, tree-ring records have often been collected to reconstruct past climate and hydrological variability from sites where trees grow near the colder or drier fringes of their distribution (**??**). 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 (Fig. ~~S1~~S2) (**??**). 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 (**????**). 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"; **?**) that is available through the database of the BACI project (**?**), the aforementioned issues may limit the information content of the ITRDB records. This ~~loss in~~ incomplete information content should, however, be balanced against the associated benefits in terms of time gain and resource savings when re-using the large ITRDB dataset.

~~When~~ If tree rings are to be used as benchmarks for LSMs, the models must demonstrate skillful ~~simulation~~ simulations of 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. The first TRW models (**?**) described processes at the cell level: cell division, cell enlargement, and cell wall thickening. Later, the carbon and water balance of trees was added (**?**) as well as ~~climatic influences~~ a parameterization of the influences of

climate on cambial activity (**?**). 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 (**?????**).

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~~the representation of processes. In these models, photosynthate availability, air temperature and soil water content were used to constrain wood cell growth and successfully reproduced observations (**???**). Further simplifications were proposed by simulating the radial growth of trees based solely on carbon allocation (**??**) rather than cell dynamics, the latter being computationally too expensive for large scale vegetation models (**???**). Hence, a variety of approaches is now available to describe TRW growth in forest models, dynamic vegetation models and LSMs, but to the best of our knowledge there is yet no land-surface component of any 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 against tree-ring records archived in the ITRDB; (2) review potential issues of using the ITRDB to benchmark LSMs; (3) propose solutions for a meaningful comparison of LSMs against ITRDB records; and (4) verify the proposed solutions by benchmarking a LSM using ~~a dataset~~ data from a European biomass network (**?**) that is not prone to sampling biases related to ~~palaeoclimatological~~ dendroclimatic research. Dependencies between these aims and the workflow of this study are detailed in Fig. ~~2.~~ **??**.

2 Background: model requirements, data limitations and benchmarks

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

The ~~linear aggregate conceptual~~ conceptual linear aggregate model of tree growth (**?**) considers that the observed TRW at year t (in mm) consists of five additive growth contributions (Fig. ~~3)~~ **??**, left column) and as such provides a framework for simulating tree-ring widths with (semi)mechanistic model approaches (Fig. **??**):

(i) Size-dependent growth is the dominant signal in raw tree-ring measurements (**?**). Conceptually~~it can be understood by considering~~ , an almost constant volume of wood due to a more or less constant primary production (**?**) being added to the trunks year after year (**?**). The annual diameter increment of the trees ~~will decrease~~ decreases 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~~ tends to reduce stand density and competition for resources~~, implying that the remaining trees can~~ . The trees left can thus increase their crown volume and ~~thus increase~~ their primary production (**?**) 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 account for size- dependent growth and stand self-thinning (**??**).

(ii) Climate-dependent growth reflects the sensitivity of tree growth to radiation, temperature, phenology, and water availability (**?**) and is accounted for in LSMs, as it represents the core purpose of this type of models. LSMs often rely on the Farquhar model for the radiation and temperature dependency of photosynthesis (**?**), the McCree - de Wit - Penning de Vries - Thornley approach for the temperature dependence of respiration (**?**)~~, and~~ . They account for a decoupling of photosynthesis and growth by the use of a labile carbon pool (**???**). Plant water availability is ~~accounted for~~ represented through either simple transfer functions or more recently by accounting for the hydraulic architecture of the simulated trees (**??**).

(iii) Endogenous disturbances refer to within-stand resource competition and are being increasingly simulated in LSMs albeit often by empirical approaches (**???**). 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).

(iv) Chronic exogenous disturbances such as increasing atmospheric ~~CO2~~ CO_2 concentration (**?**) and N-deposition (**?**) are well-developed as they are among the main purposes of using LSMs. The effect of ~~CO2~~ CO_2 fertilization on photosynthesis is accounted for in the photosynthetic submodel whereas nitrogen dynamics are accounted for through static or dynamic stoichiometric approaches (**??**).

~~(iv)~~ Although abrupt disturbances such as fires, pests and storms are increasingly being simulated by LSMs (**??**) and ~~they~~ leave marks in TRW (**??**), e.g., fire scars and missing rings, ~~these functionalities are at present~~ they are 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~~ 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. Thus simulated stand demographics ~~, rather than secular change, might~~ should be the basis for benchmarking against observations rather than secular changes such as infrequent disturbances described above.

(v) The final term in the aggregate tree-growth model constitutes all processes and interactions between processes not ~~145~~ previously accounted for in the LSM, 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 (**??**)~~, which~~ . These methods rely on the assumptions that the sampled trees capture the ~~relevant~~ common growth variability of the stand ~~and that~~ (e.g., growth responses to climate variability and resource competition), and the contribution of each major driver can be statistically identified as either signal or noise. Nevertheless, alternative ~~ecological models~~ approaches based on Liebig's Law of the minimum (**?**) have been proposed to attribute TRW to

its major drivers~~(?)~~. In practice, observed TRW records cannot always be fully decomposed in the absence of metadata because several drivers might not leave a unique fingerprint in growth. However, size effect and climate sensitivity have a much larger contribution to TRW than the other processes (**?**).

In addition to accurate process representation, ~~the model~~ LSMs will need to be driven by historical climate, atmospheric ~~CO2 con- centrations~~ CO_2 concentrations and N-deposition. In general, commonly-used century-long climate reanalyses such as NCEP (**?**), 20CR (**?**), and CERA-20C (**?**) are based on the assimilation of instrumental observations in climate simulations and are thus independent from climate estimates derived from tree rings or other proxy data. Nevertheless, random and systematic ~~error~~ errors in the reanalyses increase as data availability ~~decrease~~decreases, particularly in remote areas with a low density and temporal depth of meteorological stations. Given that local climate effects may have contributed to the TRW, it might be desirable to ~~bias-correct the~~ correct the bias in reanalysis with present day site-specific climate observations where they exist (**?**). 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 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. ~~3~~**??**). 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 ~~170~~ stand are at present only accounted for in few LSMs, for example, CLM (ED) (**?**), ORCHIDEE (**?**), and LPJ-GUESS (**?**). 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 of increment cores from tens of individual trees from the same site and species. Each record may have different starting and ending dates, and thus length (Fig. ~~4~~ **??** a and b). If a core reaches the ~~centre~~ pith of the trunk~~(i.e., pith), ,~~ annual tree diameter can be reconstructed (**?**). 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~~pith, which is often the case for large trees, the lack of information about the rings near the pith ~~will be missed adding~~ adds uncertainty to the diameter and age reconstruction (**?**). In this case, diameter increment ~~can~~ could still be reconstructed (by subtracting the measured TRW from the diameter of the tree) if trunk diameter at the time of sampling ~~is~~ was known, but this metadata is rarely recorded in dendroclimatic collections and it is not stored in the ITRDB.

Despite of its known biases, the ITRDB 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 most useful to reconstruct the climate variability prior to instrumental records. The ITRDB is thus likely to ~~overrepresent~~ over-represent

large trees (**??**) 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 (**??**) 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, this bias can be accommodated through 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.4~~illustrated by the black dotted and highest blue lines in Fig. **??**a).

