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.
[†]Deceased 29 January, 2021.

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 modeling community, as tree-ring data 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 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 We propose and evaluate an improved conceptual framework of when and how ITRDB data can, despite its tree-ring data may, despite their sampling biases, be used as century-long benchmarks for LSMs. Combining 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 independentEuropean tree-ring network of biomass plots that were sampled in a locally representative way and are thus not biased by hindcasting targets for evaluating LSMs. Four complementary benchmarks – size-related diameter growth, diameter increment of mature trees, diameter increment of young trees, and the response of tree growth to extreme events – were simulated using the LSM ORCHIDEE version r5698, and verified against observations from 11 sites in the independent, unbiased European biomass network datasets. The potential for big-tree tree selection , and (2) following sub-sampling of this selection bias in the International Tree-Ring Data Bank (ITRDB) was investigated by subsampling the 11 sites from European biomass network by only considering the 15% biggest trees. This study showed that the ITRDB data can be used with . We find

that in about 95% ~~confidence to benchmark annual radial growth during extreme climate years. In about 70%~~ of the test cases,

20 using ITRDB data would result in the same conclusions as using the European biomass network when the ~~model~~ LSM is bench-marked against the annual radial growth during extreme climate years. The ITRDB data can be used with 70% confidence when 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,~~

25 ~~they may advance the field of land surface modelling by providing a much-needed large-scale constraint on changes in the simulated~~ The proposed maximum tree diameter and annual growth increment ~~for~~ benchmarks may enable use of ITRDB data for large-scale validation of LSM-simulated response of forest ecosystems to 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~~ The results also suggest ways in which tree-ring width observations may be col-

30 lected and ~~processed~~ /or reprocessed 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 (IPCC, 2013) and as such have made a tremendous impact on global environmental policy (Paris Agreements, 2015). 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 (McGuffie A. and Henderson-Sellers, 2005). Although long-term changes that date 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~~ substantially to uncertainties in simulated future global carbon stocks in soil and vegetation (Friedlingstein et al., 2006, 2014) 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 (Fritts, 2012; Hemming et al., 2001). 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) (Babst et al., 2014a, b, 2017, 2018; Zuidema et al., 2018).

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 (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 (Fig. ~~S1~~S2) (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 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. 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 (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~~ a parameterization of the influences of climate on cambial activity (Vaganov et al., 2006). These models were capable of repro-

70 ducing 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).

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

75 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, 2004; Sato et al., 2007). Hence, a variety of approaches is now available to describe TRW growth in forest models,

80 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

85 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 (BACI, 2020) that is not prone to sampling biases related to ~~palaeoclimatological research. Dependencies~~ dendroclimatic research. The organization of this paper follows these aims, and dependencies between these aims and the workflow of this study are detailed in Fig. ~~2.~~ 9.

90 ## 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 (Cook and Kairiukstis, 1990) considers that the observed TRW at year t (in mm) consists of five additive growth contributions (Fig. ~~3)~~ 2, left column) and as such provides a framework for simulating tree-ring widths with (semi)mechanistic model approaches (Fig. 2):

95   (i)   Size-dependent growth is the dominant signal in raw tree-ring measurements (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~~ 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

100   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 (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 account for ~~size- dependent~~ size-dependent growth and stand self-thinning (Franklin et al., 2012; Wolf et al., 2011).

  (ii)   Climate-dependent growth reflects the sensitivity of tree growth to radiation, temperature, phenology, and water avail-

105   ability (Fritts, 2012) 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 (Farquhar, 1989), the McCree - de Wit - Penning de Vries - Thornley approach for the temperature dependence of respiration (Amthor, 2000)~~, and~~. They 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~~ represented through either simple transfer functions or more recently

110   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 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~~ Section 2.2 and 2.3 below).

115   (iv)   Chronic exogenous disturbances such as increasing atmospheric ~~CO2~~ CO_2 concentration (LaMarche et al., 1984) and N-deposition (Magnani et al., 2007) 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 (Vuichard et al., 2019; Zaehle and Friend, 2010).

~~(iv)~~ Although abrupt disturbances such as fires, pests and storms are increasingly being simulated by LSMs (Chen et al., 2018;

120   Yue et al., 2014) and ~~they~~ leave marks in TRW (Bowman et al., 2009; Bräuning et al., 2016), 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

125   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 (Briffa and Melvin, 2011; Cook and Kairiukstis, 1990)~~, 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~~ Alternative approaches based on Liebig's Law of the Minimum (Stine, 2019) have been proposed to attribute TRW to its major drivers~~(Stine, 2019)~~. 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~~ dependent growth and climate sensitivity have ~~a much larger contribution to TRW than the other processes (Hughes et al., 2011).~~ been observed to comprise the primary contributions of variance in the linear aggregate model (Hughes et al., 2011).

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 (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 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 (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 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~~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 ~~170~~ stand are at present only accounted for in a few LSMs, for example, CLM (ED) (Fisher et al., 2015), ORCHIDEE (Naudts et al., 2015), and LPJ-GUESS (Smith, 2001). 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~~3). If a core reaches the ~~centre~~ pith 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~~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 (Briffa and Melvin, 2011). 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 (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.4a).~~

~~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~~ Tree-ring datasets often contain cores of individuals from different cohorts. Slow and fast-growing trees within the same cohort (~~Fig~~illustrated by the grey lines in Fig. 3) is another source of bias (Melvin, 2004; Zuidema et al., 2011). ~~4b) .~~ Slow-growing trees tend to live longer than fast-growing trees in the same cohort (Mencuccini et al., 2005; Schulman, 1954). ~~Records of TRW~~ Owing to survivorship being biased towards slow-growing individuals (Bowman et al., 2013), 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 (Brienen et al., 2012). ~~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). Comparing these data against simulations requires the model to be individual-based or to align TRW records by age (Fig. 4a).~~

~~Given the above, only part of the~~

2.3 Solutions for the challenges and virtual trees

Despite its known biases, poorly described sampling protocols, or protocols that were not rigorously enforced, information contained in TRW records ~~can~~ of the ITRDB can still be used for benchmarking ~~if their sampling protocol is poorly described or not rigorously enforced. A~~ LSMs.

190 Without additional data, model-data comparison cannot correct for ~~these biases~~ the big-tree selection bias in the ITRDB, however, this bias can be accommodated through models that simulate multiple tree diameter classes by comparing the largest simulated diameter class with the observed ITRDB tree-ring records (illustrated by the black dotted and highest blue lines in Fig. 4 a)

Likewise, model-data comparison cannot correct biases from slow-grower survivorship, but we propose to enhance the con-

195 sistency between modelled and observed ~~tree-rings for a stand under study~~ TRWs by making use of site-specific virtual trees. ~~'Virture tree' is "not physically present as such but made by software to appear to be so from the point of view of a program or user" (OED online) !!!!!!!~~

~~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~~ Virtual trees are created from observations by combining data

200 from different individuals to obtain a time series of TRW with the desired property (details on the desired properties are given below) that reflects stand-level characteristics. By definition, virtual trees are not observed as data sequences in the ITRDB observations, but rather are extracted from site level data. Since for a single site, the tree-ring observations consist of samples from multiple individual trees (individual trees are shown by grey lines in Fig. 4), constructing a single virtual tree for a given site facilitates data-model comparisons. Because virtual trees are dependent on the chosen desired property, multiple

205 data-model comparisons are possible for each site. In section 2.4 below, 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~~ three virtual tree target properties:

– **Tree-age-aligned average** virtual tree:

~~(i)~~ The average virtual tree of a stand aligned by tree age is calculated as the time series for the average ring width after

210 aligning the age of the individual trees (Fig. ~~4a~~ 4 a). Age-aligned TRWs 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~~ size-related growth is the dominant driver of tree growth (see ~~subsection 2.3~~ section 2.1; subsection 2.4 (i)).

~~(ii)~~

– **Calendar-year-aligned average virtual tree:** The average virtual tree of a stand aligned by calendar year is calculated by ordering individual tree-ring series by calendar year (Fig. 4 b) 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 better cope with a difference in simulated and observed forest structure by compiling a representative and comparable tree with the simulated tree (see subsection 2.4 (ii)).

– **Tree-age-aligned largest virtual tree:** The largest virtual tree of a stand is calculated after aligning individual trees by their age (Fig. 4 c). The recommendation to remove the age trend from tree-ring records (Cook et al., 1995) 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 that has the maximum observed diameter of all trees for a given tree age (note the difference in the weight of dark grey lines to virtual trees in Fig.4 b and c). The virtual fast-growing tree thus gives a better idea of the true mean tree growth in old stands. (see subsection 2.4 (iii)).

The proposed data-model comparison largely relies on the concept of virtual trees, since these virtual trees can account better for known sampling biases of the ITRDB datasets and different aspects of TRW, and facilitate comparing simulations and observations at the stand level. The proposed definitions and uses of virtual trees which were partly customized to ORCHIDEE r5698, are evaluated in section 4. Except for LSMs with an individual tree-based stand definition (Sato et al., 2007), 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 the LSM under evaluation.

2.4 Benchmarks for comparing observed and simulated tree-ring widths

If a LSM explicitly accounts for the main contributors 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 growth 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:

(i)   **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 (examples shown in Fig. 4a and Fig. 5 a and b) and subtracting this virtual tree from the simulated TRWs of the largest diameter class (Fig. 5 c).

245 Subsequently, ~~a~~ linear regression is used to quantify the temporal trend in the residuals (~~Fig. 5d~~examples shown in 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~~ 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~~.

