Received: 4 February 2022 | Revised: 1 June 2022 | Accepted: 19 June 2022

DOI: 10.1111/gcb.16384

Global Change Biology WILEY

RESEARCH ARTICLE

Accuracy, realism and general applicability of European forest models

Mats Mahnken[1,2] | Maxime Cailleret[3,4] | Alessio Collalti[5,6,7] | Carlo Trotta[6,7] | Corrado Biondo[6,7] | Ettore D'Andrea[5] | Daniela Dalmonech[5] | Gina Marano[5,8] | Annikki Mäkelä[9] | Francesco Minunno[9] | Mikko Peltoniemi[10] | Volodymyr Trotsiuk[4] | Daniel Nadal-Sala[11,12] | Santiago Sabaté[12,13] | Patrick Vallet[14] | Raphaël Aussenac[14] | David R. Cameron[15] | Friedrich J. Bohn[16] | Rüdiger Grote[11] | Andrey L. D. Augustynczik[17] | Rasoul Yousefpour[18,19] | Nica Huber[8,20] | Harald Bugmann[8] | Katarina Merganičová[21,22] | Jan Merganic[23] | Peter Valent[23] | Petra Lasch-Born[1] | Florian Hartig[24] | Iliusi D. Vega del Valle[1] | Jan Volkholz[1] | Martin Gutsch[1] | Giorgio Matteucci[5] | Jan Krejza[25,26] | Andreas Ibrom[27] | Henning Meesenburg[28] | Thomas Rötzer[29] | Marieke van der Maaten-Theunissen[2] | Ernst van der Maaten[2] | Christopher P. O. Reyer[1]

[1]Potsdam Institute for Climate Impact Research (PIK), Leibniz Association, Potsdam, Germany

[2]Forest Growth and Woody Biomass Production, TU Dresden, Tharandt, Germany

[3]UMR RECOVER, INRAE, Aix-Marseille University, Aix-en-Provence, France

[4]Forest Dynamics Unit, Swiss Federal Research Institute WSL, Birmensdorf, Switzerland

[5]Forest Modelling Lab, National Research Council of Italy, Institute for Agriculture and Forestry Systems in the Mediterranean (CNR-ISAFOM), Perugia, Italy

[6]Department of Innovation in Biological, Agro-Food and Forest Systems (DIBAF), University of Tuscia, Viterbo, Italy

[7]Division Impacts on Agriculture, Forests and Ecosystem Services (IAFES), Fondazione Centro Euro-Mediterraneo sui Cambiamenti Climatici, Viterbo, Italy

[8]Department of Environmental Systems Science, Forest Ecology, Institute of Terrestrial Ecosystems, ETH Zurich, Zurich, Switzerland

[9]Department of Forest Sciences, Institute for Atmospheric and Earth System Research (INAR) and Faculty of Agriculture and Forestry, University of Helsinki, Helsinki, Finland

[10]Natural Resources Institute Finland (Luke), Helsinki, Finland

[11]Institute of Meteorology and Climate Research – Atmospheric Environmental Research (IMK-IFU), Karlsruhe Institute of Technology (KIT), Garmisch-Partenkirchen, Germany

[12]Ecology Section, Department of Evolutionary Biology, Ecology and Environmental Sciences, University of Barcelona (UB), Barcelona, Spain

[13]CREAF (Center for Ecological Research and Forestry Applications), Cerdanyola del Vallès, Spain

[14]LESSEM, INRAE, Univ. Grenoble Alpes, St-Martin-d'Hères, France

[15]UK Centre for Ecology and Hydrology, Penicuik, Midlothian, UK

[16]Helmholtz Centre for Environmental Research – UFZ, Leipzig, Germany

[17]International Institute for Applied Systems Analysis (IIASA), Laxenburg, Austria

[18]Forestry Economics and Forest Planning, University of Freiburg, Freiburg, Germany

[19]Institute of Forestry and Conservation, John Daniels Faculty of Architecture, Landscape and Design, University of Toronto, Toronto, Ontario, Canada

[20]Remote Sensing, Swiss Federal Research Institute WSL, Birmensdorf, Switzerland

[21]Faculty of Forestry and Wood Sciences, Czech University of Life Sciences Prague, Praha, Czech Republic

[22]Department of Biodiversity of Ecosystems and Landscape, Institute of Landscape Ecology, Slovak Academy of Sciences, Nitra, Slovakia

[23]Faculty of Forestry, Technical University in Zvolen, Zvolen, Slovak Republic

[24]Theoretical Ecology, University of Regensburg, Regensburg, Germany
[25]Global Change Research Institute CAS, Brno, Czech Republic
[26]Department of Forest Ecology, Mendel University in Brno, Brno, Czech Republic
[27]Department of Environmental Engineering, Technical University of Denmark (DTU), Lyngby, Denmark
[28]Northwest German Forest Research Institute, Göttingen, Germany
[29]Forest Growth and Yield Science, TU München, Freising, Germany

Correspondence
Mats Mahnken, Potsdam Institute for Climate Impact Research (PIK), Leibniz Association, Telegrafenberg, 14473 Potsdam, Germany.
Email: mahnken@pik-potsdam.de

Funding information
Bavarian Ministry of Science and the Arts, Grant/Award Number: bayklif; COST Action PROCLIAS - European Cooperation in Science and Technology, Grant/Award Number: CA19139; ERA-Net Cofund BiodivClim, Grant/Award Number: 344722; ERA-NET Cofund ForestValue, Grant/Award Number: 22035418 and 773324; European Regional Development Fund; German Federal Ministry Ministry of Education and Research (BMBF), Grant/Award Number: 01LS1711A and 16QK05; Italian National Operational Program for Research and Competitiveness; ALForLab, Grant/Award Number: PON03PE_00024_1; Foundation Euro-Mediterranean Centre on Climate Change

Abstract

Forest models are instrumental for understanding and projecting the impact of climate change on forests. A considerable number of forest models have been developed in the last decades. However, few systematic and comprehensive model comparisons have been performed in Europe that combine an evaluation of modelled carbon and water fluxes and forest structure. We evaluate 13 widely used, state-of-the-art, stand-scale forest models against field measurements of forest structure and eddy-covariance data of carbon and water fluxes over multiple decades across an environmental gradient at nine typical European forest stands. We test the models' performance in three dimensions: *accuracy of local predictions* (agreement of modelled and observed annual data), *realism of environmental responses* (agreement of modelled and observed responses of daily gross primary productivity to temperature, radiation and vapour pressure deficit) and *general applicability* (proportion of European tree species covered). We find that multiple models are available that excel according to our three dimensions of model performance. For the accuracy of local predictions, variables related to forest structure have lower random and systematic errors than annual carbon and water flux variables. Moreover, the multi-model ensemble mean provided overall more realistic daily productivity responses to environmental drivers across all sites than any single individual model. The general applicability of the models is high, as almost all models are currently able to cover Europe's common tree species. We show that forest models complement each other in their response to environmental drivers and that there are several cases in which individual models outperform the model ensemble. Our framework provides a first step to capturing essential differences between forest models that go beyond the most commonly used accuracy of predictions. Overall, this study provides a point of reference for future model work aimed at predicting climate impacts and supporting climate mitigation and adaptation measures in forests.

KEYWORDS
eddy-covariance, gap model, model ensemble, model evaluation, process-based modeling, terrestrial carbon dynamics

1 | INTRODUCTION

Forest models are widely used to assess the impacts of changing environmental conditions such as climate, atmospheric CO_2 concentration and nitrogen deposition on forest functioning, dynamics and structure (e.g., Reyer et al., 2013). Yet, because of our incomplete understanding of forest ecosystems and computational constraints, these models differ in the way specific processes are represented, leading to differences in their predictions (Bugmann et al., 2019; Collalti et al., 2019; Huber et al., 2021). Hence, models need to be comprehensively evaluated using different data types at different spatio-temporal scales before we can judge their structural uncertainties and suitability for answering specific questions (Marechaux et al., 2021; Oberpriller et al., 2021).

Model simulations need to be in adequate agreement with independent observations. Moreover, models have to be sensitive to environmental drivers to ensure that system responses are realistically predicted under a wide range of environmental and climatic

conditions (Collalti et al., 2016). Additionally, for spatially comprehensive assessments of climate impacts, it is also required that the models have a large range of applicability covering different ecological conditions. Ideally, models meet all these requirements.

Levins (1966) categorized these requirements as trade-offs between three dimensions: model accuracy, realism and generality. Accuracy indicates the goodness-of-fit between prediction and observation, realism refers to causally correct internal model processes, and generality represents robust applicability across space and time (Kramer et al., 2002). While it is difficult to maximize accuracy, realism and generality simultaneously, model developers have to identify an optimal point on the trade-off according to the overall aim of the model.

Many climate sensitive forest models have been developed in Europe for different applications, regions and species (e.g., Fontes et al., 2010; Pretzsch et al., 2015). Yet, it is unknown how they perform relative to the same benchmark conditions, and how their structure leads to trade-offs between accuracy, realism and generality since model inter-comparisons across large numbers of complex models are missing. While there is a large body of knowledge from extensive multi-model-data comparisons in North America, especially on carbon and water fluxes (e.g., Medlyn et al., 2015; Schaefer et al., 2012), we lack similar studies for European climate and forest conditions (Table S4). In addition, only few of these evaluation studies include forest structure variables (e.g., LAI: Richardson et al., 2012; biomass: Klesse et al., 2018). Earlier model evaluations have either focused on selected processes (e.g., NPP: Morales et al., 2005; mortality: Bugmann et al., 2019), relied on short time series of observed data (Kramer et al., 2002), or investigated only few models and sites (Horemans et al., 2017). Yet, the increasing amount of harmonized data recently becoming available across Europe (e.g., Reyer et al., 2020a, 2020b) allows for a rigorous evaluation of the state-of-the-art in forest modeling across different biogeographical regions, forest types and types of data. Such an evaluation may provide a deeper understanding of model differences and structural uncertainties, and provide crucial guidance for designing ensemble studies of climate impacts on forests.