Another bias related to the ITRDB sampling design comes from the fact that the growth rate of trees within a cohort differs between individuals (**??**) resulting in ~~slowly~~ slow and fast-growing trees within the same cohort (~~Fig. 4b~~illustrated by the grey lines in Fig. **??**b). Slow-growing trees tend to live longer than fast-growing trees in the same cohort (**??**). ~~Records of TRW~~ Owing to survivorship being biased towards slow-growing individuals (**?**), TRW records 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 (**?**). Another challenge of using ITRDB data comes from the difference between the observed and simulated forest structure. Tree-ring datasets are composed by cores of individuals from different cohorts (Fig. ~~4b~~**??**b). Comparing these data against simulations requires the model to be individual-based or to align TRW records by age (Fig. ~~4a~~**??**c).

Given the above, for sites for which the sampling protocol is poorly described or not rigorously enforced 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~~ by making use of site-specific virtual trees. ~~'Virture tree' is "not physically present as s (**?**) !!!!!!!~~

~~Virtual trees will, however, also require careful post-processing of the ITRDB data to produce an intercomparison between analogous simulations and observations . Later in this study we propose four different benchmarks based on the ITRDB data , three of which make use of a virtual tree. Nevertheless, each of these benchmarks addresses a different aspect of TRW , and therefore uses a different definition for its virtual tree:~~

~~(i) The average virtual tree of a stand aligned by tree age is calculated as the time series for~~ Virtual trees are created cautiously from observations by combining data from different individuals to obtain a time series of TRW with a desired property (see section 2.3 for details on the desired properties). By definition, virtual trees were not observed and therefore not present as such in the ITRDB observations. Since for a single site, the ~~average ring width after aligning the age of the individual trees (Fig. 4a). Age-aligned TRWs are widely used to calculate a statistic known as the mean regional curve of the sampled stand (**?**) . This assumes that common drivers regardless of time, exceed the signal from local and individual differences in tree growth (see subsection 2.3 (i)).~~

(ii) The average virtual tree of a stand aligned by calendar year is calculated by ordering individual tree-ring series by calendar year (Fig. 4b)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 2.3 (ii)).

(iii) The largest virtual tree of a stand is calculated after aligning individual trees by their age (Fig. 4c). The recommendation to remove the age trend from observations consists of samples from multiple individual trees (individual trees are shown by grey lines in Fig. **??**), constructing a single virtual tree facilitates data-model comparisons. Moreover, varying the way the tree-ring records (**?**) 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 treesfor 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 are aligned (compare Figs. **??** b and c; details are in Section 2.3(iii)).) altered the properties and intended use of the virtual trees.

The proposed data-model comparison thus largely relies on the concept of virtual treesto account , since these virtual trees can account better for known sampling biases of the ITRDB datasets, different aspect of TRW as well as for the model definition of a forest standto facilitate comparing simulations and observations. The proposed definitions and uses of virtual trees which were partly customized to ORCHIDEE r5698, are evaluated in section 4.1. Except for LSMs with an individual tree-based stand definition (**?**), benchmarking other models against ITRDB data will also have to consider the use of virtual trees and may have to adjust the proposed definitions to the peculiarities of each modelthe LSM under evaluation.

2.3 Benchmarks for comparing observed and simulated tree-ring widths

If a LSM explicitly accounts for the main factors contributing contributors to TRW, i.e., size effects and climate sensitivity (**?**), 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 1S1):

(i) Size related diameter growth. **Size related diameter growth.** The size-related growth trend in diameter increment can be assessed by calculating the average virtual tree for a stand aligned by tree age (Figexamples shown in Fig. **??**a and Fig. **??** a and b). 5 a , b)and subtracting its TRWs Age-aligned TRWs are widely used to calculate a statistic known as the mean regional curve of the sampled stand (**?**). Aligning by age assumes that size-related growth is the dominant driver of tree growth (Section 2.1). Subsequently, the average virtual tree is subtracted from the simulated TRWs of the largest diameter class (Fig. 5c). Subsequently**??**). Next, a linear regression is used to quantify the temporal trend in the residuals (Fig. 5dexamples shown

in Fig. ??d). 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.~~ **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 (**??**). In reality, growing conditions can suddenly become ~~favourable~~ favorable for trees that have previously been suppressed, resulting in fluctuating growth rates ~~.~~ (see dark-grey lines in Fig. **??**b). 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~~. For this benchmark the average virtual tree of a stand aligned by calendar year ~~, taking the average tree diameter of all samples to construct the virtual tree~~ is used and calculated by ordering individual tree-ring series by calendar year (Fig. ~~6~~ **??**b and Fig. **??**a and b) ~~.~~ ~~Under the assumption~~ and calculating the average diameter for each year. Alignment by calendar year thus reflects the real temporal evolution of the stand. Following the big-tree selection bias (Section 2.2) it can be assumed that the observed trees are representative of the biggest trees from a given site. Hence, the virtual tree can be compared with the biggest diameter class from the model. ~~Given that for the last~~ The survivorship bias of slow-growing individuals has the strongest impact when assessing TRW in century old trees (Section 2.2). When analyzing recent decades both the quick and slowly growing trees are still alive and could have been sampled, therefore, only the growth in recent decades of the virtual tree ~~are~~ should be compared to the simulations (Fig. ~~6~~ **??** c and d). This virtual tree was thus used to better cope with differences in simulated and observed forest structure. The RMSE and trend of the residuals between the virtual tree and the largest diameter class simulated are calculated (Fig. ~~6d~~**??**d). A skilled model is expected to simultaneously show no trend in the residuals and a low RMSE across many sites.

(iii) ~~Diameter~~ **Diameter increment of young trees.** The 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.~~ largest virtual tree of a stand is calculated after aligning individual trees by their age (Fig. **??**c and Fig. **??** a and b). The recommendation to remove the age trend from tree-ring records (**?**) confirms the assumption underlying 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 (See the difference in the weight of dark grey lines to virtual trees in Fig.**??** b and c). 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. ~~7~~ **??** a and b). The virtual fast-growing tree thus gives a better idea of the true mean young tree growth in old stands. The first three 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~~**??**d). 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 (**?**).

(iv) ~~Extreme growth events.~~ **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~~ like the climate analysis 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~~ 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 CRUNCEP, the climate reconstructions used ~~to drive the data rely~~ for this study, relies 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~~trees were not juvenile anymore, detrending was not required. Subsequently, individual tree records from the same site were averaged to obtain a single time series per site (Fig. ~~8a~~**??**a). 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). 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 (**?**, 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~~extreme, the similarity between simulations and observations was calculated as the RMSE with the error being the distance from the 1:1 line (Fig. ~~8~~ **??** 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 (**??**) is the land-surface model of the IPSL (Institute Pierre Simon Laplace) Earth system model (**?**). 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, incoming solar radiation, and ~~CO2~~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 steps: 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 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 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 (**??**) and the explicit canopy representation of ORCHIDEE r4262 (**???**). It is one of the branches of the ORCHIDEE model and it was further developed from **?** and **?** (Text S1), parameterized, and tested to simulate TRW series, 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 data product for the climate forcing from a merged and homogenized gridded dataset developed for modelling purposes over the 20th century, i.e., CRU-NCEP (**?**), the gridded nitrogen deposition product from CCMI (**?**), and a gridded nitrogen fertilization product for N2O (**?**) such that observed TRWs for the past century can be used to evaluate the skill of the LSM ORCHIDEE r5698 to simulate radial tree growth. A detailed overview of earlier developments (**???**) that resulted in the emerging capability of ORCHIDEE r5698 to match the aggregate tree growth model (Fig. ~~3~~**??**) is given in the supplementary material (Text S1).