250 (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 (Bellassen et al., 2010; Haverd et al., 2013). 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. 4b). This discrepancy between simulated and observed competition can be accounted for in the benchmark by using the observations to compile a virtual

255 tree of the stand aligned by calendar year, taking the average tree diameter of all samples to construct the virtual tree (Fig. ~~6~~ 4b and Fig. 6 a and b). ~~Under the assumption~~ 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~~ largest 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

260 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~~ therefore, the first five decades were excluded for the better comparison (Fig. ~~6~~ 6 c and d). The 50-year threshold in this study is somewhat arbitrary but reflects the observation that in most of the selected time series, the fastest changes in tree growth occur in the first few decades. 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. The RMSE and trend of

265 the residuals between the virtual tree and the largest diameter class simulated are calculated (Fig. ~~6d~~ 6 d-f). 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

270 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~~ in-field sampling due to the fact that the old fast-growing trees died well before sampling took place (Fig. ~~7~~ 4 c and Fig. 7 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. ~~7~~ 7 c and d) by calculating the RMSE and trend of the residuals (Fig. ~~7d). The 30-year~~ 7 d-f). As for the previous benchmark, the threshold is somewhat

275 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~~ and was set to focus the analysis on the first decades in which diameter growth is generally faster compared to later decades. Note that the thresholds for young and mature trees are separated by 20 years during which the observations are not considered in either benchmark. When needed, these thresholds could be

10

adjusted to better ~~fit the observed growthdynamics for other tree species and/or other regions~~match local tree growth, but it is recommended to keep the separation between the thresholds to account for the transition in diameter growth from young to mature trees. 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~~ **Extreme growth events.** For this benchmark, extreme growth is ~~defined~~ here defined coarsely 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~~. For the purpose of estimating extremes, we seek the average virtual trees and LSM simulations to most accurately and precisely characterize the interannual variability. In the present study, use of the 1951-2000 interval produces LSM simulations forced by climate and nutrient loading estimates best constrained by dense, recent climate observations (see Section 3), and therefore minimizing the contribution of forcing uncertainty to LSM simulation uncertainty; in the TRW observations, it also leverages non-juvenile tree-ring series of more than 50 years ~~or older in 1950 and TRW were thus past their juvenile dynamic growth phases, detrending was not required~~which do not require detrending and are expected to most reflect the climate sensitivity. Subsequently, for each site, individual tree records ~~from the same site were~~ are averaged to obtain a single time series ~~per site (Fig. 8a). The model skill to reproduce the absolute ring-width amplitude regardless of timing was tested~~ 8a). Model skill for estimating the distribution of growth arising from climate variability is evaluated 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~~ 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~~ is evaluated by comparing simulated to observed virtual standardized 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~~ of the distance from the 1:1 line (Fig. ~~8~~ 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 (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, 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

345 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~~ (Text S1), used in this study, combines the dynamic nitrogen cycle of ORCHIDEE r4999 (Vuichard et al., 2019; Zaehle and

350 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, in order to meet the aforementioned minimum requirements of simulating the~~ simulates carbon, water, energy, and nitrogen ~~cycle, while accounting~~ fluxes, allows for size-dependent allocation ~~for~~ across three diameter classes within a forest stand~~.~~

355 ~~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 (Viovy, 2016), the gridded nitrogen deposition product from CCMI (Eyring et al., 2013), and a gridded nitrogen fertilization product for N2O (Lu and Tian, 2017) such that observed TRWs for the past century can be used to evaluate the skill of the LSM ORCHIDEE r5698 to simulate~~, and has been parameterized and tested for the simulation of TRW series via radial tree growth estimation. A detailed overview of earlier developments (Krinner

360 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. ~~3~~2) is given in the supplementary material (Text S1).

3.2 Simulations: forcing and model parameterization

In this study, offline ORCHIDEE simulations for the 20th century are forced with the merged and homogenized gridded CRU-NCEP climate dataset (Viovy, 2016); the gridded nitrogen deposition product from the Chemistry-Climate Model Initiative

365 (CCMI) (Eyring et al., 2013); a gridded nitrogen fertilization product for N_2O (Lu and Tian, 2017); an observation-based time series of global atmospheric CO_2 concentrations (Keeling et al., 1996); forest management followed the reported management status for each of 11 sites simulated (Table S3) for comparison, via the virtual tree benchmarks, with observations (Section 3.3). Simulations were started from a 300-year long spinup, which was required for equilibrium with respect to the slow carbon and nitrogen pools in the soil. The spinup was concluded with a forest clear cut, such that the start year and the length of each

370 simulation matched the observed stand age for the validation dataset (Section 3.3).

During model development, two global (number of diameter classes and ratio for number of trees per each diameter class; see 3.1) and six PFT-specific parameters (f_{power}, f_σ, k_{α_r}, k_{α_s}, k_{β_r}, and k_{β_s}; Table S1) were manually adjusted to jointly reproduce the TRW data of 11 ITRDB sites: aust112 (Strumia, 2005), cana106 (Archambault and Bergeron, 2002), chin037

13

(2k Consortium, 2013), finl055 (Melvin, 2005), fran4 (Tessier, 1996), id007 (Briffa and Schweingruber, 2002), japa011 (Davi et al., 2011)
375 , mo009 (Bell et al., 2002), nepa003 (Krusic and Cook, 2005), spai055 (Briongos and Cerro-Barja, 2007), and turk027 (Griggs et al., 2006). Sites belonging to the same PFT were thus simulated by making use of a single PFT-specific parameter set. In other words, no site-specific parameters were applied. All other parameters were set to default values. As configured, the model distinguished five diameter classes for simulated trees. The smallest and largest diameter classes each contained 15% of the total number of simulated trees. The three intermediate classes contained 21%, 27% and 21% of simulated trees respectively. Agreement
380 between simulated and observed TRW was assessed visually.

3.3   Reference data of productivity-oriented TRW samplingVerification

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 (BACI, 2020). It archives at present 48 datasets covering temperate and semi boreal climates (Fig. S1S2) and being collected from a variety of research
385 efforts in Eurasia (Klesse et al., 2018). 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 (Babst et al., 2014b). 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; (Bowman et al., 2013)) 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
390 cope with the aforementioned sampling biases.

3.4   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 2.32.4). The benchmarks were applied to a
395 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 S2S3.

400 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 concentrations was used (Keeling et al., 1996) and forest management followed the reported management status of the site (Table S2). During model

3.4 ~~Verification of the benchmark~~

The European biomass network data were additionally 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 from the European biomass network. 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~~ The "big tree" verification required three additional steps (Fig. ~~9~~9): 1) The simulated TRW from the largest diameter class were transformed by modifiers ~~(see below)~~ to minimize the two metrics of each benchmark ~~; 2)the~~ (Section 2.4; Table S2). 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 S2). 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

15

modifier improved simulations for the big-tree sample and for the all-tree data. Improvement of a specific metric of a benchmark was quantified by subtracting the original value for that metric from its modified value for all-tree data. A negative value thus indicated an improvement. If all three conditions were satisfied, 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 and both benchmark metrics, 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 of remaining conceptual, avoiding 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 matched the ranked order of observed extreme TRWs.

4 Results

4.1 Verification of the concept behind benchmarking ITRDB

The verification

Verification of ORCHIDEE-based TRW simulations was applied at the 11 sites selected from European biomass network (Section 3.3), via estimation of the four benchmarks (Section 2.4) from simulations and observations, their evaluation via the two skill metrics per benchmark, for a total of 88 test case comparisons (Table 1, Fig. S3). We describe the results of big-tree bias estimation and modifier estimation in general, and then for each of the individual benchmarks in the remainder of this section.

4.1 Big-tree bias estimation

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 verifying 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~~ 1). We note that this overall result 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)

465 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% (~~

4.2 Verification by benchmark

when benchmarking size-related growth, conclusions would be similar in 16 out of 22 ~~) confidence for benchmarking LSMs.~~
470 ~~Given the reasoning underlying the verification, this suggests that for 72% of the cases the conclusions would be similar~~ ~~irrespective of~~ cases (73%), regardless 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~~ positive differences close to zero, suggesting that simulations with ORCHIDEE r5698 matched the observed size-related growth trend reasonably well, leaving limited room for improvements. One site ~~, SCH~~(SCH) showed a positive difference because it contained two slowly growing trees which lived
475 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~~S4). Except for this site, the other sites showed marginal inconsistencies or showed improved simulated output against the two datasets. The size-related trend in tree growth, thus, 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)
480 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%
485 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).~~ S5 b), despite the average simulation matched well with the average diameter trend as shown by the
490 calculated slope of residuals: 0.08 and 0.09 (Fig. S5 a). These results suggest that the mature trees benchmark is sensitive to the stand structure.

17

With 50% (11 out of 22) confidence in using the big-tree data in benchmarking LSMs, ~~this~~ the 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~~ similarity between simulations and observations ~~being similar~~, with RMSE around 10 mm ~~and;~~, and (2) the difference between big trees and small trees ~~growths~~ growth being larger in the observations (Fig. ~~S4~~S6). 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~~S7). For SOR the difficulties may have come from the model itself ~~more specifically difficulties with~~, more specifically from difficulties in simulating the carbon allocation (Fig. ~~S6~~S8). These results suggest that a variety of issues decreases the confidence in using big-tree data for young tree 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~~S9). 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~~S5 and S6. This result suggests that the extreme growth benchmark is the least demanding benchmark in terms of ~~the sampling-design~~sampling design.

5 Discussion

5.1 ~~Benchmarking~~ LSM ~~against~~ verification: beyond 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~~ verification of LSMs against tree ring width records has relied primarily on interannual variation in ~~the~~ simulated net primary productivity as a proxy for ~~TRW~~ site-level TRW chronologies (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~~ appropriate, to a certain extent, for validating the capability of LSMs to simulate interannual variability, for studying patterns and mechanism of change over longer timescales, 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~~ verification process (Bunde et al., 2013; Cedro, 2016; Nicklen . Recognizing Cook's conceptual model of aggregate tree growth (Fig. 2), we propose to move beyond the net primary production proxy by explicitly simulating and validating stem radial growth ~~and TRW enriches the benchmark since correcting for~~ demographics. By doing so, we potentially enrich the verification by including the effects of potentially confounding factors ~~including climate responses,~~ such as forest structure, age and size trends ~~(Alexander et al., 2018; Nickless et al., 2011)~~

, as well as ~~(Alexander et al., 2018; Nickless et al., 2011; Jiang et al., 2018), phenology (Shen et al., 2020) and~~ sampling biases (Babst et al., 2014a), ~~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 (Shen et al., 2020), tree size (Jiang et al., 2018), climate sensitivity (Klesse et al., 2018), and atmospheric CO2 concentrations (Zuidema et al., 2020), 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 concentrations (Li et al., 2014; Rollinson et al., 2017).~~ in addition to climate and environmental forcing (Klesse et al., 2018; Zuidema et al., 2020; Li et al., 2014; Rollinson et al., 2017)~~.~~.