The objective of this paper is to evaluate and compare 13 widely applied forest models in managed forests across an environmental gradient in Europe. The models range in complexity from empirically based to highly mechanistic approaches, while the evaluation data types range from ground-based inventories to tower-based eddy-covariance measurements. To achieve this objective, we: (i) compare model outputs to observations to quantify the accuracy of local predictions by deriving the statistical fit between observations and model output of important forest variables; (ii) determine the realism of environmental responses by assessing the agreement of observed and modeled relationships between stand productivity and climatic drivers; (iii) describe the general applicability by deriving the proportion of European forest stands that a model is able to cover; and (iv) integrate these three dimensions in a model performance framework. We hypothesize that trade-offs in our ensemble of forest models can be traced back to differences in accuracy, realism and generality as described by Levins (1966).

2 | MATERIALS AND METHODS

2.1 | Vegetation models and simulation protocol

We used simulation outputs from 13 state-of-the-art, structurally different, forest models (3D-CMCC-FEM LUE, 3D-CMCC-FEM BGC, 3PG, 3PGN-BW, 4C, BASFOR, ForClim v.3.3, FORMIND, GOTILWA+, LandscapeDNDC, PREBAS, SALEM, SIBYLA) that participated in the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP, Frieler et al., 2017; Mahnken et al., 2022). The key assumptions and formulations for simulating processes or variables between models as well as their differences are described in Table 1 (see Table S5 for a comprehensive description). All models are designed to predict long-term (multiple decades) forest growth and forest dynamics. Empirical models are geared towards one full stand rotation while gap models focus on describing successional dynamics in multi-species stands. Mechanistic models describe forest dynamics based on the dynamics of plant carbon and water exchange at a high temporal resolution. Ten of the models describe the ecosystem-atmosphere exchange of carbon, and nine of them describe the exchange of water in forest stands at a daily to annual time step. All 13 models have been applied as research tools to study climate impacts on managed forests.

The simulations followed the ISIMIP phase 2a simulation protocol (https://www.isimip.org/protocol/), which provides a consistent simulation setup based on common, harmonized data for initializing, driving and evaluating models from the PROFOUND database (Reyer et al., 2020a, 2020b). The models were initialized with observed stand characteristics (e.g., stem diameter at breast height, tree height, stand density, stand age) and then driven with locally observed weather data (e.g., surface air temperature, precipitation, vapour pressure deficit), atmospheric CO_2 concentration and nitrogen deposition data, as well as historically observed forest management interventions. Simulated management was based on observed stem numbers and thinning regimes, that is, thinning from above (higher diameter classes preferentially removed) or from below (lower diameter classes preferentially removed). Forest management was the only explicitly simulated disturbance. Drought effects were implicitly included by the driving weather data. The models were run for 13–63 years on nine forest stands across Europe that are contrasting in climate, species composition, phenology, management type and age (Table 2). Not all sites were simulated by all models due to incomplete parameterization for species. Site-specific parameter calibration on the observed data was not permitted.

2.2 | Evaluation data

The PROFOUND database (Reyer et al., 2020a, 2020b) hosts observed data from nine boreal and temperate forest stands located across Europe (Table 2). The database provides measurements of forest structure including basal area (BA), arithmetic mean diameter at breast height (DBH) and arithmetic mean tree height (*H*). On a subset

TABLE 1 Overview of main processes implemented in all forest models as well as examples of model applications

Model	Photosynthesis	Autotrophic respiration	Carbon allocation	Structure development		Mortality	Example applications	Model class
				Height	Diameter			
SALEM	NA	NA	NA	Allometric equation (1)	Diameter, density, and site index specific stand level-dependent increment model (1, 2)	Diameter-dependent specific self thinning (1)	1, 3, 4	E
SIBYLA	NA	NA	NA	Empirical: based on tree age, site specification, tree vitality and competition	Empirical: based on site specification, tree vitality and competition	Empirical: based on tree dimensions, growth and stand density	5, 6, 7	E
ForClim v.3.3	NA	NA	NA	Derived from diameter increment under consideration of light availability and climate specific maximum tree height	Modified carbon budget model (8) considering environmental constraints	Age-related, stress-related	v.3.3: 9; for most recent version v.4.0.1 see 10, 11	H
FORMIND	Light-use efficiency (12)	Maintenance respiration + dynamic growth respiration	Dynamic allocation based on phenology, temperature, light and water availability	Allometric equations	Dependent on carbon allocation to stem mass and current DBH of the tree	Carbon-based stress mortality	13, 14, 15	H
3PG	Light-use efficiency (16)	Constant fraction of GPP	Dynamic allocation based on age, size, soil water, VPD	Allometric equation from DBH, competition, etc.	Dependent on carbon allocation to stem mass and current DBH of the tree	Age-dependent + stress-related + self-thinning	17, 18, 19	H
3PGN-BW	Light-use efficiency (16)	Maintenance respiration + dynamic growth respiration	Dynamic allocation based on environmental modifiers	Allometric equations	Dependent on carbon allocation to stem mass and current DBH of the tree	Age-dependent + stress-related + self-thinning with stochastic component	20, 21	H
BASFOR	Light-use efficiency	Fixed ratio NPP/GPP	Branch and stem fractions constant, leaf and root fractions functions of water- and nitrogen status	Function of stem dry matter	Function of stem dry matter and height	NA	22, 23	H
PREBAS	Light-use efficiency (24, 25)	Maintenance respiration + growth respiration (26)	Dynamic allocation based on pipe-model and functional balance theories and crown allometry	Follows from carbon allocation (27)	Follows from carbon allocation (27)	Competition	28, 29, 30, 31	H

TABLE 1 (Continued)

Model	Photosynthesis	Autotrophic respiration	Carbon allocation	Structure development		Mortality	Example applications	Model class
				Height	Diameter			
3D-CMCC-FEM LUE	Light-use efficiency (16)	Maintenance respiration + dynamic growth respiration (32, 33)	Dynamic allocation based on phenology, light and water availability (sensu 34)	Allometric equations from DBH	Allometric equations from stem biomass	Age-dependent + self-thinning + NSC pool depletion + stochastic component	35, 36, 37, 38	H
3D-CMCC-FEM BGC	Farquhar, von Caemmerer and Berry (39, 40)	Maintenance respiration + dynamic growth respiration (32, 33)	Dynamic allocation based on phenology, light and water availability (sensu 34)	Allometric equations from DBH	Allometric equations from stem biomass	Age-dependent + self-thinning + NSC pool depletion + stochastic component	41, 42	P
4C	Light-use efficiency (12, 43)	Constant fraction of GPP	Dynamic allocation based on pipe-model and functional balance theories	Function of foliage mass and crown architecture	Dependent on carbon allocation	Self-thinning + carbon starvation + age-related (44)	45	P
GOTILWA+	Farquhar (39)	Maintenance respiration + dynamic growth respiration	Dynamic allocation based on pipe-model and functional balance	Allometric equations from DBH	Follows from carbon allocation	NSC pool depletion + loss of active sapwood	46, 47, 48	P
Landscape-DNDC (PSIM)	Farquhar (39)	Maintenance respiration (49) + growth respiration (fixed fraction)	Sink-source approach driven by phenology (50)	Based on stem carbon allocation and density-dependent height:diameter relations (51)	Based on stem carbon allocation and density-dependent height:diameter relations (51)	Fixed fraction + density related limits (52)	53, 54, 55	P

Note: Models are classified according to their complexity into empirical (E), hybrid (H) and process-based (P) types. This classification is based on expert judgment to provide a rough overview of model complexity; in reality, these models align along a continuum from more empirical to more process-based models. References are indicated by numbers. *References*: 1: Aussenac et al. (2021); 2: Toïgo et al. (2015); 3: Vallet and Pérot (2018); 4: Toïgo et al. (2018); 5: Fabrika and Ďurský (2005); 6: Hlásny et al. (2014); 7: Merganic et al. (2020); 8: Moore (1989); 9: Mina et al. (2015); 10: Huber et al. (2020); 11: Huber et al. (2021); 12: Haxeltine and Prentice (1996a); 13: Bohn et al. (2014); 14: Rödig et al. (2017); 15: Bohn et al. (2018); 16: Monteith et al. (1977); 17: Landsberg and Waring (1997); 18: Gupta and Sharma (2019); 19: Trotsiuk et al. (2020); 20: Xenakis et al. (2008); 21: Augustynczik and Yousefpour (2021); 22: van Oijen et al. (2014); 23: Cameron et al. (2013); 24: Mäkelä et al. (2008); 25: Peltoniemi et al. (2015); 26: Mäkelä (1997); 27: Minunno et al. (2019); 28: Kalliokoski et al. (2018); 29: Kalliokoski et al. (2019); 30: Holmberg et al. (2019); 31: Forsius et al. (2021); 32: McCree and Setlick (1970); 33: Thornley (1970); 34: Friedlingstein et al. (1999); 35: Collalti et al. (2014); 36: Collalti et al. (2016); 37: Collalti et al. (2018); 38: Marconi et al. (2017); 39: Farquhar et al. (1980); 40: de Pury and Farquhar (1997); 41: Collalti et al. (2019); 42: Collalti, Tjoelker, et al. (2020); 43: Haxeltine and Prentice (1996b); 44: Botkin et al. (1972); 45: Gutsch et al. (2018); 46: Sabaté et al. (2002); 47: Keenan et al. (2011); 48: Nadal-Sala et al. (2019); 49: Canell and Thornley (2000); 50: Grote (1998); 51: Grote et al. (2020); 52: Grote et al. (2011); 53: Lindauer et al. (2014); 54: Schweier et al. (2017); 55: Dirnböck et al. (2020).