3.2 Reference data of productivity-oriented TRW sampling

The European biomass network contains TRW samples from "fixed-plot sampling". The database was established within multiple research projects and made publicly available through the EU Horizon-2020 project BACI (**?**). It archives at present 48 datasets covering temperate and semi boreal climates (Fig. ~~S1~~S2) and being collected from a variety of research efforts in Eurasia (**?**). 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 on stand density, to be archived in the BACI database (**?**). The European biomass network 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~~; (?)~~) may still be present. The records from the European biomass network are thus suited to evaluate the validity of using virtual trees constructed from ITRDB records to cope with 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 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~~S3.

For the simulations, the LSM ORCHIDEE r5698 was used. This model version accounts for the aforementioned minimum requirements for LSMs to mechanistically simulate TRW (section 2.1). ORCHIDEE r5698 was run for 11 individual pixels, each containing one of the selected sites. An observation-based time series of atmospheric ~~CO2~~ CO_2 concentrations was used (**?**) and forest management followed the reported management status of the site (Table ~~S2~~S3). During model development, three global (e.g., the number of diameter classes) as well as two PFT-specific parameters (e.g. recruitment success) were manually adjusted to jointly reproduce the TRW data of 10 ITRDB sites (aust112, cana106, chin037, finl055, fran4, id007, japa011, mo009, nepa003, spai055, and turk027) for which the match between observations and simulations was assessed visually. All runs in this study used the default parameter values except for the five parameters that were manually tuned.

In this study, 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 benchmark

The European biomass network data were used to verify, whether the big-tree selection bias that is present in the ITRDB data invalidates its use for benchmarking LSMs. The verification checked whether changes in parameter values or model process representation that would be required to make the model output better match the ITRDB data, would also result in a better match between the model output and the all tree data. If this would be the case, benchmarking LSMs against ITRDB data would result in model changes that would enhance the model's capability to simulate tree growth thus justifying the conclusion that despite the known biases, ITRDB data can be used for LSM benchmarking.

The verification, therefore, used the data from the European biomass network in two different ways: 1) all trees in the European biomass network dataset 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~~ 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 the top 15% of the trees based on their diameter, and 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. 9??): 1) The simulated TRW from the largest diameter class were transformed by modifiers (see below) to minimize the two metrics of each benchmark. The different benchmarks may use different metrics, i.e., the RMSE and slope of the residuals were used as the metrics for benchmarking size-related growth trend, growth of mature trees, and growth of young trees, whereas extreme growth and TRW amplitude were used as the metrics for benchmarking extreme growths (Table S1); 2) the same modifiers were then applied to all simulated diameter classes~~both metrics for~~ , and all four benchmarks. Hence, for each benchmark its two metrics were calculated using ~~the~~ all-tree data, and 3) the actual verification tested whether for a given metric and a given benchmark the modifier improved simulations for the big-tree sample ~~as well as~~ and for the all-tree data. Improvement of a specific metric of a benchmark was quantified by subtracting the ~~pre-modified~~ original value for that metric from its ~~post-modified value for the~~ modified value for all-tree data. A negative value thus indicated an improvement. If this was the case, the benchmarks of the big-tree and all-tree data were said to be consistent, implying that using this benchmark in combination with the ITRDB data would reveal the same model shortcomings as benchmarking ORCHIDEE against TRW data from all-tree networks. Across the 11 sites and for each of the four proposed benchmarks, sites where the test improved for both datasets were counted to estimate the confidence in using ITRDB in benchmarking LSMs.

Since this study aims to propose benchmarks making use of the ITRDB study rather than improving the ORCHIDEE model, the modifiers were applied to the model output directly. This approach has the advantage to remain conceptual by staying away from the need to optimize specific model parameters or rewrite or add processes in the model code. Different modifiers were used to accommodate the differences between the metrics: 1) the RMSE or amplitude of a benchmark was minimized by multiplying the simulated TRW with a modifier (Fig. 9??), 2) the slope of the residuals of a benchmark was minimized by subtracting a trend-modifier from the simulated growth trend and 3) the years of the simulated TRW were rearranged such that they match the ranked order of observed extreme TRWs.

4 Results

4.1 Verification of the concept behind benchmarking ITRDB

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 (Fig. ~~10~~??). Despite its simplicity, the use of modifiers 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 big-tree data. Applying the same modifiers to the all-tree data improved the match between the simulations and observations in 72% of the test cases (63 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 84% (27 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 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. ~~S2~~S3). 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~~ At 5 out of 11 sites, the all-tree data and the big-tree data yield different results. Two sites (HD2 and TIC) for which ~~the~~ for both metrics inconsistencies between the big-tree and all-tree data ~~results in different benchmarking results from the big-tree data~~were observed, 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~~ five size classes similar to the model. ~~On the other hand, ZOF had~~ The site labelled as ZOF has a bimodal size distribution ~~, which has~~ with the biggest number of trees in the 1st and 4th ~~bins~~ diameter class (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~~ two other 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. S3 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. ~~S3 b).~~ S4 b), despite the average simulation matched well with the average diameter trend as shown by the calculated slope of residuals: 0.08 and 0.09 (Fig. S4 a). 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~~ young trees 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. ~~S4~~S5). 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. ~~S5~~S6). For SOR the difficulties may have come from the model itself more specifically from difficulties with the carbon allocation (Fig. ~~S6~~S7). 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 470 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. ~~S7~~S8). 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. ~~S3 and~~ S4 and S5. This result suggests that the extreme growth benchmark is the least demanding benchmark in terms of the sampling-design.

5 Discussion

5.1 Benchmarking LSM against tree-ring width

The wealth of approaches available for modelling tree-ring growth has been largely overlooked by the global land-surface community. Until now, benchmarking LSMs against ring-width records still relies mostly on interannual variation in the simulated net primary productivity as a proxy for TRW (**????**). Although such an indirect approach is valid to certain extent to benchmark the capability of LSMs to simulate interannual variability, the observations will need to be detrended to remove the size-related growth signal, adding considerable uncertainty to the benchmark (**????**). Moving beyond the net primary production proxy by explicitly simulating stem radial growth and TRW enriches the benchmark since correcting for potentially confounding factors including climate responses, forest structure, age and size trends (**??**), as well as sampling biases (**?**)~~, can be performed~~. 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 model of aggregate tree growth (Fig. ~~3~~**??**). Given that tree-ring width is largely explained by phenology (**?**), tree size (**?**), climate sensitivity (**?**), and atmospheric ~~CO2~~ CO_2 concentrations (**?**), 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 ~~CO2~~ 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 sampling biases (**??**). 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 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 (**?**). 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~~S1):

(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 (**?**). Although this trend is removed in many dendrochronological studies to amplify the climate signal contained in TRW (**?**), 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 this 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 (**??**). 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 ~~CO2~~ 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 neither of endogenous nor of ~~either endogenous nor every~~ exogenous disturbance such as fire, pest and disease outbreak or ~~dead~~ death 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. Since the last benchmark for extreme growth events was not intended to test the capability of LSMs to simulate growth trends, the slope of residuals was not included. A skilled model is expected to simulate not only the timing of extreme growth but also the magnitude of it, the benchmark therefore was designed to evaluated both using RMSE.