~~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 (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~~ Targeting both ~~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 (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~~ and age-structured information in observations and simulations (Fig. 3), we have proposed the use of four verification benchmarks created from observations and potentially simulated by LSMs, each of them defined by two complementary metrics (~~Table 1;~~ Fig. 2; Table S2).

(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~~ testing 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.~~ simulated stand density (Fig. 4a).

(ii)   The mature trees benchmark tests the capability of the model in simulating annual growth of ~~a~~ mature forest. Since this ~~benchmarks~~ benchmark aligns the observations by calendar year (Fig. 4b), it may reflect the effects of long-term environmental changes, if there were any and ~~the~~ if the observational record is long enough ~~as to experience~~ to express 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~~change, 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 ~~.~~(Fig. 4c). 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 either the timing of ~~either endogenous nor every exogenous disturbance~~ endogenous or exogenous disturbances, such as fire, pest and disease outbreak~~or dead~~, and 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 .~~ By forming a benchmark from the extrema of the empirical distribution of incremental growth in mature trees (e.g. as evident from Fig. 5a; see also Figs. 9, S8, and S9), we create a direct comparison with the simulated demographics of trees, as observed over a contemporaneous time interval. 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 ~~.~~(Figs. 5, 6, and 7). 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~~ 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 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. As a skilled model is expected to simulate not only the timing of extreme growth but also the magnitude of it, the metrics for this benchmark therefore were designed to both be evaluated using RMSE (Fig. 9).

5.2   ~~Verification of the benchmarks~~

5.2   Toward LSM verification using the ITRDB

20

~~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~~

585 ~~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~~

590 ~~deficiencies. The conclusion is supported by the verification (Table 2) where, except at site GIU, each site showed different weaknesses and strengths. The novel~~ Regardless of the approach to LSM verification, the largest publicly available archive of tree-ring records, the ITRDB, is prone to sampling biases (Klesse et al., 2018; Zhao et al., 2019). Although it may be difficult to correct the data for these biases, our benchmarks present two solutions for comparing LSM output to ITRDB observations of raw ring width. Simulating a size-structured population of trees enables comparing the observations relative to a benchmark

595 for a tall simulated tree, which compensates for the tendency of dendroclimatic sampling to select the oldest trees in a stand, which may 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, use of the tall tree benchmark protects against comparing the observed mean of a biased sample to the observed mean in unbiased simulation demographics. The second solution relies on the observation that the variation due to size-related growth by far exceeds the variation due

600 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 ~~thus provide new targets for evaluating LSMs' performance as each of the eight metrics could be used in the objective function of any dataassimilation technique (Peylin et al., 2016) to rigorously account for the information contained in TRW records .~~

605 ~~If the ITRDB is~~ provide a tool to start using ITRDB TRWs as a much-needed large-scale constraint on the maximum tree diameter and annual growth for the transition from pre-industrial to ~~be used, the number of benchmarks and the circumstances that can be used with confidence becomes more limited. The verification~~ present-day environmental conditions.

Our verification approach estimated the level of confidence for each benchmark from the fraction of cases for which scaling simulations to observational benchmarks for big-tree data would result in improving the model performance for observational

610 benchmarks for all-tree data. In other words, big-tree-biased verification data should not degrade model performance relative to all-tree verification data, and such tests were performed using the European biomass network dataset (BACI, 2020). The results, then, might inform the use of the ITRDB, which is suspected to contain large-tree bias, for large-scale verification of LSMs on decadal to centennial timescales (Fig. S1b).

Verification results (Table ~~2~~1) show that if the output of ORCHIDEE is benchmarked against data with ~~a~~ imposed 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~~produce similar conclusions as reached from use of all-tree data (Table 1, columns 1-2). This level of confidence ~~, i.e., 70%, is~~ is perhaps sufficient to support benchmarking a LSM against tens to hundreds of ITRDB sites ~~at the same time~~in aggregate. The same level of confidence is likely too low to benchmark a LSM~~(or an ecosystem model)~~, or any 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 verifiably improve the model. Given the ~~spatial extent of~~ limited spatial extent, species content, and environmental range of the European biomass network used in this study, the levels of confidence represent temperate and hemi-boreal forests, a subset of the ITRDB range and the actual distribution of forests (Fig. ~~S1)and the~~ S2). The validity of the use of the ITRDB data from boreal and tropical ~~forest~~forests will need to be verified ~~when~~ as suitable data become available.

~~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~~Across all species, benchmarking against extreme events, mature trees, and the size-related ~~trend~~growth appeared to be the least ~~demanding in terms of the biases present in the data used for benchmarking~~sensitive to big-tree bias (Table 1; bolded 63/88 cases). 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~~(Table 1, compare bolded values reported in columns 1-4 and 7-8 to those in columns 5-6). This finding ~~limits the use of the ITRDB data for~~suggests limited utility of ITRDB data to verify simulations using the young tree benchmarks. Because the true diameter distribution is not contained in ~~this database~~the ITRDB, 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 ~~.~~(Figs. S4 to S7). 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. The higher fraction of modifier-improved ORCHIDEE simulations for beech (87%) relative to spruce (64%) (Table 1, compare bolded values in rows 1-7 to bolded values in rows 8-11) suggest that the validity of the assumptions underlying the use of ITRDB data partly depends on tree species. Unfortunately, in this study, the variety in species in the was too limited to generalize this result in terms of plant functional types.

~~Benchmarking, even against ITRDB data is useful, irrespective of the proposed benchmark when the simulated TRWs differ substantially from~~ Even when benchmarks calculated from observations and simulations are not in agreement, they may nevertheless be used to identify ways in which to improve the observational database or the simulation model. Considering the

comparison of observed to simulated interquartile range (Fig. S9), we see that ~~the observed TRWs. For example, at the~~ GIU site ~~, where the simulated TRW is much smaller than the observed (Fig. S8 a),~~ is amongst the poorest simulated sites. However, ~~we found that~~ the simulation could be improved for ~~all~~ both metrics of all four benchmarks through the use of modifiers (Table 1, third row), despite the difference in simulated and observed stand structure (Fig. ~~S8 b). However, as shown above, inconsistencies between benchmarking the~~ S10). Yet we also found that inconsistencies between big-tree data and ~~the~~ all-tree ~~data start to appear~~ based benchmarks may appear even when the simulated ~~TRWs are better approaching the observations (Fig. S 2- 5~~ mean TRWs approach observed means (Figs. S4 to S7). This ~~implies~~ suggests that: 1) ITRDB data 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 observed stand structure and growth rates could hamper the use of ~~especially these benchmarks~~ young and mature tree benchmarks in particular.

6 ~~Outlook~~

5.1 Outlook

Tree-ring records ~~could~~ of incremental growth, suitably described in terms of benchmarks that more fully describe the information richness of these datasets, might complement well-established but short-term benchmarks for LSMs (Randerson et al., 2009), such as forest inventory data (Bellassen et al., 2010; Naudts et al., 2015), ~~FLUXNET sites~~ eddy covariance measurements (Blyth et al., 2010; Williams et al., 2009), Free Air ~~CO2~~ CO_2 Enrichment experiments (De Kauwe et al., 2013) and satellite observations of vegetation activity (Chen et al., 2011; Demarty et al., 2007). The novel benchmarks proposed here may also provide new targets for evaluating LSMs' performance, as the 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 datasets. The value of tree-ring records ~~can be further enriched~~ for LSM verification might be further enhanced by: (i) developing new~~and unbiased networksto complement the ITRDB,~~ unbiased networks, such as the European biomass network, to both complement and identify biases in the ITRDB, (ii) adding their stable isotope ratios ~~as benchmarks~~ to verification benchmarks that may be simulated by isotope-enabled LSMs (Levesque et al., 2019; Barichivich et al., 2020) and (iii) combining their use with high-frequency but ~~short- term~~ short-term eddy covariance measurements (Pappas et al., 2020; Teets et al., 2018), experimental data from plant growth under pre-industrial ~~CO2~~ CO_2 concentrations (Temme et al., 2015), and proxies of atmospheric composition (Campbell et al., 2017).

6 Conclusion

We have proposed and evaluated the use of four benchmarks and two metrics that leverage observed demographics to provide more nuanced verification targets for LSMs that simulate both demographic responses and their environmental forcing on decadal to centennial timescales. Using a small but relatively unbiased European biomass network datasets, we identify the extent to which presumed biases in the much larger ITRDB might degrade the validation of LSMs. We find that size, mature tree and extreme growth verification benchmarks are relatively insensitive to big-tree bias, but use of young tree benchmarks for verification of LSMs may require development of new unbiased observational TRW datasets, and/or innovative use of independent verification data.

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, SL and MNE 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.