Abbreviations: DBH, diameter at breast height; GPP, gross primary productivity; NA, not included explicitly; NSC, non-structural carbon; VPD, vapour pressure deficit.

TABLE 2 Features of evaluation sites in the PROFOUND database used in this study

Site	Dominant species	Forest type	MAP (mm/year)	MAT (°C)	Elevation (m a.s.l)	Country	Lat.	Long.	Structure	Flux
Hyytiälä	*Pinus sylvestris*	Even-aged	604	4.4	185	FI	61.85	24.23	1995–2011	1996–2014
Solling beech	*Fagus sylvatica*	Even-aged	1113	6.8	500	DE	51.77	9.57	1967–2014	NA
Solling spruce	*Picea abies*	Even-aged	1113	6.8	508	DE	51.77	9.57	1967–2014	NA
Collelongo	*Fagus sylvatica*	Even-aged	1179	7.2	1560	IT	41.85	13.59	1992–2012	1996–2014
Bily Kriz	*Picea abies*	Even-aged	1434	7.4	875	CZ	49.30	18.32	1997–2015	2000–2008
Kroof	*Fagus sylvatica*, *Picea abies*, deciduous species	Mixed	849	8.2	502	DE	48.25	11.4	1997–2010	NA
Sorø	*Fagus sylvatica*	Even-aged	774	9.0	40	DK	55.49	11.65	1997–2013	1996–2012
Peitz	*Pinus sylvestris*	Even-aged	608	9.2	50	DE	51.92	14.35	1948–2011	NA
Le Bray	*Pinus pinaster*	Even-aged	920	13.4	61	FR	44.72	−0.77	1986–2009	1996–2008

Note: lat.: latitude; long.: longitude; Structure: structure variable time coverage; Flux: flux variable time coverage; NA: no flux variable observations.
Abbreviations: MAP, mean annual precipitation; MAT, mean annual temperature.

of five sites, carbon and water fluxes measured at eddy-covariance towers are available (Table 2) including gross primary productivity (GPP), ecosystem respiration (Reco), net ecosystem exchange (NEE) and actual evapotranspiration (AET).

For the carbon flux data, there are multiple products available for the same variable due to varying underlying estimation techniques (Pastorello et al., 2020). We used the data derived with constant friction velocity (USTAR) threshold where the reference is selected based on model efficiency for processing NEE (NEE_CUT_REF; https://fluxnet.org/data/fluxnet2015-dataset/data-processing/, Pastorello et al., 2020) and the daytime (DT) method (Lasslop et al., 2010) for partitioning NEE into GPP and Reco. The first year of carbon flux measurements at each site was discarded since the majority of data points had a quality flag of "poor". Daily AET was derived from measured latent heat flux (LE) to the atmosphere by $AET = LE / \lambda$, with $\lambda = (2.501 - 0.00237 \times T_{air}) \times 10^6$, where T_{air} is the mean daily temperature (Foken, 2008). Annual AET was aggregated as the sum of daily AET derived from the measured daily latent heat flux.

2.3 | Evaluation framework

We evaluated the models in three dimensions based on the framework by Levins (1966) and further specified by Kramer et al. (2002): the accuracy of local predictions, realism of environmental responses and general applicability. We defined the *accuracy of local predictions* as the agreement between observed and predicted data of relevant forest variables at the annual time scale; the *realism of environmental responses* as the agreement of simulated to observed relationships between daily climatic drivers and gross primary productivity; and the *general applicability* as the proportion of European forests a model can represent based on parameterized tree species. In addition to the individual models, we evaluated the model ensemble as the arithmetic mean time series of all individual model predictions available for a given site and variable. We used the statistical computing language R (R Core Team, 2020) for all analyses.

Uncertainty in model predictions arises from model structural uncertainty, parameter uncertainty and input data uncertainty (Collalti et al., 2019; Lindner et al., 2014). Here, we focused on evaluating compound model uncertainty originating from all uncertainty sources except for input data uncertainty, which is shared across all models. The coverage of sites and variables is model-specific and the temporal resolution of model predictions varies from daily to monthly to annual. The models used their individual default species-specific parameter settings for the simulations.

2.3.1 | Accuracy of local predictions

The accuracy of local predictions was quantified for the primary variables of interest on an annual resolution: BA, DBH increment (DBHinc), H increment (Hinc), GPP, Reco, NEE, AET. DBHinc and

Hinc were evaluated instead of DBH and *H* to eliminate the temporal autocorrelation that is associated with these variables, resulting from the incremental nature of diameter and height growth. In this way, we covered increments as well as the structure through BA (which is strongly dominated by temporal autocorrelation). DBHinc and Hinc were computed as the average annual change of stand scale mean DBH and *H*, respectively, for the period between two consecutive observations, since there were no measurements available for every year at all sites and the uncertainty in single year increment measurements is high. The same approach was applied to derive increments from the simulated data. DBHinc and Hinc integrate individual tree increments related to growth as well as changes of the stand scale mean DBH and *H* resulting from the removal of certain trees during management interventions and/or natural tree mortality.

Following Gauch et al. (2003), we computed multiple metrics describing different aspects of the disagreement between predictions and observations. The mean squared deviation (MSD) and its components, squared bias (SB), lack of correlation (LC) and non-unity slope (NU), were computed for each model-site-variable combination. These metrics describe three sources of error: a systematic error (SB), random errors (LC) and linear patterns in the residuals (NU):

$$\mathrm{MSD} = \frac{\sum_{n=1}^{N}(X_n - Y_n)^2}{N} = \mathrm{SB} + \mathrm{NU} + \mathrm{LC}, \tag{1}$$

where X = simulated data, Y = observed data and $n = \{1,2, \ldots N\}$, with N = number of data pairs.

$$\mathrm{SB} = (\overline{X} - \overline{Y})^2, \tag{2}$$

$$\mathrm{NU} = (1-b)^2 \times \left(\frac{\sum_{n=1}^{N} x_n^2}{N}\right), \tag{3}$$

with $b = \sum_{n=1}^{N} x_n y_n / \sum_{n=1}^{N} x_n^2$, which is the slope of the least-square-regression between Y and X. The deviations from the mean are described by $y_n = Y_n - \overline{Y}$ (analogous: $x_n = X_n - \overline{X}$).

$$\mathrm{LC} = (1 - r^2) \times \left(\frac{\sum_{n=1}^{N} y_n^2}{N}\right), \tag{4}$$

with $r^2 = \left(\sum_{n=1}^{N} x_n y_n\right)^2 / \sum_{n=1}^{N} x_n^2 \sum_{n=1}^{N} y_n^2$ which is the square of the correlation between Y and X.

The quantification of these three completely independent components of the MSD allowed us to derive which components drive the inaccuracies most strongly.

For cross-variable and cross-site comparability, we normalized the MSD (norm. MSD; and analogous SB, LC, and NU) with the observed variance of a given variable at a specific site:

$$\mathrm{norm.\,MSD} = \frac{\mathrm{MSD}}{\frac{1}{N}\sum_{n=1}^{N}(Y_n - \overline{Y})^2}, \tag{5}$$

$$\mathrm{norm.\,SB} = \frac{\mathrm{SB}}{\frac{1}{N}\sum_{n=1}^{N}(Y_n - \overline{Y})^2}, \tag{6}$$

$$\mathrm{norm.\,LC} = \frac{\mathrm{LC}}{\frac{1}{N}\sum_{n=1}^{N}(Y_n - \overline{Y})^2}, \tag{7}$$

$$\mathrm{norm.\,NU} = \frac{\mathrm{NU}}{\frac{1}{N}\sum_{n=1}^{N}(Y_n - \overline{Y})^2}. \tag{8}$$

Then, we aggregated the norm. MSD over all sites by computing the arithmetic mean of norm. MSD for a given model-variable combination. To derive a unique accuracy of local predictions score (A) for each model, we first computed the coefficient of determination as R^2 =1 − norm. MSD for each variable (cf. Moffat et al., 2010). Then, we calculated the arithmetic mean of the R^2 values across all structure variables and all carbon and water variables ($R^2_{\mathrm{structure}}$ and $R^2_{\mathrm{carbon\ and\ water}}$) and re-projected the resulting values to the range from 0.1 to 1 to derive $A_{\mathrm{structure}}$ and $A_{\mathrm{carbon\ and\ water}}$. Overall A was then derived analogous to $A_{\mathrm{structure}}$ and $A_{\mathrm{carbon\ and\ water}}$ but with all variables available for a model. The predictive skill of a forest model was higher than the predictive skill of the observed mean in terms of the overall absolute error if norm. MSD <1.