5.2 Verification of the benchmarks

Even when the ITRDB data with a big-tree bias are used, the model improvements and parameter adjustments should improve the growth of all simulated trees, not only the big-trees contained in the ITRDB data. Our verification approach estimated the level of confidence for each benchmark. This level confidence quantified the number of cases where improving the model or adjusting its parameters based on benchmarking against big-trees would result in improving the model performance for all trees.

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

If the ITRDB is to be used, the number of benchmarks and the circumstances that can be used with confidence becomes more limited. The verification results (Table ~~2~~??) show that if the output of 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. This level of confidence, i.e., 70%, is sufficient to support benchmarking a LSM against tens to hundreds of ITRDB sites at the same time. The same level of confidence is likely too low to benchmark a LSM (or an ecosystem model) against a single ITRDB data series as there is a 30% chance that parameter tuning or model improvements following the benchmarking will not improve the model. Given the spatial extent of European biomass network, the levels of confidence represent temperate and hemi-boreal forests (Fig. ~~S1~~S2) and the validity of the use of the ITRDB data from boreal and tropical forest will need to be verified when suitable data become available.

~~Higher chances for~~ 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, implying that benchmarking against young trees will benefit most from using data free from the big-tree selection bias. Higher chances to improve ORCHIDEE when beech (87%) compared to spruce (64%) as a benchmark suggest that the validity of the assumptions underlying the use of ITRDB data ~~, partly depend~~ partly depends on tree species. In this study, the variety in species was too limited to generalize

this ~~results~~ result 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~~. S3 to S5). 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. ~~S8~~ S9 a), the simulation could be improved for all metrics of all four benchmarks despite the difference in simulated and observed stand structure (Fig. ~~S8~~ S9 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 2-5~~S3 to S6). 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.

6 Outlook

Tree-ring records could complement well-established but short-term benchmarks for LSMs (**?**), such as forest inventory data (**??**), ~~FLUXNET sites~~ eddy covariance measurements (**??**), Free Air ~~CO2~~ CO_2 Enrichment experiments (**?**) and satellite observations of vegetation activity (**??**). 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 (**??**) and (iii) combining their use with high-frequency but short- term eddy covariance measurements (**??**), experimental data from plant growth under pre-industrial ~~CO2~~ CO_2 concentrations (**?**), and proxies of atmospheric composition (**?**).

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 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 BACI data. JJ and SL wrote the manuscript, all authors contributed to revising and editing the different versions of the manuscript.

9 Competing interests

The authors declare that they have no conflict of interest.

Acknowledgements. JJ, PP, MJM and SL were funded by the VERIFY project under the European Union's Horizon 2020 research and innovation ~~programme~~ program 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~~ program 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. SL would like to thank Antonio Lara (Universidad Austral de Chile) for early discussions on the topic. Valérie Daux (Laboratoire des Sciences du Climat et de l'Environnement) is acknowledged for commenting on a previous version of the manuscript which improved its clarity.

~~Benchmark Metrics Targeted process understanding Solutions for meaningful model comparison with ITRDB Figure~~

~~Size dependent growth RMSE Slope of the residuals~~

~~Long-term size-related growth Within-stand competition~~

~~Select the biggest tree of the simulation Construct an average virtual tree aligned by tree age Fig. 5~~

~~Diameter increment of mature trees RMSE Slope of the residuals~~

~~Long-term tree growth after establishment Within-stand competition~~

~~Select the biggest tree of the simulation Construct an average virtual tree aligned by calendar year Fig. 6~~

~~Diameter increment of young trees RMSE Slope of the residuals~~

~~Short-term (i.e. 30-year) tree growth during establishment Size-related growth~~

~~Select the biggest tree of the simulation Construct a fast-growing virtual tree Fig.7~~

~~Extreme growth Extreme events Amplitude~~

~~Yearly climate sensitivity~~

~~Select the biggest tree of the simulation Define extreme growth using 25% smallest and 75% largest observations Fig.8~~

11.1

Table 1. 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 consistent with the optimisation result of the big-tree data.

Benchmark	**Size dependent growth**		**Diameter increment of mature trees**		**Diameter increment of young trees**		**Extreme growth**	
Metrics	**RMSE (mm)**	**Slope of residuals (mm/yr)**	**RMSE (mm)**	**Slope of residuals (mm/yr)**	**RMSE (mm)**	**Slope of residuals (mm/yr)**	**Ampli-tude (mm)**	**Extreme growth (scaled)**
Picea abies	**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)	**DEO (-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 (-0.0008)	**SOB (-0.001)**	**SOB (-20.22)**	SOB (0.50)	SOB (9.32)	SOB (1.29)	**SOB (-0.08)**	**SOB (-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)**

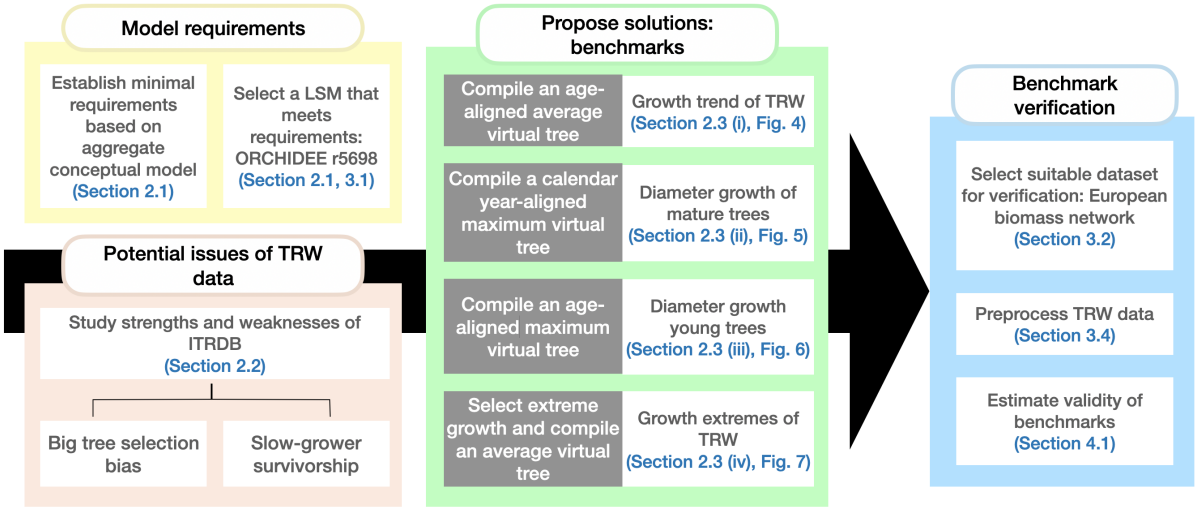

Figure 1. Workflow of ~~the~~ this study and the dependencies between the different sections of the manuscript. Each colour represents a different aim of the study. The arrow shows ~~how~~ that the outcomes of the first three aims ~~are~~ have to be combined to ~~achieve~~ verity the ~~final aim, i.e., verify the~~ proposed benchmarks.