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

Size dependent growth RMSE Slope of the residuals

705 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

710 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

715 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	
Metrics	RMSE (mm)	Slope of residuals (mm/yr)	RMSE (mm)	Slope of residuals (mm/yr)	RMSE (mm)	Slope of residuals (mm/yr)
Picea abies ~~(-0.60)~~	**DEO (-0.005)**	~~DEO (0.000)~~ DEO (0.000)	**DEO (-75.97)**	~~DEO(0.78)~~ DEO (0.78)	~~DEO(8.11)~~ DEO (8.11)	~~DEO(1.47)~~
	DVN (-0.182)	~~DVN (0.000)~~ DVN (0.000)	**DEO (-161.69)**	~~DVN(0.70)~~ DVN (-0.69)	**DVN (~~-11.68~~-11.96)**	**DVN (~~-0.43~~-0...)**
	GIU (-0.600)	**GIU (-0.007)**	**GIU (-131.25)**	**GIU (-0.68)**	**GIU (~~-39.70~~-39.96)**	**GIU (-1.30...)**
	~~HD2 (0.009)~~ HD2 (0.009)	**HD2 (-0.002)**	~~HD2 (15.60)~~ HD2 (15.60)	~~HD2 (0.53)~~ HD2 (0.53)	~~HD2(1.95)~~ HD2 (1.95)	~~HD2 (0.56)~~
	~~SCH (0.029)~~ SCH (0.029)	**SCH (-0.004)**	**SCH (-182.46)**	**SCH (-0.96)**	~~SCH (57.56)~~ SCH (57.55)	~~SCH (5.49)~~
	SOB (~~-0.0008~~-0.008)	~~SOB (-0.001)~~ SOB (0.001)	**SOB (-20.22)**	~~SOB (0.50)~~ SOB (0.49)	~~SOB (9.32)~~ SOB (9.32)	~~SOB (1.29)~~
	TIC (-0.151)	**TIC (-0.001)**	~~TIC (24.47)~~ TIC (24.47)	~~TIC (1.52)~~ TIC (1.52)	**TIC (-10.33)**	**TIC (~~-0.27~~-0...)**
Fagus sylvatica	**CAN (-0.046)**	**CAN (~~-0.003~~-0.004)**	**CAN (-74.03)**	**CAN (-1.27)**	**CAN (~~-5.81~~-5.91)**	CAN (0.16)
	SOR (0.007)	**SOR (~~-0.004~~-0.005)**	**SOR (-116.26)**	**SOR (-1.62)**	SOR (2.69)	SOR (-1.13...)
	~~GIU~~ TER (**~~-0.600~~-0.060**)	~~GIU~~ TER (**~~-0.0007~~-0.000**)	~~GIU~~TER (**~~-131.25~~-3.73**)	~~GIU~~TER (**~~-0.68~~-0.09**)	~~GIU~~TER (**~~-39.70~~-15.93**)	GIU(-1.30)
	~~GIU~~ ZOF (**~~-0.600~~-0.183**)	~~GIU~~ ZOF (**~~-0.0007~~-0.000**)	~~GIU~~ZOF (**~~-131.25~~-42.72**)	~~GIU(-0.68)~~ ZOF (0.02)	~~GIU~~ZOF (**~~-39.70~~-11.98**)	~~GIU~~ZO... (**~~-1.30~~-0...**)
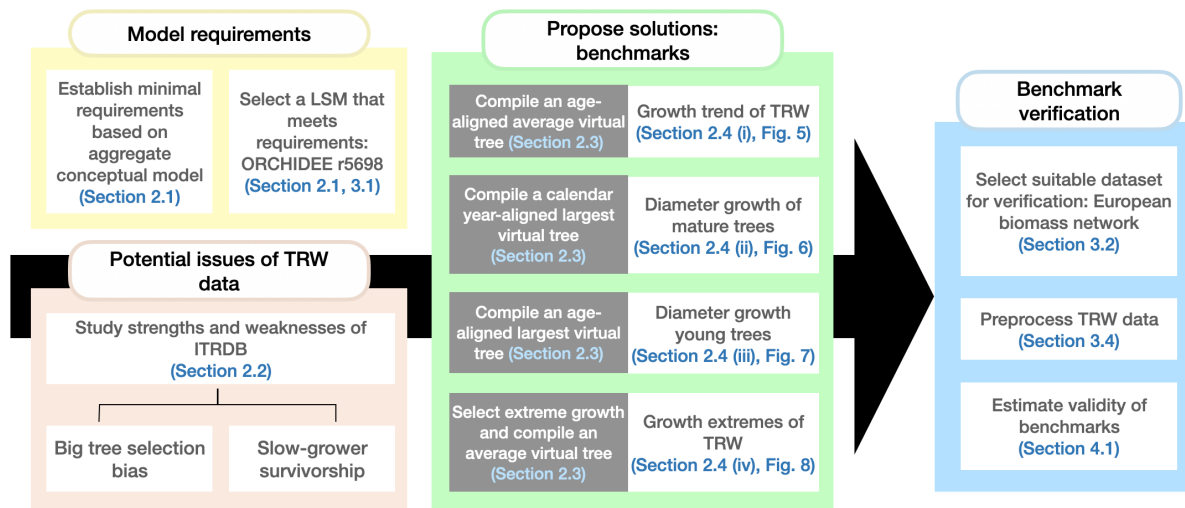
Workflow of ~~the~~ this study and the dependencies between the different sections of the manuscript. Each colour represents a different aim of the study. The arrow shows ~~how~~ that the outcomes of the first three aims ~~are~~ have to be combined to ~~achieve the final aim, i.e.,~~ verify the proposed benchmarks. In this study, "virtual tree" refers to a tree that does not exist in the data but is constructed from the data in varying ways. Section 2.3 provides detailed descriptions of the different virtual trees used.