2.3.2 | Realism of environmental responses

The realism of environmental responses was derived by quantifying the agreement of simulated to observed relationships between climatic drivers and productivity, that is, GPP, since GPP is sensitive to several interacting climatic drivers (Zhang et al., 2017, 2019; Zhou et al., 2021). Only those models that output daily GPP could be evaluated for their realism of environmental responses. We considered mean daily temperature (temp), daily global incoming radiation (rad) and daily mean vapour pressure deficit (vpd) as forcing variables on the daily GPP. For each of the five FLUXNET sites, we assessed the realism of the environmental responses for the relation of GPP to temp, rad and vpd of every model. The observations were filtered for FLUXNET quality flags 0 (measured) and 1 (good quality gap-filled). Additionally, the data was filtered for days with temp >5°C (cf. Franklin et al., 2013; Rehfeldt et al., 2006) to ensure that the bulk of the data lie within the growing season, because this is the most important period in which the model needs to exhibit realistic responses of productivity to environmental drivers.

First, we visually compared the form of the observed and simulated relationships between GPP and the three forcing variables including their interactions by deriving general additive models (GAMs) for the 0.5 quantile. We selected the 0.5 quantile (the median) to represent the average response, analogous to regular GAMs. The advantage of using quantile regression is its higher robustness against outliers, which are present in the type of ecological data used here. The computation was done using the R library qgam (Fasiolo et al., 2017). The quantile GAMs have the form

$$\begin{aligned}\mathrm{GPP} = f_1(\mathrm{temp}) + f_2(\mathrm{rad}) + f_3(\mathrm{vpd}) + f_4(\mathrm{temp}, \mathrm{rad}) \\ + f_5(\mathrm{temp}, \mathrm{vpd}) + f_6(\mathrm{rad}, \mathrm{vpd}) + f_7(\mathrm{temp}, \mathrm{rad}, \mathrm{vpd}),\end{aligned} \quad (9)$$

using tensor product (te) smooth functions f_i. We selected the default smoothing parameters, which have been set to generate a reasonable performance on average data (see Wood, 2017), as to not introduce any element of subjectivity into the analysis regarding expected forms of the relationships.

Second, to formally compute model scores for the realism of environmental responses, the residuals between daily simulated and observed *GPP* were derived from the GAMs. We computed simple linear regression models relating the residual daily GPP from the GAMs to each of the three forcing variables. The GAM predictions were obtained by fixing two of the three independent variables to their arithmetic mean value. The linear regressions take the form

$$\mathrm{GPP}_{\mathrm{sim,rad_{fixed},\ vpd_{fixed}}} - \mathrm{GPP}_{\mathrm{obs,rad_{fixed},\ vpd_{fixed}}} = \beta_1 + \alpha_1 \times \mathrm{temp}, \quad (10)$$

$$\mathrm{GPP}_{\mathrm{sim,\ temp_{fixed},vpd_{fixed}}} - \mathrm{GPP}_{\mathrm{obs,\ temp_{fixed},vpd_{fixed}}} = \beta_2 + \alpha_2 \times \mathrm{rad}, \quad (11)$$

$$\mathrm{GPP}_{\mathrm{sim,temp_{fixed},rad_{fixed}}} - \mathrm{GPP}_{\mathrm{obs,temp_{fixed},rad_{fixed}}} = \beta_3 + \alpha_3 \times \mathrm{vpd}. \quad (12)$$

Similar GPP–environment relationships in observed and simulated data were characterized by small residuals, or at least by a lack of patterns in the residuals across the environmental drivers. Hence, small absolute slopes in the linear regression of the residuals indicated an agreement of observed to simulated relationships. For each environmental variable we re-projected the mean absolute slope across all models and sites $\overline{|\alpha_i|}$ to the range between 0 and 1 ($\overline{|\alpha_i|}'$) to account for differences in the magnitude of the variable units (temp: °C; rad: J/cm^2; vpd: kPa). Then, we derived the realism of environmental responses for each model as the mean of the re-projected slope $\left(\frac{\overline{|\alpha_1|}' + \overline{|\alpha_2|}' + \overline{|\alpha_3|}'}{3}\right)$ of these linear regressions.

2.3.3 | General applicability

We interpreted the general applicability of the models as the application range across tree species. As opposed to the accuracy of local predictions and the realism of environmental responses, this quantification was independent of the actual simulations and solely based upon the tree species represented in the models. We computed the share of European forests covered by dominant tree species each model is currently parameterized for. Data on tree species group coverage across Europe were derived from Brus et al. (2011). In case a model covered only subsets of a tree species group (e.g., only *Larix decidua* and not *L. kaempferi* for genus *Larix*), we assumed the forest area of that species group to be covered fully by the model. We only expect a minor overestimation of the area covered by a model because the tree species groups with many species are the ones that are less dominant in Europe. In this way, we derived a rough approximation of the share of European forests where a given model could be applied without considering the actual predictive skill that the model would have in these forests.

2.3.4 | Standardization and aggregation

The results for the accuracy of local predictions, the realism of environmental responses and the general applicability were projected back to a range from 0.1 to 1, which can be interpreted as relative differences across models. We would like to stress that the designation of 0.1 to a model does not indicate a failure or lack of performance but rather that the model had the lowest metric value (relative performance) across the models that were investigated here. We selected 0.1 as the lower boundary simply to avoid misinterpretation that may be intuitively associated with the number zero.

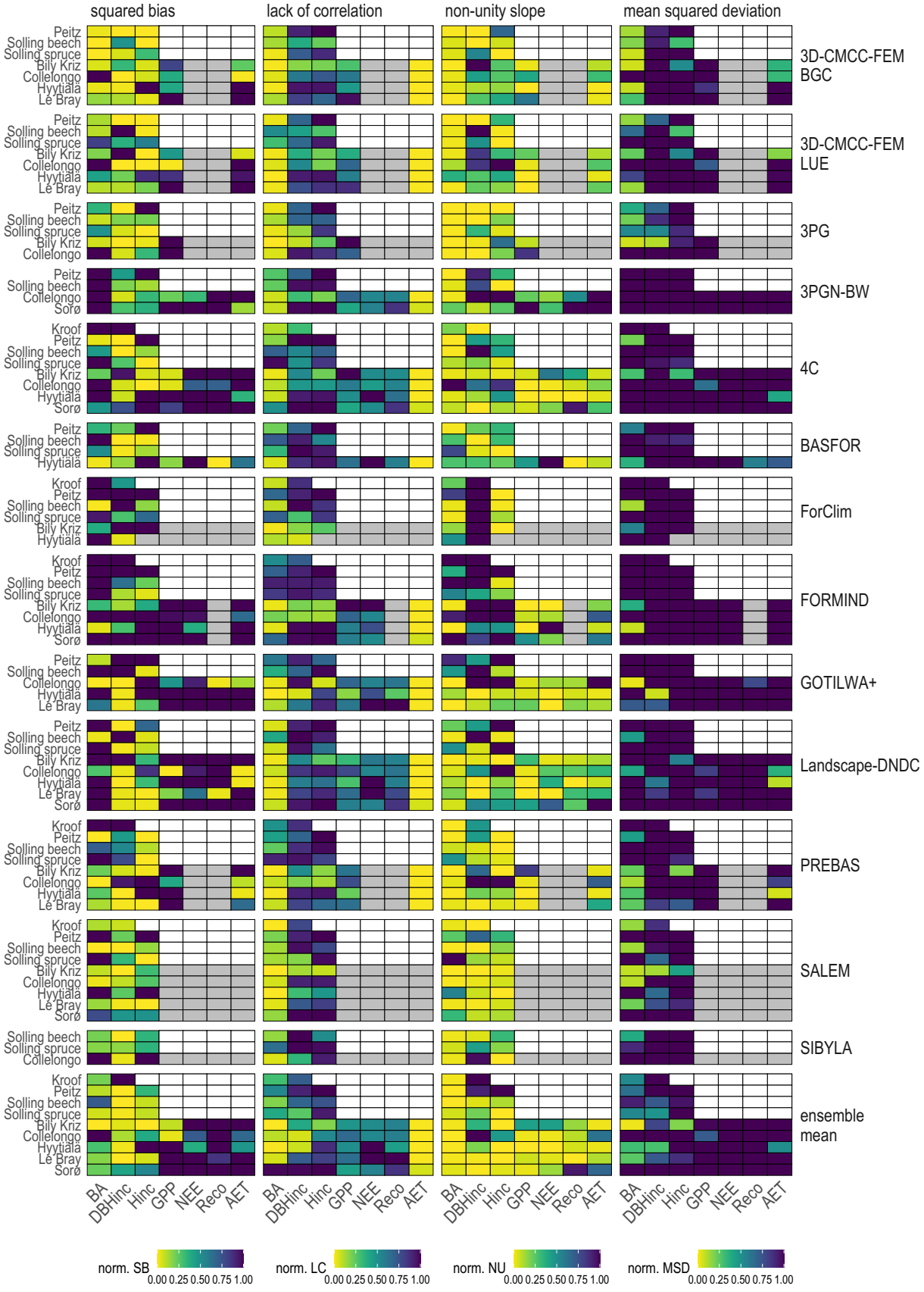
3 | RESULTS

3.1 | Accuracy of local predictions

There was no model that was able to predict all variables at all sites with high accuracy and only few models showed a high accuracy of local predictions for all variables at one site (SALEM at Bily-Kriz, 3PG at Solling-spruce and 3D-CMCC-FEM BGC at Solling-beech). At the same time, every model predicted at least one variable at one site with an adequate accuracy of local predictions except for 3PGN-BW which showed consistently lower predictive skill than the average of observations. (Figure 1).