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

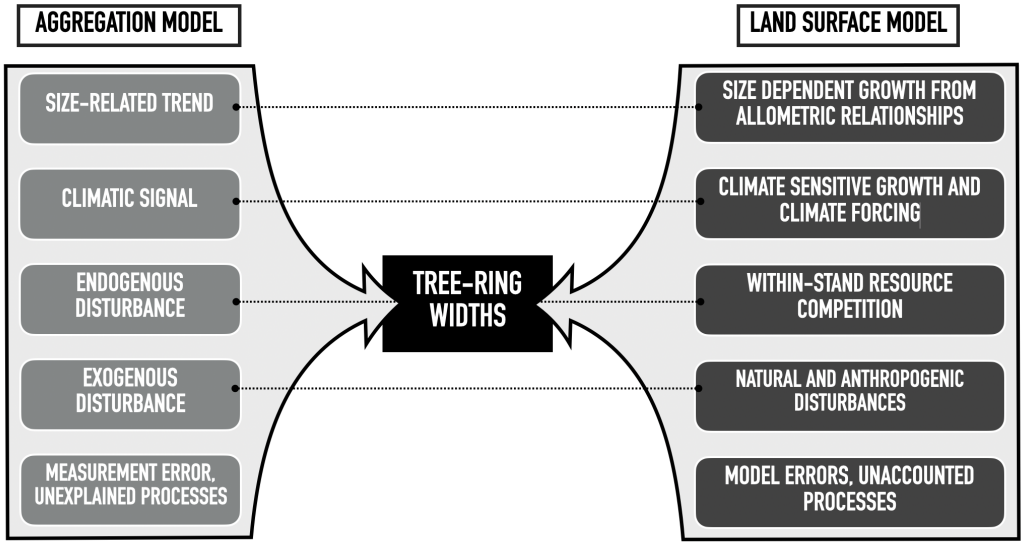

Figure 2. Main drivers of the linear aggregate conceptual model of tree-ring growth and the equivalent processes in land-surface models. The dotted lines connect the related components. Note that both the aggregation and the land-surface model come with errors, uncertainties and unaccounted processes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

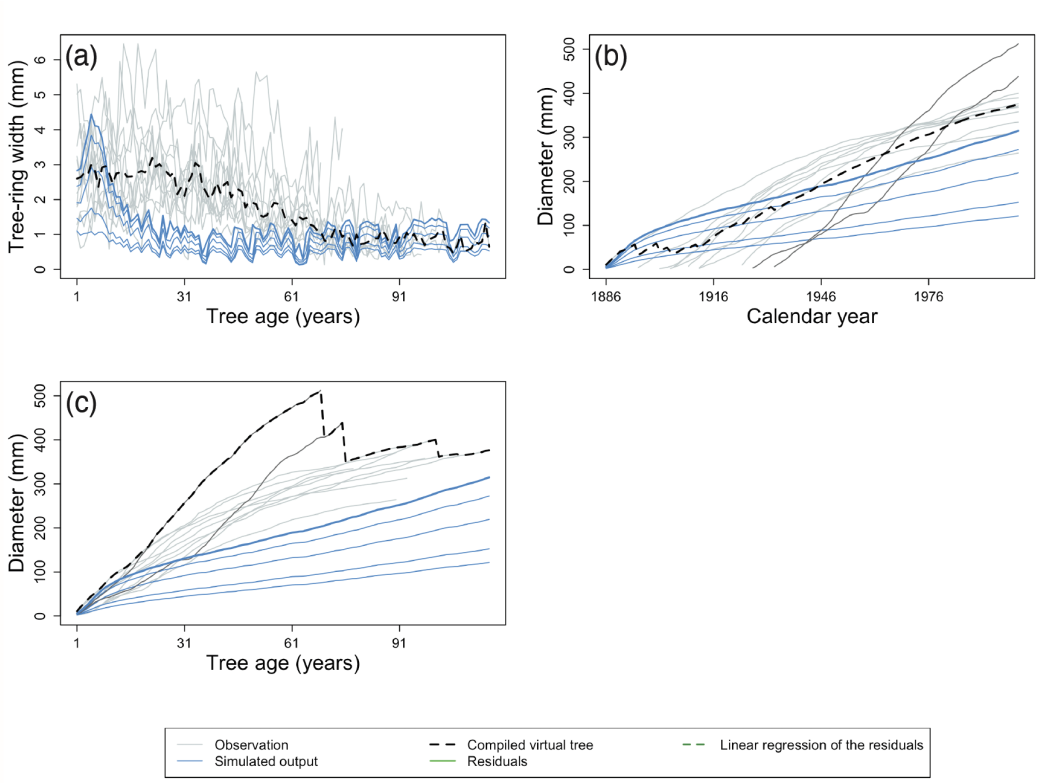

Figure 3. ~~**Solutions and virtual trees to account for challenges to use ITRDB datasets when evaluating land-surface models against ITRDB tree-ring data.**~~ **Example of the solutions including a virtual tree, to account for challenges to use ITRDB datasets when evaluating land-surface models against ITRDB tree-ring data.** (a) Data-model comparison may overcome the big-tree selection bias by comparing only the simulated biggest diameter class (bold blue line) for evaluation rather than all diameter classes (thin blue lines), with the compiled average 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~~ Using all trees without further consideration ~~this~~ in the calculation of the average virtual tree (black dotted line) would lead to underestimating tree growth at the time of stand establishment ~~and thus result~~ resulting in a flawed ~~test when compared against~~ comparison with the simulated tree growth (blue ~~line~~lines)~~as the virtual tree (black dotted line) is much smaller during stand establishment~~. (c) ~~Aligning~~ However, aligning observations by the age of individual trees before compiling maximum virtual tree (black dotted line) results in a very different virtual tree compared to Fig 3b. The maximum virtual tree 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~~ pine forest in Germany archived as germ214 (**?**) (Table ~~S2~~S3), and the model outcome is the simulation for observed sites from ORCHIDEE r5698.