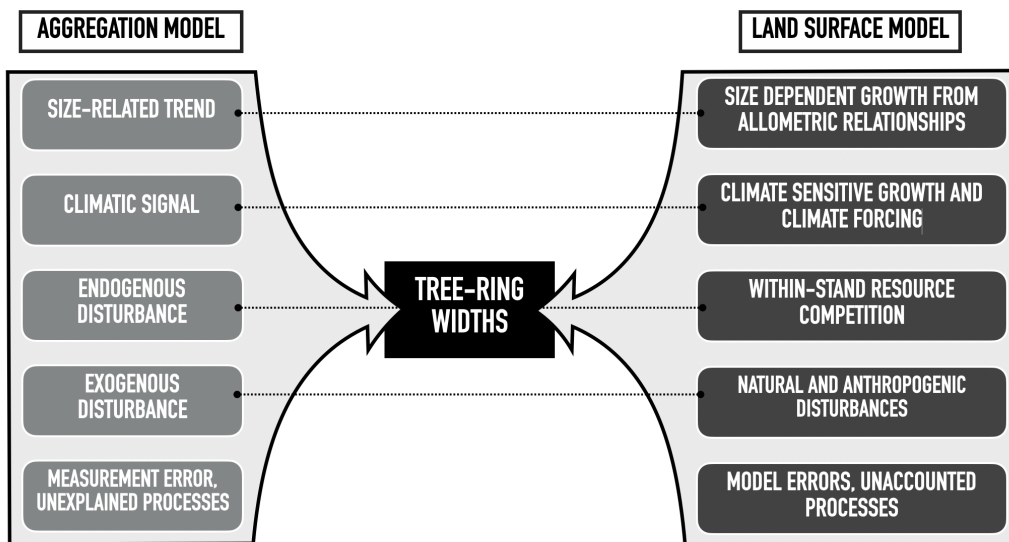**Figure 1.** ~~Conceptual illustration of the expected reduction in model uncertainty following the use of tree-ring width records to benchmark land-surface models.~~ ~~Note that the anticipated uncertainty reduction assumes that a large part of the model uncertainty comes from the model formulation 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 model of tree-ring growth and the equivalent processes in land-surface models.~~ Main drivers of the linear aggregate conceptual model of tree-ring growth (Cook and Kairiukstis, 1990) 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.
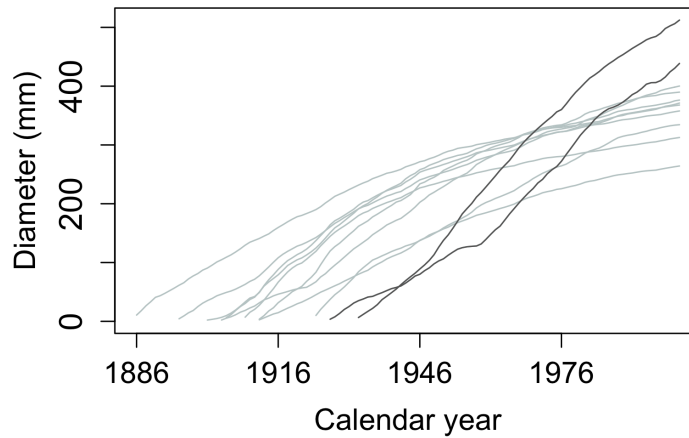
Figure 3. ~~Solutions and virtual trees to account for challenges to use ITRDB datasets when evaluating land-surface models against ITRDB tree-ring data.~~ **Example of a typical data record in the ITRDB dataset.** Each dataset in ITRDB consists of increment cores from multiple trees (tens to several hundred, depending on the dataset) with varying ages and growth rates. In this figure, observations are from a pine forest in Germany archived as germ214 (Neuwirth et al., 2007) (Table S3). The diameter of individual trees was reconstructed by summing the annual tree-ring from the dataset. Note the presence of both fast-growing relatively young trees (dark-gray lines) and slow-growing older trees (light-gray lines).
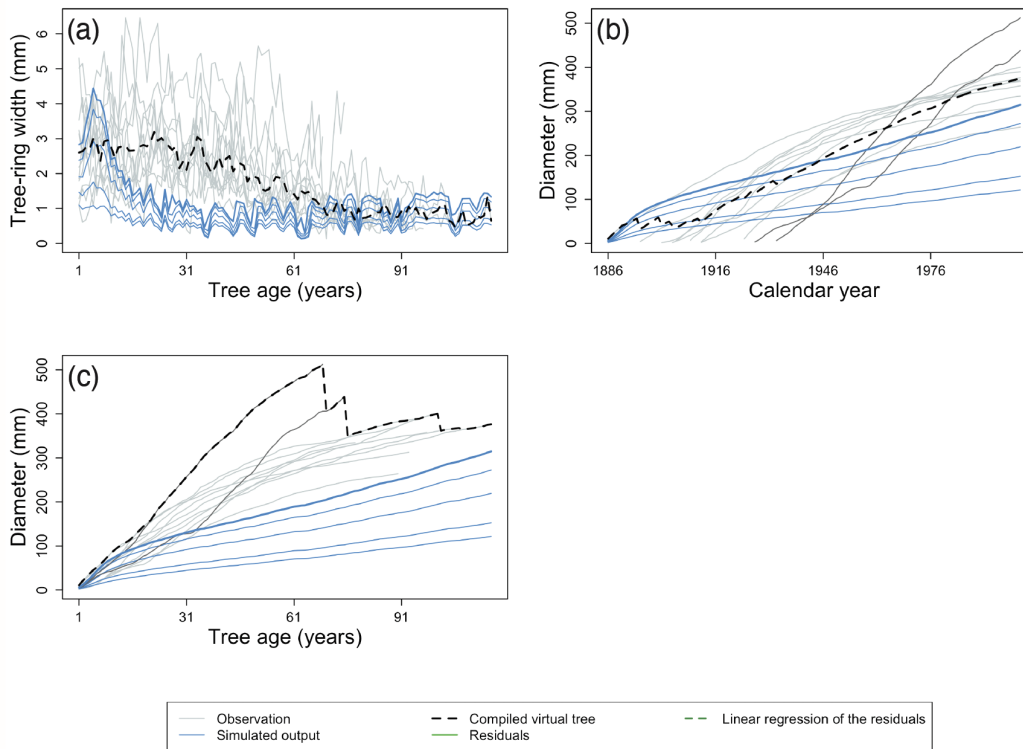
Figure 4. Using virtual trees to account for challenges related to the use of ITRDB datasets when evaluating land-surface. (a) Data-model comparison may overcome the big-tree selection bias by comparing only the simulated ~~biggest~~ largest diameter class (bold blue line) for evaluation rather than 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 largest virtual tree (black dotted line) results in a very different virtual tree compared to Fig 3b. The largest 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 ~~(?)~~ (Neuwirth et al., 2007) (Table ~~S2~~S3). Simulation were run with ORCHIDEE r5698.
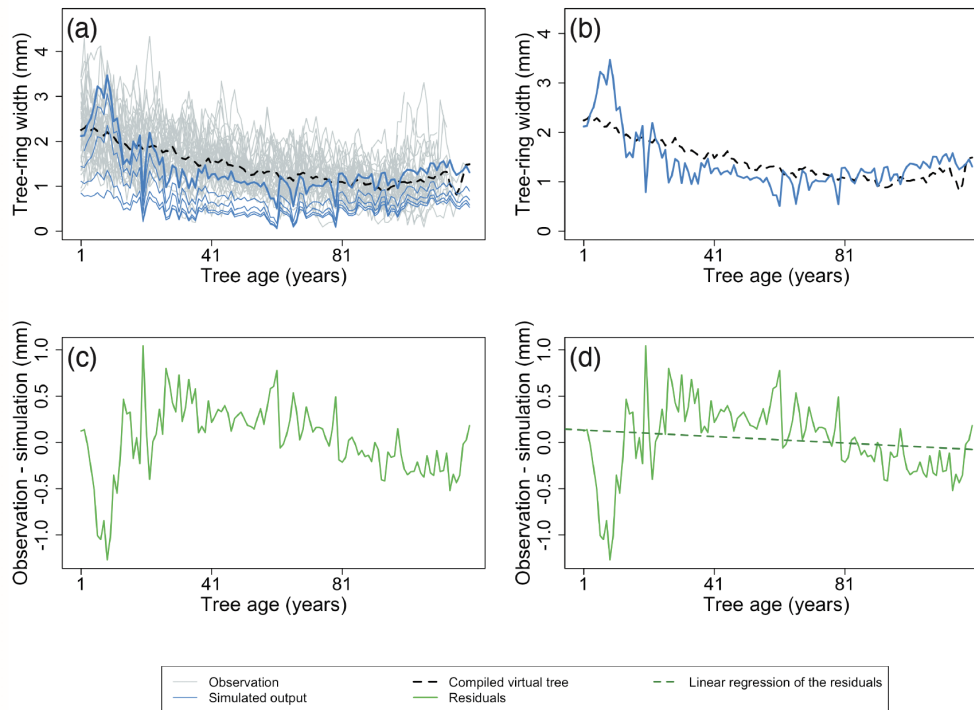
Figure 5. Example of the major steps for calculating the metrics of the benchmark for the size-related trend in diameter increment. The size-related trend in diameter increment can be assessed by calculating a time series for the average ring width after aligning the age of the individual trees (a, b). Observations are shown as grey lines and simulation as blue lines. The ~~biggest~~ largest diameter class from the simulation 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 a *Pinus sylvestris* site in Finland archived as finl052 ~~(NOAA-WDS for Paleoclimatology, 2020a)~~ (Meriläinen et al., 2004). For this example, calculated RMSE (b) is 0.39 (mm), and the slope of residuals (d) is -0.002 (mm/yr).
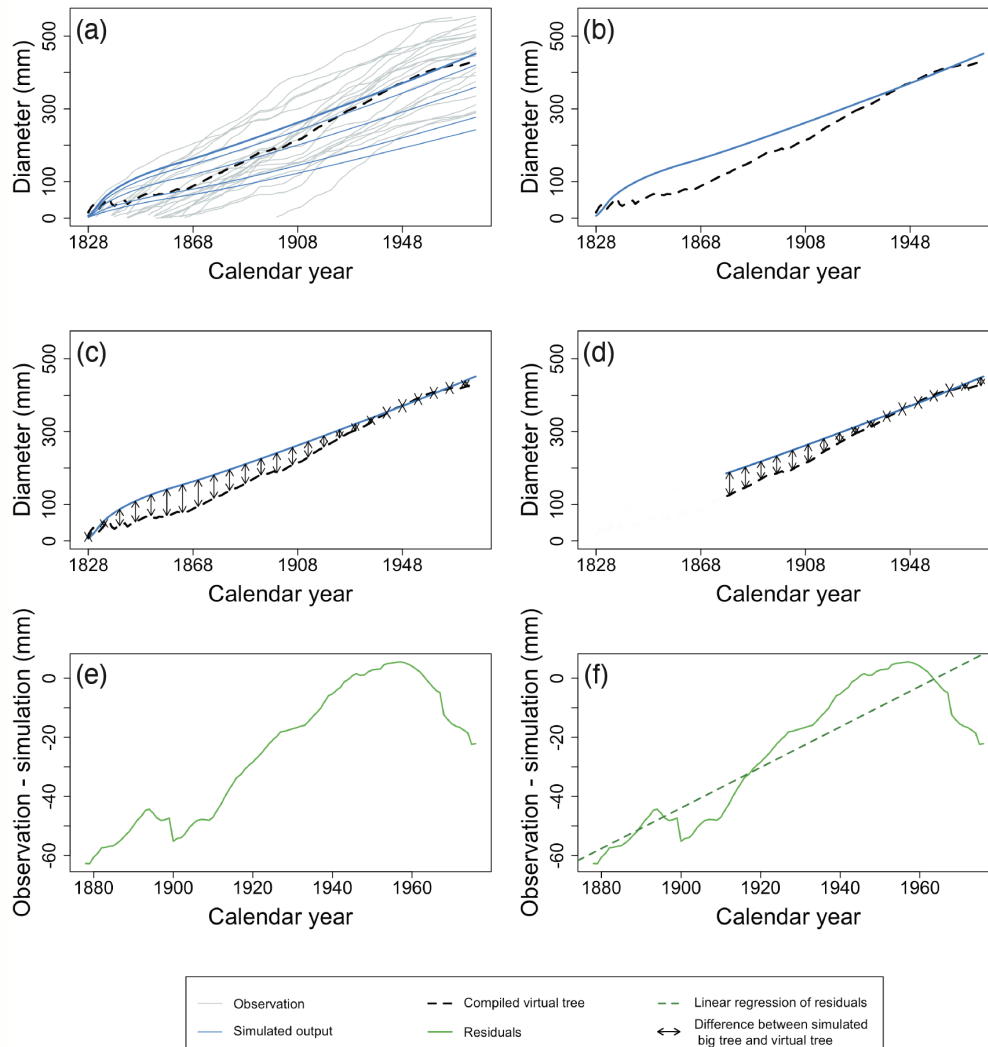
Figure 6. Example of the major steps in calculating the metrics of the benchmark for the diameter increment in mature trees.
Individual tree records are ordered by calendar year and for each year the average observed diameter is calculated (a). Observations are ~~685~~
shown as grey lines and simulation as blue lines. The ~~biggest~~ largest diameter class from the simulation is presented by the bold blue line.
Black dotted lines represent the yearly average of observations. Note that X-axis in Fig. ~~6~~ 6 is different from Fig. ~~5.~~ 5. Under the assumption
that the observed trees are the biggest trees from a given site, the virtual tree is compared with the ~~biggest~~ largest 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 a *Pinus sylvestris* site in Scotland archived as brit021 (Schweingruber, 1995).
In this case, RMSE (d) and the slope of residuals (f) were calculated as 33.65 (mm) and 0.68 (mm/yr), respectively.
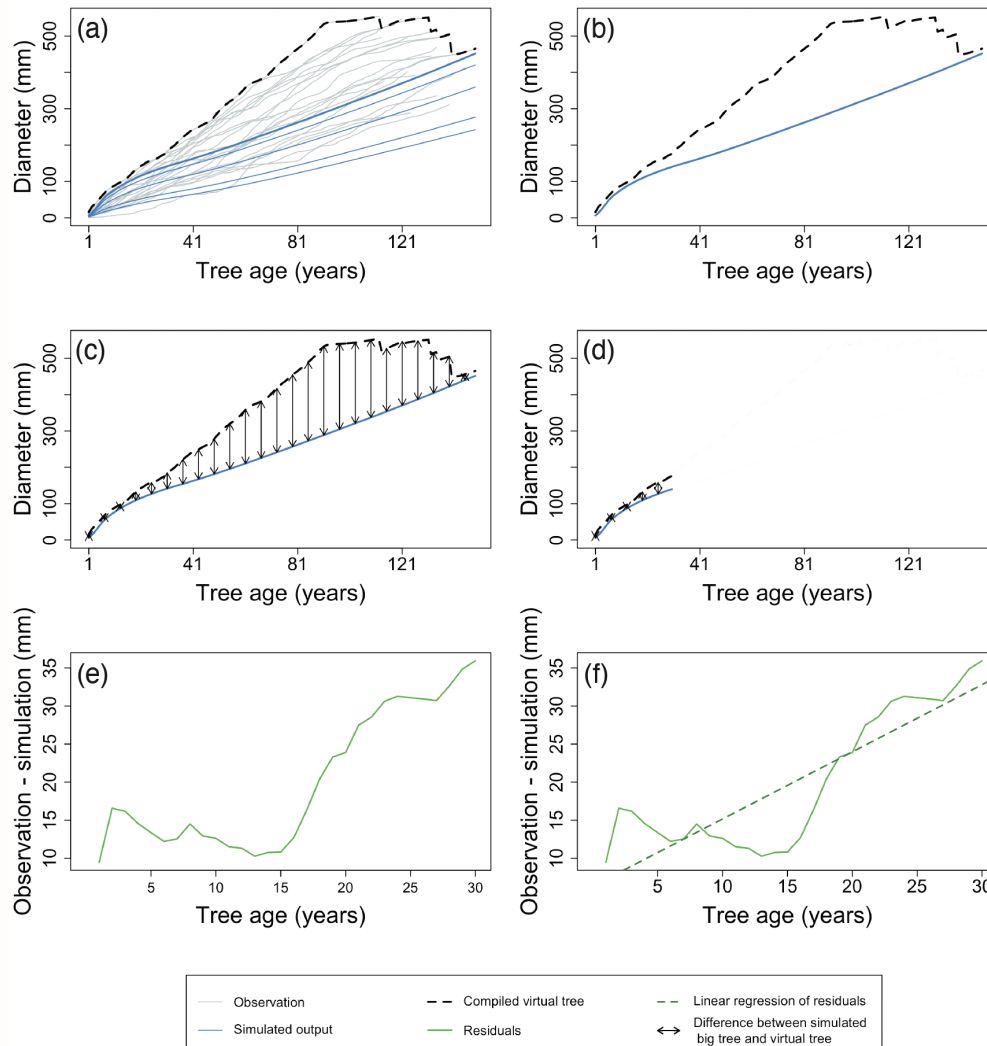
Figure 7. Example of the major steps in calculating the metrics of the benchmark for diameter increment in young trees. After aligning the TRW records of the individual trees by their age, a virtual tree is constructed by taking the maximum observed diameter of all trees for each year (a). Observations are shown as grey lines and simulation as blue lines. The ~~biggest~~ largest diameter class from the simulation 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 (Schweingruber, 1995). The calculated RMSE (d) was 21.86 (mm) and the slope of residuals (f) was 0.88 (mm/yr) for this example.
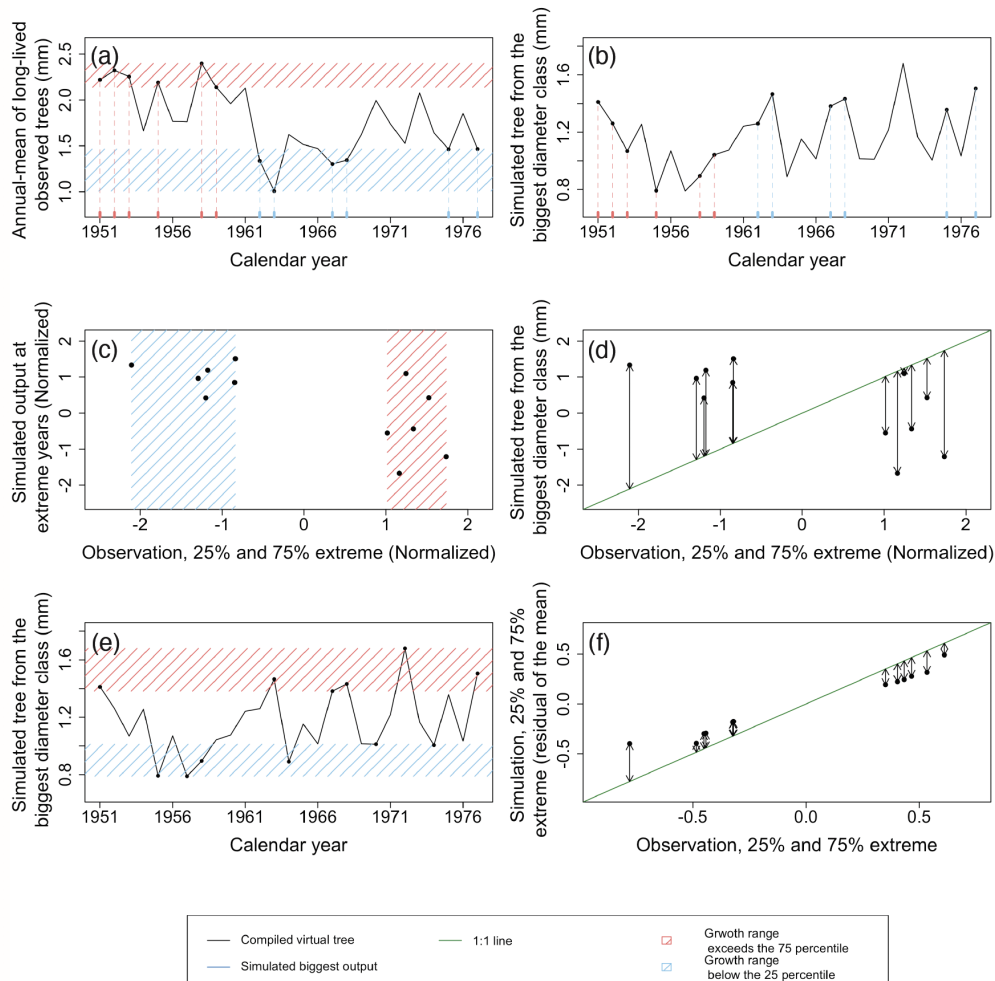
Figure 8. Example of the major steps in calculating the metrics of the benchmark for extreme growth events. In this benchmark, extreme growth is defined as the first and last quartiles in TRW ordered by calendar year and averaged over the individual trees records (a). Red shaded area and ticks represent 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 (NOAA-WDS for Paleoclimatology, 2020b). In this test case, RMSE for extreme years (d) was 0.57 (mm) and RMSE for extreme growth (amplitude; f) was 0.03 (scaled).
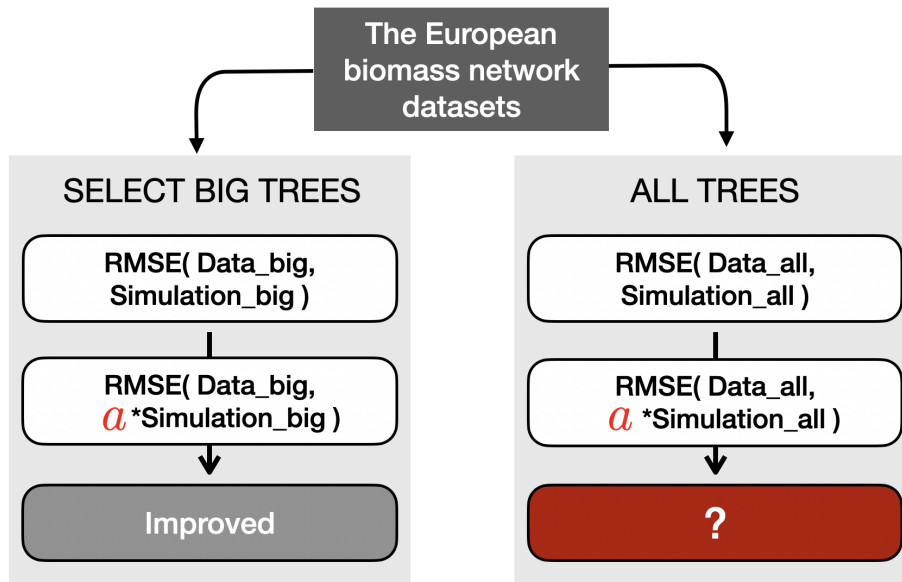
Figure 9. Schematic representation of the verification process for the RMSE-metric. Before the verification, two types of datasets were prepared: big-tree data (limited to the 15% biggest trees) and all-tree data. In this example, the simulated TRW was multiplied with a verified modifier such that it minimized the RMSE between the simulated and observed TRW for the 15% biggest trees (see ~~section 3.4~~ Section 3.2 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~~ 3.3 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.