Partitioning the accuracy differences between models into the three MSD components showed that the offset between model prediction and observed data had varying origins (Figure 1). Random errors (LC) made up the largest share of the overall error except for BA and AET. Systematic errors (SB) of the structure variables may have been a result of offsets in model initialization from the reference data (Figures S4–S9). Flux variables were also prone to SB due to systematic over- or under-estimation. Persistent underestimation of GPP was evident in GOTILWA+ and FORMIND as well as for a range of models at Hyytiälä, while 3PG persistently overestimated

FIGURE 1 Metrics for the accuracy of local predictions for all site-model-variable combinations. On the *y*-axis are the sites, the *x*-axis shows variables, vertical panels are different models and horizontal panels show the different metrics. Colors visualize the normalized metric values, where yellow indicates high agreement and blue indicates low agreement of observed and predicted data. Cells in the column for mean squared deviation (right) in dark blue (norm. MSD ≥1) indicate cases where the observed average has a higher predictive skill than the model predictions. White cells indicate cases with no evaluation data available whereas grey cells indicate cases that are not provided by the model. The model coverage of sites and variables depends on the model application range. norm. LC, normalized lack of correlation; norm. MSD, normalized mean squared deviation; norm. NU, normalized non-unity slope; norm. SB, normalized squared bias.



GPP and Landscape-DNDC overestimated GPP at Bily Kriz. Most models underestimated AET in Le Bray, while overestimation was evident at Bily Kriz (Figures S10–S17). Predicted-observed offsets from linear patterns in the residuals (NU) were generally low except for BA and DBHinc simulated by FORMIND, DBHinc simulated by ForClim v.3.3 as well as Reco and AET for 3PGN-BW.

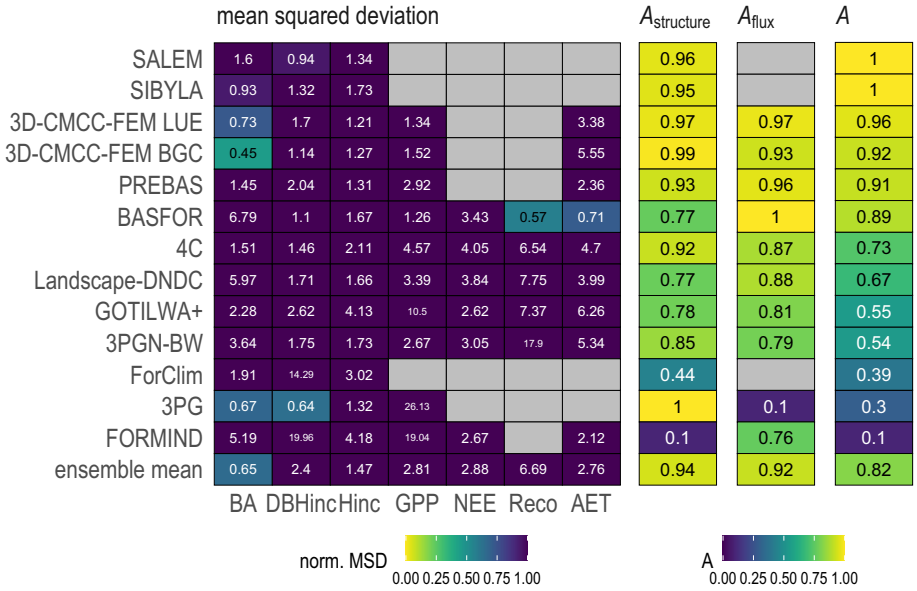


FIGURE 2 Aggregated metrics for accuracy of local predictions for all model-variable combinations assessed (aggregated across sites). Numbers indicate the metric value and colors visualize the normalized metric values, where yellow indicates high agreement and blue indicates low agreement of observed and predicted data.

Forest structure variables displayed a higher overall accuracy of local predictions than the carbon and water variables. On average, simulated BA showed the highest accuracy of local predictions. This is partly related to the temporal autocorrelation of the variable. Annual carbon variables had the lowest accuracy of local predictions, while NEE had the lowest accuracy of the annual carbon variables. No model had a better predictive skill at any site than the observed mean NEE. None of the sites' observed data could be predicted with a high accuracy of local predictions for all carbon and water variables simultaneously by any given model.

The models varied regarding the overall accuracy of local prediction score (*A*, Figure 2). Only few models had a consistently better predictive skill for single variables than the observed mean (norm. MSD <1): SALEM for DBHinc, 3D-CMCC-FEM BGC, 3D-CMCC-FEM LUE and SIBYLA for BA, 3PG for BA and DBHinc and BASFOR for Reco and AET. Although 3PG had a high predictive skill for structure variables, the predictions for GPP had the lowest predictive skill of any model. While some models performed consistently well for one or two variables over multiple sites, other models performed worse than the observed mean for all variable-site combinations. The ensemble mean ranked sixth for accuracy of local predictions of forest structure variables and fourth for carbon and water fluxes. Overall, the ensemble mean had a higher accuracy of local predictions than eight of the individual models.

3.2 | Realism of environmental responses

Observed relationships of daily GPP to temp, rad and vpd followed plausible patterns for all models, while the distinct patterns differed from site to site (Figure 3). Increasing temp and increasing rad were related to increasing daily GPP, except for temp relationship at higher temp values in Bily Kriz, while an increase in vpd was related to decreasing daily GPP. Most models were able to reproduce these observed patterns. Distinct site-specific patterns however were not predicted well at all sites by all models. Strong non-linear patterns were observed for the temp relationship in GOTILWA+ at Collelongo and for the vpd relationship of 4C at Sorø. These patterns may result from outliers in poorly sampled regions in the environmental variable space at the tails of the distribution in combination with model responsiveness to other drivers such as water availability, which was not analyzed here due to the lack of observed data at the sites. Models tended to overestimate daily GPP at high vpd. High daily GPP at high levels of vpd for 4C at Bily-Kriz and Sorø and many models at Le Bray and Hyytiälä indicated unrealistic productivity responses.

The slopes of the linear regressions of the daily GPP residuals (sim. GPP − obs. GPP) to environmental variables indicated varying agreement of observed and simulated environmental responses across models and sites (Table 3; Figure S2). The temp and rad response had the lowest average absolute slope at Le Bray and Sorø had the lowest average absolute slope for vpd (Table S2).

On average, the ensemble mean showed the most realistic environmental responses while Landscape-DNDC and 3D-CMCC-FEM BGC also show more realistic responses of daily GPP to different environmental drivers than other models in our ensemble. Yet, there is no individual model that shows the most realistic responses of GPP to all three environmental variables at all sites. Some models feature intermediate realism of environmental responses to all environmental variables, for example, 3D-CMCC-FEM LUE. The most realistic response to rad was obtained by the ensemble mean. In the ensemble, Landscape-DNDC had the most realistic GPP response to vpd, while GOTILWA+ had the most