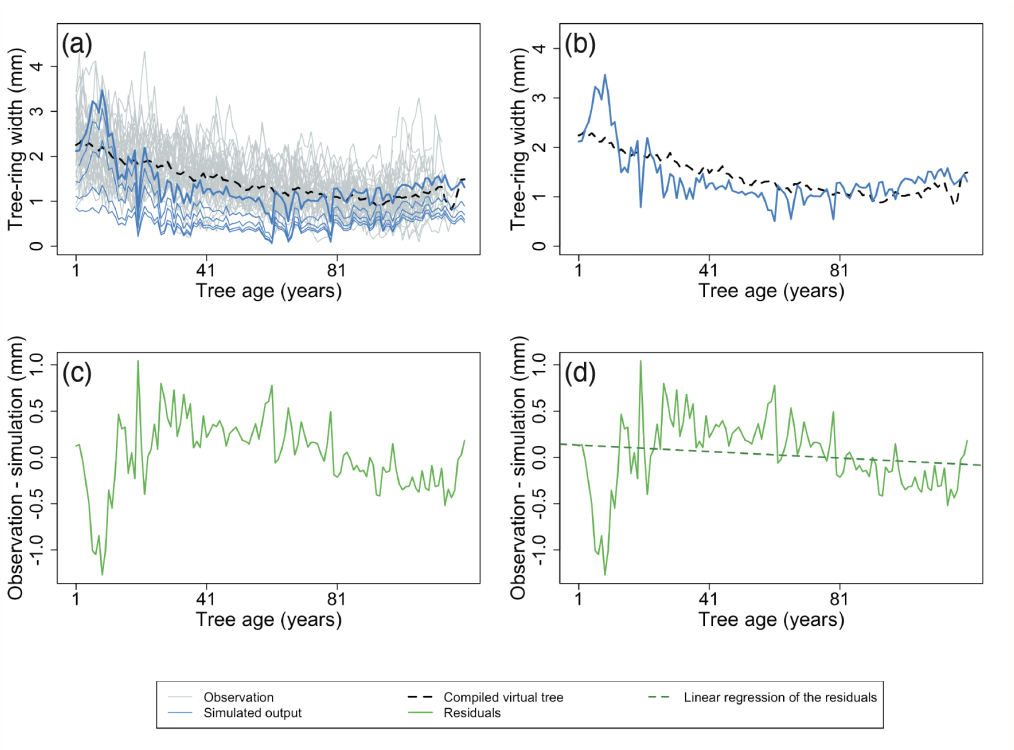

Figure 4. Example of the major steps for calculating the metrics of the benchmark for the size-related trend in diameter increment. The size-related trend in diameter increment can be assessed by calculating a time series for the average ring width after aligning the age of the individual trees (a, b). 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; RMSE between blue line and black dotted line) and normalized by the length of time series to calculate the difference in observed and simulated growth trends. Observations and simulation are from *Pinus sylvestris* site in Finland archived as finl052 (**?**). For this example, calculated RMSE (b) is 0.39 (mm), and the slope of residuals (d) is -0.002 (mm/yr).

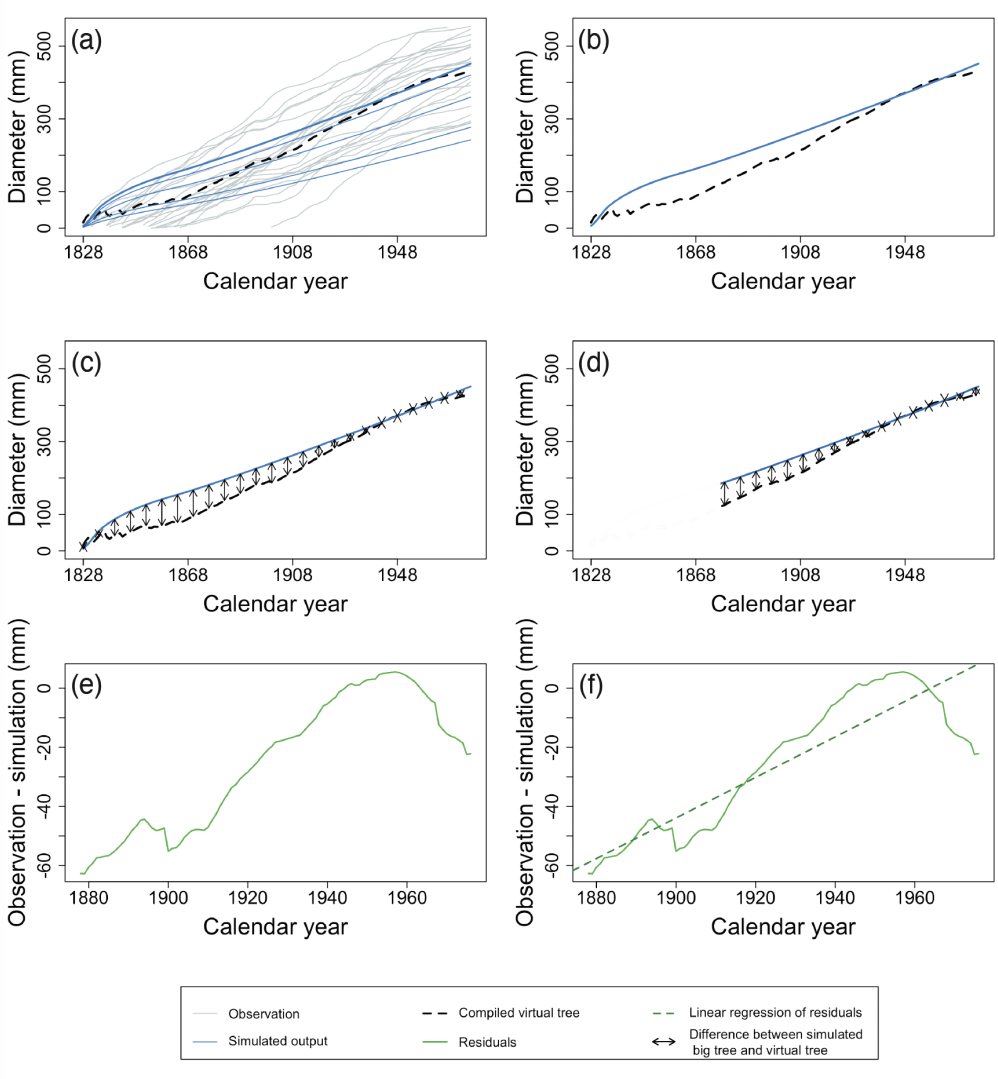

Figure 5. Example of the major steps in calculating the metrics of the benchmark for the diameter increment in mature trees. Individual tree records are ordered by calendar year and for each year the average observed diameter is calculated (a). Observations are ~~685~~ shown as grey lines and simulation as blue lines. The biggest class is presented by the bold blue line. Black dotted lines represent the yearly average of observations. Note that X-axis in Fig. ~~6~~ **??** is different from Fig. ~~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 (ten decades, in this example) of the virtual tree growth are compared to the simulations (d). The RMSE (~~grey~~ d; black arrows) and trend 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. Observations and simulation are from *Pinus sylvestris* site in Scotland archived as brit021 (**?**). In this case, RMSE (d) and the slope of residuals (f) were calculated as 33.65 (mm) and 0.68 (mm/yr), respectively.