References

2k Consortium, P.: Continental-scale temperature variability during the past two millennia, Nature Geoscience, 6, 339–346, doi: 10.1038/N-GEO1797, chin037 dataset accessed at https://www.ncdc.noaa.gov/paleo-search/study/14888, 2021, 2013.

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, 265–276, https://doi.org/10.1007/s00468-017-1629-0, 2018.

Amthor, J. S.: The McCree–de Wit–Penning de Vries–Thornley Respiration Paradigms: 30 Years Later, Annals of Botany, 86, 1–20, https://doi.org/10.1006/anbo.2000.1175, 2000.

Archambault, S. and Bergeron, Y.: Lac Duparquet - THOC - ITRDB CANA106 - RWL, doi: 10.25921/rmbz-ga96, 2002.

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, 307–322, https://doi.org/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, 153–161, https://doi.org/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, 8, https://doi.org/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, Quaternary Science Reviews, 197, 1–20, 2018.

BACI: European biomass network, http://www.baci-h2020.eu, 2020.

Bakker, J. D.: A new, proportional method for reconstructing historical tree diameters, Canadian Journal of Forest Research, 35, 2515–2520, https://doi.org/10.1139/x05-136, 2005.

Barichivich, J., Peylin, P., Launois, T., Daux, V., Risi, C., Jeong, J., and Luyssaert, S.: A triple tree-ring constraint for tree growth and physiology in a global land surface model, Biogeosciences Discussions, 2020, 1–36, https://doi.org/10.5194/bg-2020-446, 2020.

Bell, R., Magre, F., and Senter, D.: Jefferson County Missouri - JUVI - ITRDB MO009 - RWL, doi: 10.25921/d4mc-nh83, 2002.