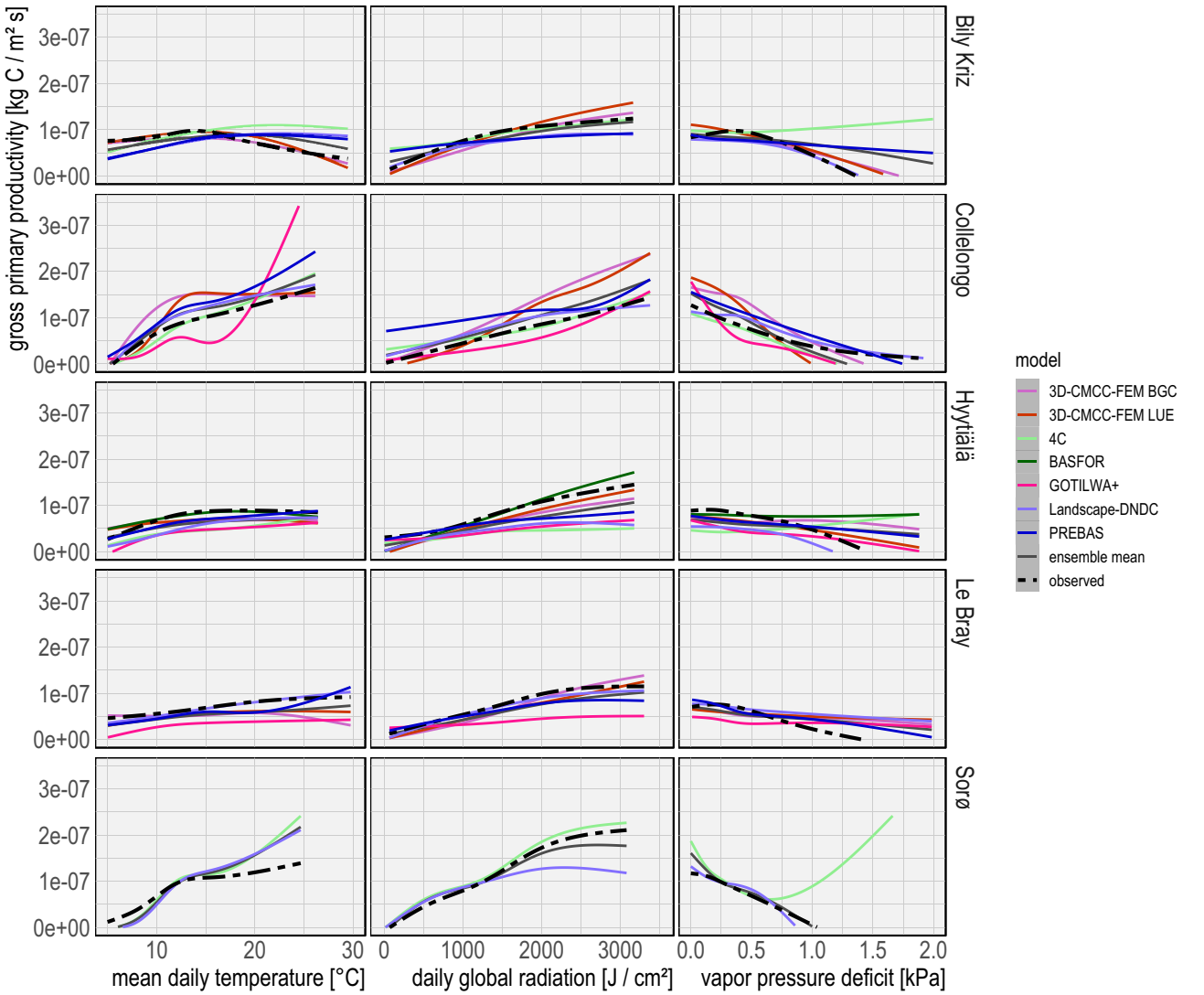


FIGURE 3 Relationship between climate variables and gross primary productivity (GPP) in model simulations and observed flux tower data. Quantile general additive models are displayed (as lines) by fixing two of the three independent variables to their arithmetic mean value.

TABLE 3 Realism of environmental responses per model and environmental variable derived from multiple linear regression slopes of residuals from simulated to observed daily GPP

Model	Mean absolute slope ($\overline{\lvert\alpha_i\rvert}$) (re-projected mean absolute slope, $\overline{\lvert\alpha_i\rvert}'$)			Realism of environmental responses
	temp	rad	vpd	
ensemble mean	1.887×10^{-9} (0.601)	0.913×10^{-9} (0.000)	4.488×10^{-8} (0.511)	1.00
Landscape-DNDC	2.121×10^{-9} (0.716)	1.587×10^{-11} (0.677)	1.686×10^{-8} (0.000)	0.70
3D-CMCC-FEM BGC	1.376×10^{-9} (0.352)	1.396×10^{-11} (0.485)	3.847×10^{-8} (0.612)	0.63
GOTILWA+	0.654×10^{-9} (0.000)	1.909×10^{-11} (1.000)	3.856×10^{-8} (0.615)	0.45
PREBAS	1.602×10^{-9} (0.462)	1.908×10^{-11} (0.998)	2.631×10^{-8} (0.268)	0.33
BASFOR	1.351×10^{-9} (0.340)	1.319×10^{-11} (0.408)	5.215×10^{-8} (1.000)	0.31
3D-CMCC-FEM LUE	1.865×10^{-9} (0.590)	1.412×10^{-11} (0.501)	4.412×10^{-8} (0.772)	0.18
4C	2.705×10^{-9} (1.000)	1.198×10^{-11} (0.286)	3.995×10^{-8} (0.654)	0.10

Note: The mean absolute slope and re-projected mean absolute slope in brackets (see Equations 10–12 and $\overline{\lvert\alpha_i\rvert}$ as well as $\overline{\lvert\alpha_i\rvert}'$ in Section 2) describe the models disagreement between observed and modelled productivity responses to changes in the environmental variable (lower values indicate lower disagreement). The realism of the environmental responses score is the average of $\overline{\lvert\alpha_i\rvert}'$ across environmental variables re-projected to the range 0.1–1 (higher values indicate higher realism of environmental responses). Note that for the models not listed here, the realism of environmental responses was not derived because of missing representation of daily GPP (see Section 2.3.4).

TABLE 4 Tree species groups parameterized in complex forest models as an indicator for the general applicability across European tree species groups

	Abies spp.	*Alnus* spp.	*Betula* spp.	*Carpinus* spp.	*Castanea* spp.	*Eucalyptus* spp.	*Fagus* spp.	*Fraxinus* spp.	*Larix* spp.	Other broadleaves	Other conifers
ensemble mean	X	X	X	X	X	X	X	X	X	X	X
3D-CMCC-FEM BGC	X		X		X		X		X	X	X
3D-CMCC-FEM LUE	X		X		X		X		X	X	X
Landscape-DNDC	X		X			X	X	X	X	X	
ForClim v.3.3	X	X	X	X	X		X	X	X	X	X
3PG	X		X				X	X	X	X	
3PGN-BW	X		X				X	X	X	X	
SALEM	X						X				
4C	X		X			X	X				
FORMIND			X				X	X			
SIBYLA	X						X				
PREBAS			X			X	X				
BASFOR							X				
GOTILWA+						X	X	X			
cover Europe (%)	3.59	1.05	4.12	0.35	0.97	0.44	10.55	0.45	0.20	3.05	0.28

Note: X indicates cases in which the model has a parameterization for at least one species in the species group. Tree species group cover ("cover Europe") indicates the relative share of forest area covered by that species group/model according to Brus et al. (2011). The general applicability per model is the coverage of European forests re-projected to a range of 0.1 to 1 (see Section 2.3.4).

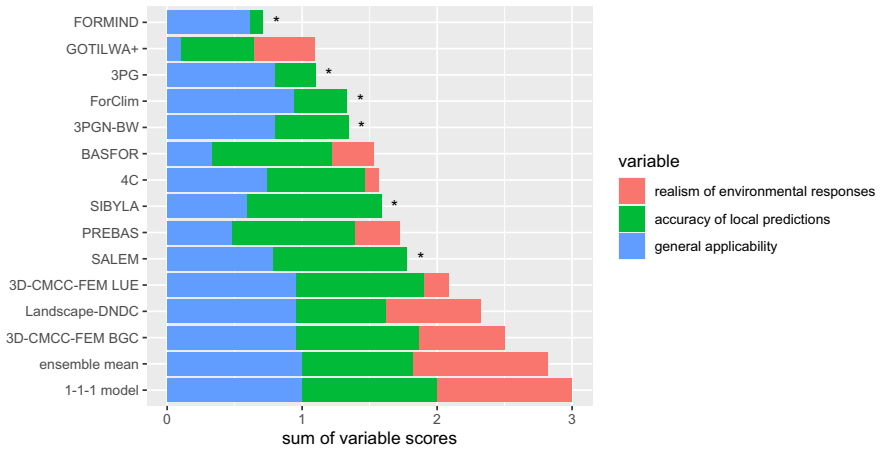


FIGURE 4 Model performance along accuracy of local predictions, realism of environmental responses and general applicability. The highest theoretical total score along three dimensions is 1–1-1 ("1–1-1 model"). "*" note that for SALEM, SYBILA, 3PGN-BW, ForClim v.3.3, 3PG and FORMIND realism of environmental responses could not be calculated. For further information regarding the interpretation of individual metrics, compare Section 2.

realistic GPP response to temp. At the same time GOTILWA+ had the least realistic GPP response to rad, 4C had the least realistic GPP response to temp and BASFOR had the least realistic GPP response to vpd.

3.3 | General applicability

The most common tree species and species groups in Europe are *Pinus sylvestris*, *Picea* spp., *Fagus sylvatica*, and *Q. robur* and *Q. petraea*, which dominate around 75% of Europe's forests (Brus et al., 2011). Almost all models covered these species with species-specific parameterizations. Only PREBAS and BASFOR were missing *Q. robur* and *Q. petraea*, whereas GOTILWA+ was missing *Picea* spp. and *Q. robur* and *Q. petraea*. Additionally, most models covered other species that are less common in Europe; hence, most models had species parameterized that represented the dominant tree species on 73%–98% of Europe's forest cover. The two models covering the least of Europe's forest cover are BASFOR and GOTILWA+ with 66% and 54%. The ensemble mean had the highest general applicability because it combined the species covered by all models. (Table 4).

3.4 | Model performance along the three dimensions of the model performance framework

Besides the analysis of model performance, the accuracy of local predictions, realism of environmental responses and general

Pinus spp.	Other *Quercus* spp.	*Picea* spp.	*Pinus pinaster*	*Pinus sylvestris*	*Populus* spp.	*Pseudotsuga menziesii*	*Quercus robur, Q. petrarea*	*Robinia* spp.	Cover Europe (%)	General applicability
X	X	X	X	X	X	X	X	X	100.0	1.00
X	X	X	X	X			X		97.34	0.95
X	X	X	X	X			X		97.34	0.95
X	X	X	X	X	X	X	X		97.29	0.95
X	X	X		X	X	X	X		96.93	0.94
X		X		X		X	X		90.02	0.80
X		X		X		X	X		90.02	0.80
X	X	X	X	X		X	X		88.88	0.78
X		X		X	X	X	X	X	86.97	0.74
		X		X	X	X	X		80.16	0.61
		X		X			X		78.87	0.59
		X	X	X	X			X	73.38	0.48
		X		X					66.04	0.33
X	X		X	X				X	54.11	0.10
3.17	4.11	22.73	2.57	32.75	0.15	0.16	9.24	0.06		

applicability in isolation, we also analyzed the relations between the three dimensions. Figure 4 shows that the ensemble mean had the highest overall score across the three dimensions. 3D-CMCC-FEM BGC, Landscape-DNDC and 3D-CMCC-FEM LUE performed best across the three dimensions, followed by PREBAS, 4C, BASFOR and GOTILWA+. The models covering only two dimensions of model performance ranked as follows: SALEM, SIBYLA, 3PGN-BW, ForClim v.3.3, 3PG and FORMIND.