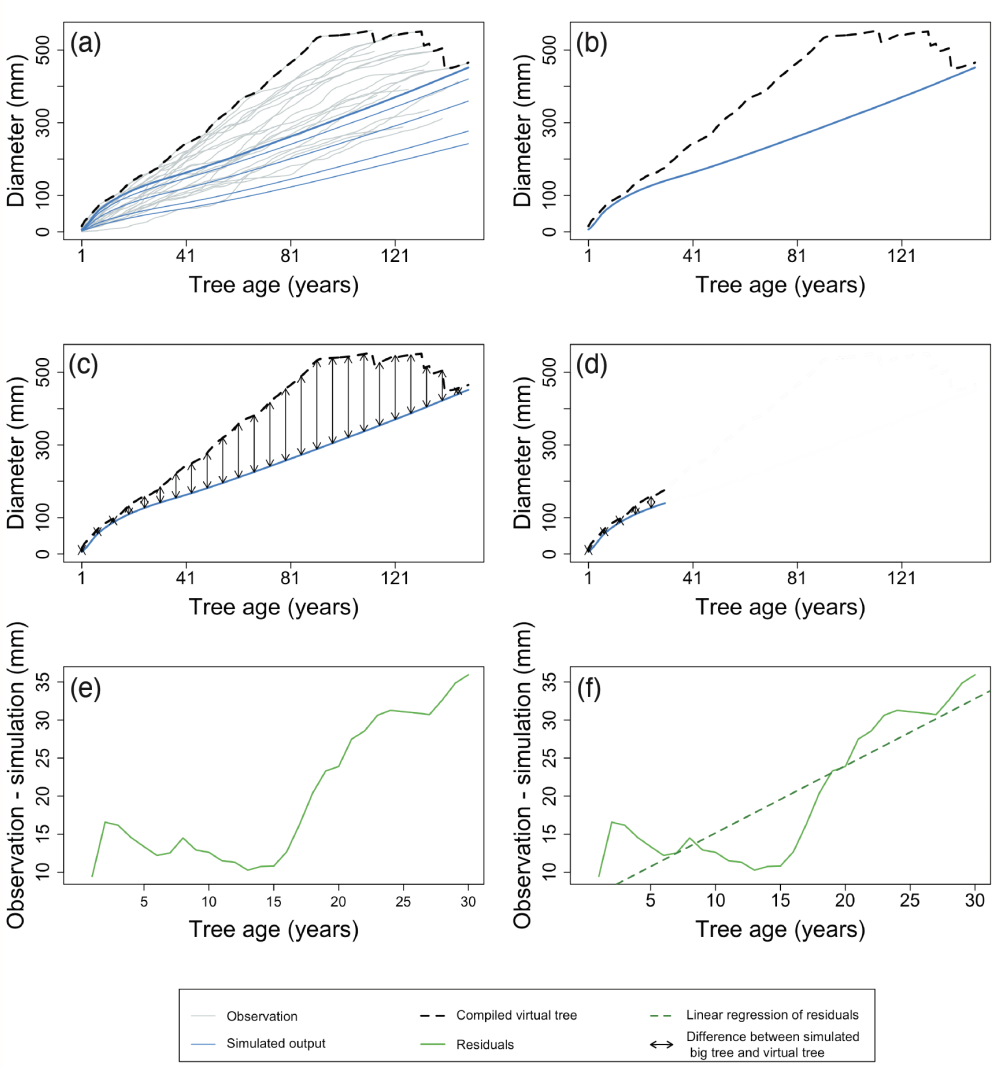

Figure 6. Example of the major steps in calculating the metrics of the benchmark for diameter increment in young trees. After aligning the TRW records of the individual trees by their age, a virtual tree is constructed by taking the maximum observed diameter of all trees for each year (a). Observations 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. 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 *Pinus sylvestris* site in Scotland archived as brit021 (**?**). The calculated RMSE (d) was 21.86 (mm) and the slope of residuals (f) was 0.88 (mm/yr) for this example.

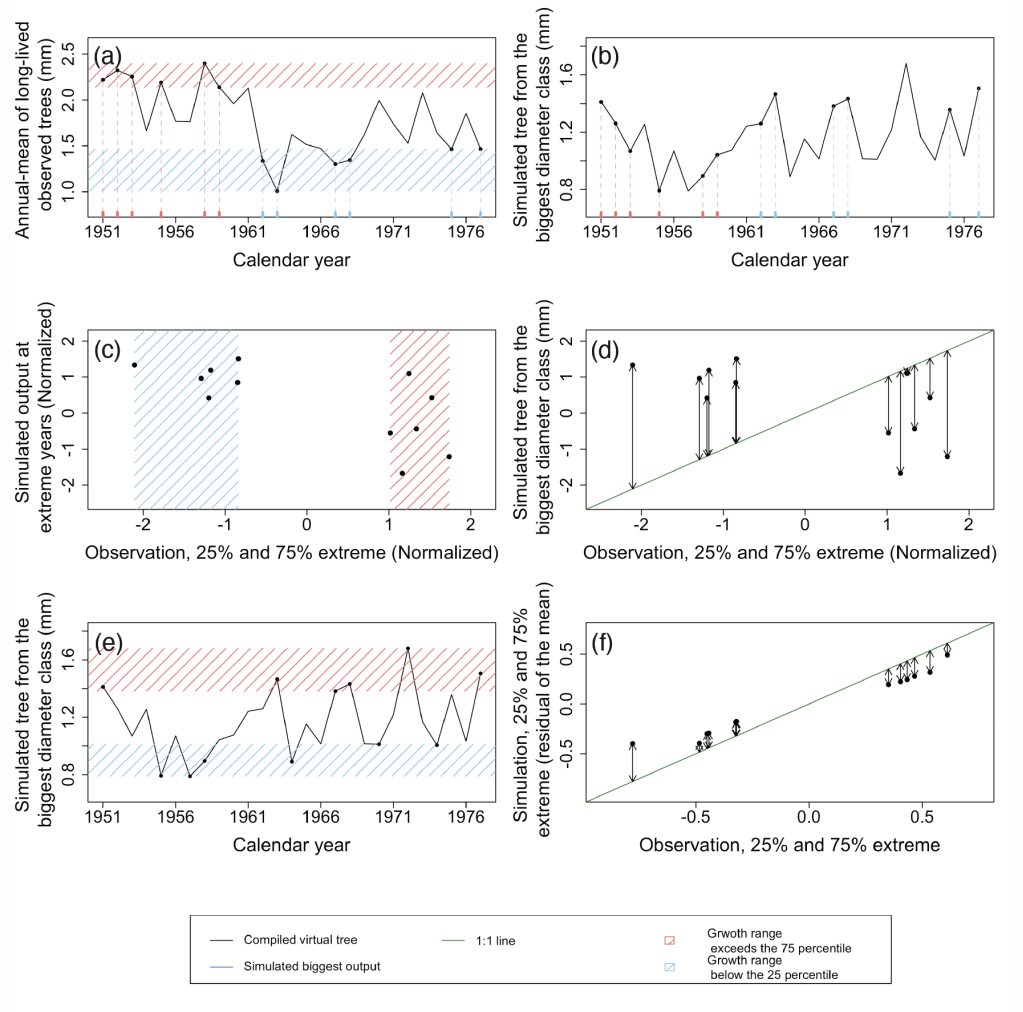

Figure 7. Example of the major steps in calculating the metrics of the benchmark for extreme growth events. In this benchmark, extreme growth is defined as the first and last quartiles in TRW ordered by calendar year and averaged over the individual trees records (a). Red shaded area and ticks represent 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 assess the absolute amplitude. Possible uncertainties from using reconstructed climate forcing, were avoided by limiting the calculations of both metrics to the past five decades for which climate observations are available. Observations and simulation are from *Pinus sylvestris* site in Spain archived as spai006 (**?**). In this test case, RMSE for extreme years (d) was 0.57 (mm) and RMSE for extreme growth (amplitude; f) was 0.03 (scaled).