Bellassen, V., Le Maire, G., Dhôte, J., Ciais, P., and Viovy, N.: Modelling forest management within a global vegetation model—Part 1: Model structure and general behaviour, Ecological Modelling, 221, 2458–2474, https://doi.org/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, Journal of Hydrometeorology, 11, 509–519, https://doi.org/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, 2193–2222, https://doi.org/10.5194/gmd-7-2193-2014, 2014.

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

Bowman, D. M. J. S., Balch, J. K., Artaxo, P., Bond, W. J., Carlson, J. M., Cochrane, M. A., D'Antonio, C. M., DeFries, R. S., Doyle, J. C., Harrison, S. P., Johnston, F. H., Keeley, J. E., Krawchuk, M. A., Kull, C. A., Marston, J. B., Moritz, M. A., Prentice, I. C.,

Roos, C. I., Scott, A. C., Swetnam, T. W., van der Werf, G. R., and Pyne, S. J.: Fire in the Earth System, Science, 324, 481–484, https://doi.org/10.1126/science.1163886, 2009.

Bräuning, A., De Ridder, M., Zafirov, N., García-González, I., Petrov Dimitrov, D., and Gärtner, H.: TREE-RING FEATURES: INDICA-
TORS OF EXTREME EVENT IMPACTS, IAWA Journal, 37, 206–231, https://doi.org/10.1163/22941932-20160131, 2016.

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 Biogeochemical Cycles, 26, n/a–n/a, https://doi.org/10.1029/2011GB004143, 2012.

Brienen, R. J. W., Gloor, M., and Ziv, G.: Tree demography dominates long-term growth trends inferred from tree rings, Global change
biology, 23, 474–484, https://doi.org/10.1111/gcb.13410, 2017.

Briffa, K. and Schweingruber, F.: Cascade Radar St. Payette - PCEN - ITRDB ID007 - RWL, doi: 10.25921/6pcv-me64, 2002.

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, Dendroclimatology, 11, 113–145, https://doi.org/10.1007/978-1-4020-
5725-0_5, 2011.

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

Briongos, J. and Cerro-Barja, A.: La Camarilla el Provencio Cuenca Undisturbed - PIPN - ITRDB SPAI055 - RWL, doi: 10.25921/p9h0-
g517, 2007.

Bunde, A., Büntgen, U., Ludescher, J., Luterbacher, J., and von Storch, H.: Is there memory in precipitation?, Nature Climate Change, 3,
174–175, https://doi.org/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, 84–87, https://doi.org/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, Climate of the Past, 15, 1503–1536, https://doi.org/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, p. 24,
https://doi.org/10.2312/GFZ.b103-16042, 2016.

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, Journal of
Geophysical Research: Atmospheres, 116, https://doi.org/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, Geoscientific Model Development Discussions, pp. 1–35, https://doi.org/10.5194/gmd-
2016-26, 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), Geoscientific Model Development,
11, 771–791, https://doi.org/10.5194/gmd-11-771-2018, 2018.

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

795   Reanalysis Project, Quarterly Journal of the Royal Meteorological Society, 137, 1–28, https://doi.org/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, Springer Netherlands, Dordrecht, https://doi.org/10.1007/978-94-015-
7879-0, 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

800   development for palaeoclimatic studies, The Holocene, 5, 229–237, https://doi.org/10.1177/095968369500500211, 1995.

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, Global and Planetary Change, 60, 289–305, https://doi.org/https://doi.org/10.1016/j.gloplacha.2007.03.004, 2008.

Davi, N., D'Arrigo, R., Jacoby, G.C.; Buckley, B., and Kobayashi, O.: Shiretoko - PCGN - ITRDB JAPA011 - RWL, doi: 10.25921/wnc6-
1r75, 2011.

805   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, Global Change Biology, 19, 1759–1779, https://doi.org/10.1111/gcb.12164, 2013.

De Schepper, V. and Steppe, K.: Development and verification of a water and sugar transport model using measured stem diameter variations,

810   Journal of Experimental Botany, 61, 2083–2099, https://doi.org/10.1093/jxb/erq018, 2010.

Deleuze, C. and Houllier, F.: Simple process-based xylem growth model for describing wood microdensitometric profiles, Journal of Theo-
retical Biology, 193, 99–113, https://doi.org/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, Annals
of Forest Science, 61, 327–335, https://doi.org/10.1051/forest:2004026, 2004.

815   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, Geophysical Research Letters, 34, https://doi.org/10.1029/2007GL030014, 2007.

Drew, D. M., Downes, G. M., and Battaglia, M.: CAMBIUM, a process-based model of daily xylem development in Eucalyptus, Journal of
Theoretical Biology, 264, 395–406, https://doi.org/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

820   Interface within the LMD Atmospheric General Circulation Model, Journal of Climate, 6, 248–273, https://doi.org/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, Ecological Modelling, 185, 407–436,
https://doi.org/10.1016/j.ecolmodel.2005.01.004, 2005.

825   Eyring, V., Lamarque, J.-F., Hess, P., Arfeuille, F., Bowman, K., Chipperfiel, M. P., Duncan, B., Fiore, A., Gettelman, A., and Giorgetta,
M. A.: Overview of IGAC/SPARC Chemistry-Climate Model Initiative (CCMI) community simulations in support of upcoming ozone
and climate assessments, SPARC newsletter, 40, 48–66, 2013.

Farquhar, G. D.: Models of Integrated Photosynthesis of Cells and Leaves, Philosophical Transactions of the Royal Society B: Biological
Sciences, 323, 357–367, https://doi.org/10.1098/rstb.1989.0016, 1989.

830 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), Geoscientific Model Development, 8, 3593–3619, https://doi.org/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 Physiology, 32, 648–666, https://doi.org/10.1093/treephys/tpr138, 2012.

835 Friedlingstein, P., Betts, R., Bopp, L., Bloh, W. V., Brovkin, V., Doney, S., Eby, M., Fung, I., Govindasamy, B., 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., Schnurr, R., Strassmann, K., Thompson, S., J Weaver, A., Yoshikawa, C., and Zeng, N.: Climate –carbon cycle feedback analysis, results from the C4MIP model intercomparison, Journal of Climate, 19, 3337–3353, https://doi.org/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
840 projections due to carbon cycle feedbacks, Journal of Climate, 27, 511–526, https://doi.org/10.1175/JCLI-D-12-00579.1, 2014.

Friend, A. D., Eckes-Shephard, A. H., Fonti, P., Rademacher, T. T., Rathgeber, C. B., Richardson, A. D., and Turton, R. H.: On the need to consider wood formation processes in global vegetation models and a suggested approach, Annals of Forest Science, 76, https://doi.org/10.1007/s13595-019-0819-x, 2019.

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

845 Fritts, H. C., Shashkin, A., and Downes, G. M.: A simulation model of conifer ring growth and cell structure, Tree-ring analysis: biological, methodological and environmental aspects.. CABI publishing: Wallingford, UK, pp. 3–32, 1999.

Griggs, C., Kuniholm, P., and Petrucci, A.: Devecikonak Forest - QUSP - ITRDB TURK027 - RWL, doi: 10.25921/axza-3w94, 2006.

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, 235–238, https://doi.org/10.1177/095968369700700212, 1997.

850 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, Geophysical Research Letters, 40, 5234–5239, https://doi.org/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, Frontiers in Plant Science, 8, https://doi.org/10.3389/fpls.2017.00182, 2017.

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

Hess, C., Niemeyer, T., Fichtner, A., Jansen, K., Kunz, M., Maneke, M., von Wehrden, H., Quante, M., Walmsley, D., von Oheimb, G., and Härdtle, W.: Anthropogenic nitrogen deposition alters growth responses of European beech (Fagus sylvativa L.) to climate change, Environmental Pollution, 233, 92–98, https://doi.org/10.1016/j.envpol.2017.10.024, 2018.

860 Hirata, R., Hirano, T., Saigusa, N., Fujinuma, Y., Inukai, K., Kitamori, Y., Takahashi, Y., and Yamamoto, S.: Seasonal and inter-annual variations in carbon dioxide exchange of a temperate larch forest, Agricultural and Forest Meteorology, 147, 110–124, https://doi.org/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, 67–78, https://doi.org/10.1007/s00468-005-0014-6, 2006.

865 Hughes, M. K., Swetnam, T. W., and Diaz, H. F.: Dendroclimatology, vol. 11 of *Developments in Paleoenvironmental Research*, Springer Netherlands, Dordrecht, https://doi.org/10.1007/978-1-4020-5725-0, 2011.

IPCC: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, https://doi.org/10.1017/CBO9781107415324, 2013.

870 Jiang, X., Huang, J.-G., Cheng, J., Dawson, A., Stadt, K. J., Comeau, P. G., and Chen, H. Y.: Interspecific variation in growth responses to tree size, competition and climate of western Canadian boreal mixed forests, Science of The Total Environment, 631-632, 1070–1078, https://doi.org/10.1016/j.scitotenv.2018.03.099, 2018.

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

875 NCEP/NCAR 40-Year Reanalysis Project, Bulletin of the American Meteorological Society, 77, 437–471, https://doi.org/10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2, 1996.

Keeling, C. D., Chin, J. F. S., and Whorf, T. P.: Increased activity of northern vegetation inferred from atmospheric CO_2 measurements, Nature, 382, 146–149, https://doi.org/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

880 Frank, D. C.: A Combined Tree Ring and Vegetation Model Assessment of European Forest Growth Sensitivity to Interannual Climate Variability, Global Biogeochemical Cycles, 32, 1226–1240, https://doi.org/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, Scientific Reports, 9, 2758, https://doi.org/10.1038/s41598-019-39373-1, 2019.

885 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 Biogeochemical Cycles, 19, https://doi.org/10.1029/2003GB002199, 2005.

Krusic, P. and Cook, E.: Above Gheri - ABSB - ITRDB NEPA003 - RWL, doi: 10.25921/xtp6-zh97, 2005.