4 | DISCUSSION

This study evaluates a large number of complex forest models in an unprecedented model comparison study against a large number of observations: 72 (carbon and water variables) to 128 (forest structure variables) site-years with multiple data sources covering forest structure, carbon and water variables. We developed a model performance framework based on Levins (1966) concept to evaluate accuracy, realism and general applicability of the participating models against this data. Overall, we find that no individual model outperforms the others across all three dimensions and that the model ensemble performs mostly well.

We provide a deeper understanding for model-data mismatches and model applicability in managed European forests that goes beyond currently available knowledge from model intercomparison projects (MIPs). In contrast to manipulatory experiments, such as free-air carbon dioxide enrichment (FACE) MIPs (e.g., De Kauwe et al., 2013, 2014; Medlyn et al., 2015; Walker et al., 2015) and rainfall exclusion/irrigation MIPs (e.g., Paschalis et al., 2020), we evaluate model behaviour against field observations in common managed forests as they are predominant in Europe. Moreover, we not only evaluate carbon and water fluxes such as in eddy covariance MIPs (e.g., Dietze et al., 2011; Huntzinger et al., 2013; Richardson et al., 2012; Schaefer et al., 2012; Stoy et al., 2013; Wei et al., 2014) but also evaluate the forest structure, which is the key target of forest management operations. Likewise, we go beyond comparison of models to tree-ring reconstruction data to evaluate growth (e.g., Klesse et al., 2018; Rollinson et al., 2017, 2021) by assessing BA, DBHinc and Hinc, although on shorter time scales.

4.1 | Model performance

4.1.1 | Accuracy of local predictions

3PG and 3D-CMCC-FEM BGC simulate the structure variables most accurately, while BASFOR and 3D-CMCC-FEM LUE do so for the carbon and water variables. The main difference between 3D-CMCC-FEM BGC and 3D-CMCC-FEM LUE is the representation of photosynthesis (Table 1), with the BGC version featuring a more process-based approach. The BGC version performs better for the structure variables than the LUE version, while the LUE version is more accurate than the BGC version regarding carbon flux variables at the annual scale. This unexpected trade-off cannot be explained in a straight-forward manner by the differences in the model versions, but indicates that more empirical photosynthesis models (LUE

version) do not necessarily produce less accurate predictions of annual flux variables than more process-based approaches (BGC version). 3PG is rather simple compared to the other models applied here (Table 1), but it still produces accurate predictions of DBHinc for the subset of sites in this study that are truly monospecific and even-aged. Apparently, less detailed but more robust model formulations are an advantage when simulating these types of forests. Likewise, the other models that focus on forest dynamics alone rather than also simulating biogeochemical fluxes, such as SALEM and SIBYLA, also show a high accuracy of local predictions for structure variables. The development of forest structure in these more empirical models is based on more empirically based formulations (i.e., allometric functions) while the other models' structure development emerges from a combination of carbon allocation to different biomass compartments and allometric functions (Table 1). While the more empirically based formulations simulate highly accurate developments of forest structure, the accuracy of local predictions for structure variables is more heterogeneous across models with tree structure emerging from carbon allocation. Hence, the specific model formulation of how carbon is allocated to form structure is important. Nevertheless, in the more complex models also, other processes interact with the carbon available for structure development, for example, phenology and the linked total amount of sequestered carbon. ForClim v.3.3 and FORMIND show a lower accuracy of local predictions for structure variables mainly because the predictions of DBHinc have a large offset to observations. These offsets result from the simulated thinning regime and, in the case of ForClim v.3.3, a bias in the allocation (which has been addressed in v4.01, Huber et al., 2020). Low accuracy of BA among all models may be explained by simulated mortality reducing stand density below the observed stem numbers (Figure S9). BASFOR, which is also among the less complex models of our ensemble, produces accurate predictions of carbon and water variables while it predicts the structure variables with low accuracy. Such systematic errors regarding structure variables may also result from specifics in model initialization (Figures S4–S9), for example, BASFOR initialized trees with a planting procedure while most models were initialized with observed data of adult stands. In models that operate at the forest stand-scale rather than the tree level, systematic errors may also arise from the underestimation of BA if it is calculated internally from a multimodal DBH distribution and stem number. For example, Landscape-DNDC and 3PGN-BW initialized mean DBH assuming a mean weighted by basal area and not an arithmetic mean, leading to systematically higher BA, DBH and H (but not growth) at sites with a heterogenous diameter distribution as is the case in particular in Sorø. Finally, the systematic over- as well as underestimation of flux variables shown by most models at least for some sites may be an effect of an insensitivity for specific environmental conditions defined by either model structure or the generic parameter sets used in this study.

Generally, the models predicted structure variables more accurately than annual carbon and water variables, except for BASFOR and FORMIND. Earlier findings by Kramer et al. (2002) and Morales et al. (2005) suggested that forest models have an adequate accuracy regarding daily carbon and water fluxes. Yet, on the multi-annual time scale, Horemans et al. (2017) found larger uncertainties for NEE than on the daily time scale. Our findings using a much larger ensemble of models confirm these earlier findings. Carry-over effects from preceding years, which are usually not well represented in models, may be a reason for the inaccurate year-to-year variation of carbon fluxes in the models (Aubinet et al., 2018).

Moreover, besides the reasons for individual model-data mismatches discussed above, the quality of the observed data may affect all models collectively. Systematic and unsystematic observation errors affect the reference data to which the models are compared to, for example, uncertainty from the method used to partition NEE into GPP and Reco (Oikawa et al., 2017). Checking the agreement of estimates from these different methods, we found that GPP estimated with the DT partitioning method (Lasslop et al., 2010) is highly correlated with GPP estimated with the nighttime method (NT, Reichstein et al., 2005) in the evaluation data with no apparent bias (Figure S3). Consequently, using DT- or NT-based GPP estimates led to only minor changes in the results. Moreover, abiotic or biotic disturbances that affect the reference data but are not represented in model simulations may affect model accuracy (Finzi et al., 2020; Trugman et al., 2021). Furthermore, the understory contribution to the carbon balance was not assessed in any of the models but contributes to the measured carbon balance (Dirnböck et al., 2020).

Additionally, uncertainties in model forcing data may contribute to model-data mismatches. For example, the climate data used to drive the simulations was sometimes observed at or close to the forest stand, but in some cases only inferred from the nearest climate station (Reyer et al., 2020b), which may introduce additional uncertainties, for example, due to orographic effects. Likewise, even though the stands are managed using standard silvicultural treatments (Reyer et al., 2020b), specific, local forest management actions may not be perfectly covered by the models' approximation of the management.

Overall, we find that simpler models, like SALEM, SIBYLA, 3PG, BASFOR and PREBAS did not necessarily perform worse than more complex models like 3D-CMCC-FEM BGC, 3D-CMCC-FEM LUE, 4C, Landscape-DNDC or GOTILWA+. The ensemble mean has an intermediate overall accuracy. Hence, in most cases there are more accurate individual models available for each site-variable combination. Moreover, the range of annual model predictions did not always overlap with observations. Hence, assessing the range of the model ensemble and assuming that the "true" value lies within that range is not always advisable. This was most pronounced for Hinc at Hyytiälä, Le-Bray, Solling-beech, Solling-spruce and Sorø, Reco at Collelongo and Sorø, NEE at Collelongo, Bily-Kriz and Sorø as well as DBHinc, GPP and AET at Le Bray. Hence, in some cases all models overestimate or underestimate the observed data, which points either to general issues in model structure and/or parameterization across all models, or it may relate to issues with the reference data outlined above. Identifying the specific reasons for the systematic mismatch at these sites for these variables is challenging. However,

it may be related to the management at the sites and specific site properties that are not reflected in the models. For example, a mismatch in the modelled to observed size distribution of removed trees during management has a large effect on the accuracy of local predictions of DBHinc and Hinc. Other site properties, such as large amounts of downed woody debris (e.g., Collelongo as described by Morales et al., 2005) may influence the carbon balance in reality but are not reflected in the models.

4.1.2 | Realism of environmental responses

Earlier findings by Kramer et al. (2002) showing realistically simulated relationships of daily GPP to daily mean temperature and global radiation can be confirmed by our large ensemble. In addition, we find that models exhibit also realistic responses of GPP to vpd. Properly capturing GPP responses to vpd has proven to be fundamental to reproduce annual productivity patterns (Medlyn et al., 2011), especially in stands where the most limiting environmental driver for GPP shifts from water availability to vpd along the year (e.g., Nadal-Sala et al., 2021), and given that vpd-driven limitation of productivity is expected to increase under global warming (e.g., Novick et al., 2016). In this regard, our lumped GAM analysis is not able to fully determine the exact driver that is limiting GPP at a given moment, and therefore, interactive effects of constraining environmental drivers cannot be fully discarded. Hence, the impact of vpd on GPP for each individual model remains unassessed, with the realism of this key response potentially being masked by its positive correlation with temperature and radiation.