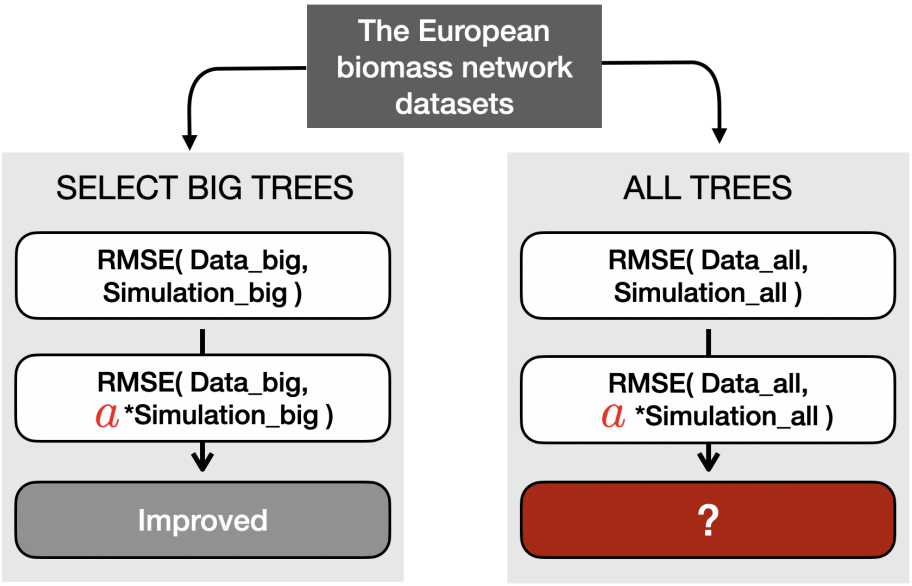

Figure 8. Schematic representation of the verification process for the RMSE-metric. Before the verification, two types of datasets were prepared: big-tree data (limited to the 15% biggest trees) and all-tree data. In this example, the simulated TRW was multiplied with a verified modifier such that it minimized the RMSE between the simulated and observed TRW for the 15% biggest trees (see section 3.4 for details). 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 modifiers which are detailed in section 3.4 follow a similar approach.

~~**Details of the four benchmarks for four out of the 11 sites selected from the European biomass network.** Each column denotes a single site. The DEO and DVN represent Norway spruce forests. The CAN and SOR sites are Beech forest. Each row denotes a different benchmark. The first row corresponds to the benchmark explained in Fig. 5 (d), the second row to Fig. 6 (f), the third row to Fig. 7(f), and the fourth and fifth rows to Fig. 8 (d) and (f), respectively.~~

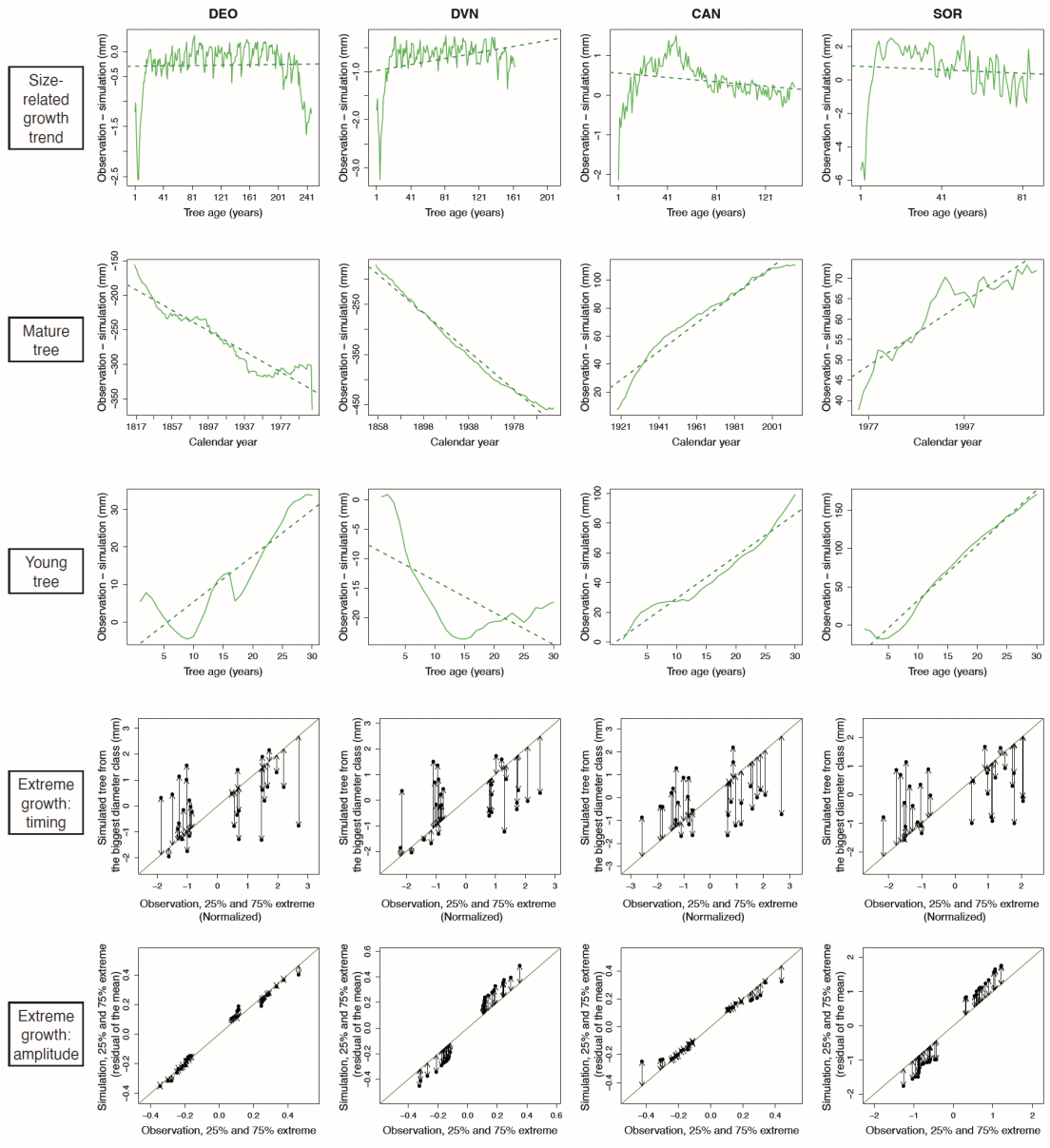

Figure 9. **Details of the four benchmarks for four out of the 11 sites selected from the European biomass network.** Each column denotes a single site. The DEO and DVN represent Norway spruce forests. The CAN and SOR sites are Beech forest. Each row denotes a different benchmark. The first row corresponds to the benchmark explained in Fig. **??** d, the second row to Fig. **??** f, the third row to Fig. **??** f, and the fourth and fifth rows to Fig. **??** d and f, respectively.

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