Laloyaux, P., de Boisseson, E., Balmaseda, M., Bidlot, J.-R., Broennimann, S., Buizza, R., Dalhgren, P., Dee, D., Haimberger, L., Hersbach,

890 H., Kosaka, Y., Martin, M., Poli, P., Rayner, N., Rustemeier, E., and Schepers, D.: CERA-20C: A Coupled Reanalysis of the Twentieth Century, Journal of Advances in Modeling Earth Systems, 10, 1172–1195, https://doi.org/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, 225, 1019–1021, https://doi.org/10.1126/science.225.4666.1019, 1984.

Leuzinger, S., Manusch, C., Bugmann, H., and Wolf, A.: A sink-limited growth model improves biomass estimation along boreal and alpine

895 tree lines, Global Ecology and Biogeography, 22, 924–932, https://doi.org/10.1111/geb.12047, 2013.

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, Nature Communications, 10, 742, https://doi.org/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,

900 and growth, Biogeosciences, 11, 6711–6724, https://doi.org/10.5194/bg-11-6711-2014, 2014.

Lu, C. and Tian, H.: Global nitrogen and phosphorus fertilizer use for agriculture production in the past half century: shifted hot spots and nutrient imbalance, Earth System Science Data, 9, 181–192, https://doi.org/10.5194/essd-9-181-2017, 2017.

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.,

905   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, 3857–3874, https://doi.org/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.,
910   Valentini, R., and Grace, J.: The human footprint in the carbon cycle of temperate and boreal forests, Nature, 447, 849–851, https://doi.org/10.1038/nature05847, 2007.

McGuffie A., K. and Henderson-Sellers: Practical Climate Modelling, in: A Climate Modelling Primer, John Wiley & Sons, Ltd, https://doi.org/doi:10.1002/0470857617.ch6, 2005.

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

915   Melvin, T.: Vytamoselka - PISY - ITRDB FINL055 - RWL, doi: 10.25921/kayq-q277, 2005.

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, Ecology Letters, 8, 1183–1190, https://doi.org/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.,
920   Yousefpour, R., Krejza, J., Collalti, A., and Reyer, C. P. O.: Forest carbon allocation modelling under climate change, Tree Physiology, https://doi.org/10.1093/treephys/tpz105, 2019.

Meriläinen, J., Lindholm, M., Timonen, M., and Kolström, T.: Kukelo Ahmovaara Juuka - PISY - ITRDB FINL052 - RWL, doi: 10.25921/jjxq-dh12, 2004.

Misson, L.: MAIDEN: a model for analyzing ecosystem processes in dendroecology, Canadian Journal of Forest Research, 34, 874–887,
925   https://doi.org/10.1139/x03-252, 2004.

Moorcroft, P. R., Hurtt, G. C., and Pacala, S. W.: A method for scaling vegetation dynamics: The ecosystem demography model (ED), Ecological Monographs, 71, 557–586, https://doi.org/10.1890/0012-9615(2001)071[0557:AMFSVD]2.0.CO;2, 2001.

Nash, S. E.: Fundamentals of tree-ring research. James H. Speer., Geoarchaeology, 26, 453–455, https://doi.org/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.,
930   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, Geoscientific Model Development, 8, 2035–2065, https://doi.org/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
935   design on tree-ring-based quantification of forest growth, Global Change Biology, 20, 2867–2885, https://doi.org/10.1111/gcb.12599, 2014.

Neuwirth, B., Schweingruber, F. H., and Winiger, M.: Spatial patterns of central European pointer years from 1901 to 1971, Dendrochronologia, 24(2-3), 79–89, doi: 10.1016/j.dendro.2006.05.004, germ214 dataset accessed at https://www.ncdc.noaa.gov/paleo/study/16747, 2020, 2007.

940   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, Global Change Biology, 25, 911–926, https://doi.org/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, South African Journal of Science, 107, 86–95, https://doi.org/10.4102/sajs.v107i5/6.356, 2011.

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

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

OED online: "virtual, adj. and n.".., https://oed.com/view/Entry/223829?redirectedFrom=virtual.

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
950   North America Using Model-Data Comparison, Ecosystems, 21, 1042–1057, https://doi.org/10.1007/s10021-017-0203-3, 2018.

Panthi, S., Fan, Z.-X., van der Sleen, P., and Zuidema, P. A.: Long-term physiological and growth responses of Himalayan fir to environmental
change are mediated by mean climate, Global Change Biology, 26, 1778–1794, https://doi.org/10.1111/gcb.14910, 2020.

Pappas, C., Maillet, J., Rakowski, S., Baltzer, J. L., Barr, A. G., Black, T. A., Fatichi, S., Laroque, C. P., Matheny, A. M., Roy, A., Sonnentag,
O., and Zha, T.: Aboveground tree growth is a minor and decoupled fraction of boreal forest carbon input, Agricultural and Forest
955   Meteorology, 290, 108 030, https://doi.org/10.1016/j.agrformet.2020.108030, 2020.

Paris Agreements: Paris agreement, in: Report of the Conference of the Parties to the United Nations Framework Convention on Climate
Change (21st Session, 2015: Paris). Retrieved December, vol. 4, p. 2017, HeinOnline, 2015.

Peylin, P., Bacour, C., MacBean, N., Leonard, S., Rayner, P., Kuppel, S., Koffi, E., Kane, A., Maignan, F., Chevallier, F., Ciais, P., and Prunet,
P.: A new stepwise carbon cycle data assimilation system using multiple data streams to constrain the simulated land surface carbon cycle,
960   Geoscientific Model Development, 9, 3321–3346, https://doi.org/10.5194/gmd-9-3321-2016, 2016.

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, 373–
385, https://doi.org/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,
965   R., Covey, C., Running, S. W., and Fung, I. Y.: Systematic assessment of terrestrial biogeochemistry in coupled climate-carbon models,
Global Change Biology, 15, 2462–2484, https://doi.org/10.1111/j.1365-2486.2009.01912.x, 2009.

Rollinson, C. R., Liu, Y., Raiho, A., Moore, D. J. P., McLachlan, J., Bishop, D. A., Dye, A., Matthes, J. H., Hessl, A., Hickler, T., Pederson,
N., Poulter, B., Quaife, T., Schaefer, K., Steinkamp, J., and Dietze, M. C.: Emergent climate and CO 2 sensitivities of net primary
productivity in ecosystem models do not agree with empirical data in temperate forests of eastern North America, Global Change Biology,
970   23, 2755–2767, https://doi.org/10.1111/gcb.13626, 2017.

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, Geoscientific
Model Development, 9, 223–245, https://doi.org/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
975   approach, Ecological Modelling, 200, 279–307, https://doi.org/10.1016/j.ecolmodel.2006.09.006, 2007.

Schulman, E.: Longevity under Adversity in Conifersca, Science, 119, 396–399, http://www.jstor.org/stable/1682970, 1954.

Schweingruber, F. H.: Drimmie Schottland - PISY - ITRDB BRIT021 - RWL, https://www.ncdc.noaa.gov/paleo/study/4399, doi:
10.25921/ctw7-nj47, 1995.

Shen, Y., Fukatsu, E., Muraoka, H., Saitoh, T. M., Hirano, Y., and Yasue, K.: Climate responses of ring widths and radial growth phenology of Betula ermanii, Fagus crenata and Quercus crispula in a cool temperate forest in central Japan, Trees, https://doi.org/10.1007/s00468-019-01948-w, 2020.

Smith, B.: LPJ-GUESS-an ecosystem modelling framework, Department of Physical Geography and Ecosystems Analysis. INES, Sölvegatan, 12, 22 362, 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 Physiology, 26, 257–273, https://doi.org/10.1093/treephys/26.3.257, 2006.

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

Strumia, G.: Weinerwald - QUPE - ITRDB AUST112 - RWL, doi: 10.25921/18tv-he69, 2005.

Teets, A., Fraver, S., Weiskittel, A. R., and Hollinger, D. Y.: Quantifying climate–growth relationships at the stand level in a mature mixed-species conifer forest, Global change biology, 24, 3587–3602, 2018.

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 CO_2, Ecology and Evolution, 5, 4949–4961, https://doi.org/10.1002/ece3.1687, 2015.

Tessier, L.: Mimet (mt. L' Eloile) - PISY - ITRDB FRAN4 - RWL, doi: doi.org/10.25921/e5zh-ry13, 1996.

Vaganov, E. A., Hughes, M. K., and Shashkin, A. V.: Growth dynamics of conifer tree rings: images of past and future environments, vol. 183, Springer, New York, 2006.

Viovy, N.: CRUNCEP data set Version 5.3.2, version 5.3.2 is not available but version 7 is accessible at: https://rda.ucar.edu/datasets/ds314.3/!description, 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, Geoscientific Model Development, 12, 4751–4779, https://doi.org/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 physiology, 35, 305–318, https://doi.org/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, 1341–1359, https://doi.org/10.5194/bg-6-1341-2009, 2009.

Wilson, B. F. and Howard, R. A.: A computer model for cambial activity, Forest Science, 14, 77–90, https://doi.org/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 Biogeochemical Cycles, 25, https://doi.org/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, 2747–2767, https://doi.org/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 Biogeochemical Cycles, 24, https://doi.org/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, 410–425, https://doi.org/10.1007/s10021-017-0157-5, 2018.

1020 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, Journal of Biogeography, 46, 355–368, https://doi.org/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, 29–38, https://doi.org/10.1007/s00468-010-0473-2, 2011.

1025 Zuidema, P. A., Poulter, B., and Frank, D. C.: A Wood Biology Agenda to Support Global Vegetation Modelling, Trends in Plant Science, 23, 1006–1015, https://doi.org/10.1016/j.tplants.2018.08.003, 2018.

Zuidema, P. A., Heinrich, I., Rahman, M., Vlam, M., Zwartsenberg, S. A., and Sleen, P.: Recent CO_2 rise has modified the sensitivity of tropical tree growth to rainfall and temperature, Global Change Biology, 26, 4028–4041, https://doi.org/10.1111/gcb.15092, 2020.