While 3D-CMCC-FEM BGC shows relatively realistic daily GPP response, the closely related model 3D-CMCC-FEM LUE has the second least realistic GPP response. The single difference between these two models is the description of photosynthesis that is more process-based for 3D-CMCC-FEM BGC, which used the Farquhar, von Caemmerer and Berry biochemical photosynthesis model (Farquhar et al., 1980) and the 3D-CMCC-FEM LUE, which uses the Monteith empirical approach (Monteith et al., 1977). While the BGC version shows more realistic daily environmental responses of GPP, the LUE version is more accurate at the annual scale. Since the BGC version was constructed to provide daily estimates of productivity while the LUE version was originally constructed to provide estimates at the monthly time scale, and compensating for possible over and under estimations, this performance relation can be expected. Biases originating from missing site-specific calibration and, given the higher number of parameters in biochemical photosynthesis models, increased uncertainty in the daily outputs of the BGC version could explain the worse performance at the annual scale. The issue related to the temporal scale in modeling GPP has already been discussed by Collalti et al. (2016) and Lasch-Born et al. (2020).

Overall, the individual models complemented each other with regard to the realism of environmental responses of productivity. On average, the ensemble mean produced more realistic daily GPP responses to environmental variables than any of the individual models. This is due to overestimating and underestimating individual models that cancel out when aggregated into an ensemble mean. Nevertheless, the ensemble mean's performance relative to individual models strongly depends on whether the underlying models are balanced (over- as well as underestimation) and represent different model structures.

4.1.3 | General applicability

Following our rather simple definition of the general applicability of models, we find that most of the models are able to simulate a relatively large share of European forests. However, simply being able to simulate tree species or plant functional types does not warrant that models are able to simulate all potential mixtures, site conditions or management systems (Bravo et al., 2018; Grote et al., 2011; Pretzsch et al., 2015). Still, it is encouraging to see that the models generally cover the main species that are currently of commercial and ecological relevance in Europe, and hence from this point of view, most models are suitable to be applied in climate impact studies covering different European forests. The ensemble covers almost all European forest tree species because the individual models complement each other especially for the less common tree species.

However, as forests may become more species rich and structurally complex in the future as part of forest adaptation to climate change (de Wergifosse et al., 2022; Huber et al., 2020; Pardos et al., 2021) the general applicability of the models may be further challenged. Additionally, the relative importance of tree species may shift in the future because of altered climatic conditions (Buras & Menzel, 2019). Although the most important European species in projected future abundance are already covered by the models (*P. sylvestris*, *Picea abies*, *Quercus* spp., *Fagus sylvatica*), shifting disturbance regimes may reinforce the species abundance shift. In that case, models may need to include species that are less abundant today, hence rarely parameterized, but may become more abundant in the future.

4.1.4 | Trade-offs between the three dimensions of the model performance framework

Even though our framework of model performance does not theoretically prevent models from scoring high in all three dimensions, we did not expect that any model would do so, but that trade-offs between accuracy of local predictions, realism of environmental responses and general applicability were present. While our results confirm that there is no “silver bullet”, we could not find explicit trade-offs such as a systematic negative relation between general applicability and accuracy of local predictions either. Models that have a high general applicability score such as 3D-CMCC-FEM BGC also perform well in terms of accuracy of local predictions and realism of environmental responses. In general, the scores of the three dimensions of model performance seem to be balanced for most

models although at different overall levels. One of the exceptions is the model GOTILWA+, which has a relatively low score for accuracy of local predictions but a comparably high score for realism of environmental responses. Such results may originate from parameter uncertainties in the initial model setup, as physiological and allometric parameters for a given species have not been calibrated, though they have been observed to be highly site-dependent (e.g., allometric and photosynthetic parameters) and varying also with forest developmental stages (Collalti et al., 2019). Also, the lack of trade-offs between accuracy of local predictions, realism of environmental responses and general applicability may be an artifact of the way we derived the realism of environmental responses. The potential trade-off in the framework provided by Levins (1966), and further elaborated by Weisberg (2007), may not be apparent in the suggested framework here, because we did not strictly follow the definitions of accuracy, realism and generality since they are inherently difficult to assess and not meant to be operationalized for actual simulation models. Operationalizing the framework for complex forest models may have distorted the relation between the three dimensions as defined by Levins (1966). Furthermore, although a balance between the three dimensions is advisable, it may not always be necessary. For example, qualitatively correct insights about forest growth and dynamics under global change may be sufficient to guide adaptation planning, for example, insights about the growth dominance of one species over the other, indicating that realism and generality may be more important for this purpose than accuracy.

Another key aspect that might explain the differences in performance among models is that some models were initially developed for other scopes. Some models have been developed to simulate forest growth and fluxes in the short-term (i.e., the variables of interest here), but others to simulate forest growth and demography over the medium- to long-term (decadal to centennial) and, thus, focusing more on processes such as reproduction and mortality (not analyzed here). For instance, a specific strategy for model development in ForClim is that each model development step should lead to better predictions of long-term (centennial) forest dynamics and/or of potential natural vegetation (simulations over >1000 years) (Didion et al., 2009). Testing for these model capabilities would probably lead to a different model ranking than presented here. Furthermore, some models have been developed with the primary aim to capture multi-decadal dynamics in complex multi-species stands (e.g., SIBYLA, FORMIND, ForClim), but eight of the nine stands used here were rather homogenous single-species stands (Table 2), which may be, in theory, easier to simulate using mechanistic biogeochemistry models.

4.2 | Limitations of the model performance framework

Most model evaluation studies to date have assessed the accuracy of local predictions (e.g., Irauschek et al., 2021). Yet, in addition to the agreement of predicted and observed variables of primary interest, complementary evaluation procedures may be implemented for a more comprehensive assessment of the models (see Wagener et al., 2022). Realistic secondary patterns, such as the responses of productivity to environmental drivers are crucial, especially when assessing models that are being used for climate impact studies. Likewise, given the rapid expansion of model uses and users, the general applicability is important to help the latter to assess whether the model is likely to be useful for comprehensive impact studies across a large range of tree species. Our model performance framework is a first attempt to operationalize Levins' (1966) ideas within the context of climate impact assessments with complex vegetation models.

Our approach for quantifying the accuracy of local predictions is a robust way for assessing the agreement of predicted-observed data for models with different numbers of variable outputs. Models that provide more output variables for assessment in the performance framework are not necessarily less accurate. Nevertheless, those models that assess variables which are generally more difficult to accurately predict will have lower levels of accuracy than those models only assessing variables that are less difficult to predict. Future applications of the framework could explore different weightings of the variables depending on the difficulty in predicting them and the availability of data to test them. Furthermore, we acknowledge that model predictions are also useful if they have less predictive skill than the observed mean because there are many instances where no data are available to derive the mean for a given variable. Here, we used the observed mean as threshold to identify especially well performing models and not to penalize poorly performing models.

Besides an accurate representation of historical data, forest models should be characterized by a realistic response of productivity to environmental drivers under varying climatic conditions. However, to assess model realism more comprehensively, all processes represented in the model need to be assessed, rather than only the productivity response (see also Huber et al., 2020). Therefore, even though we test the models with carbon and water variables, further refinements of the model performance framework should include testing other variables for their realism to environmental responses such as structure and mortality variables or autotrophic and soil respiration to test model realism across a broader range of processes. Likewise, model comparisons in which the models have been forced to mimic experimental changes in environmental variables such as shifting of atmospheric CO_2 concentrations in FACE experiments (Walker et al., 2021; Zaehle et al., 2014) or rainfall manipulation experiments (Paschalis et al., 2020) could help us to learn further about the model's realism of environmental responses. Whether the model includes flexible traits (Berzaghi et al., 2019) and whether it is able to mimic natural adaptive processes (Collalti, Ibrom, et al., 2020) could be a further element of testing the realism of environmental response. Moreover, the quantification of realism could be restricted to periods when one environmental driver (e.g., temperature, radiation or vapour pressure deficit) is driving the GPP response as to not confound interacting effects of different environmental drivers

(e.g., Nadal-Sala et al., 2021). Additionally, models that assume identical allometric relationship for a single species regardless of environmental conditions, are expected to be less accurate than models accounting for site differences by different allometric coefficients or incorporating environmental drivers (Cysneiros et al., 2021). Moreover, evaluating process rates (e.g., GPP) in contrast to model states (e.g., BA) requires a higher realism of environmental responses to produce accurate predictions, because model states are dominated more strongly by long-term model assumptions on stand dynamics (such as mortality definitions, carbon allocation, allometric relationships, management regime). Overall, to test realism properly, one should test the response of the models to different forcing conditions, and compare the (qualitative) responses of the models to our general understanding of the processes and observed data describing these responses.

Generality, as the robust model applicability across space and time, is challenging to assess since extensive data are needed to apply and evaluate models across large spatial and temporal scales. We did not derive the general applicability across time but focused on the general applicability in space. In addition to the quantification of temporal generality, information on whether the models are able to simulate mixed forests with a complex structure, comprising both managed and natural dynamics, could be used to widen the presented general applicability metric.

Finally, because we investigated the model performance based on current model parameterization without further site specific parameter calibration, the resulting uncertainty is originating from both model structure and model parameterization. The model performance is reflecting the current state of the model only. However, model parameterization and calibration have the potential to increase the performance along all three dimensions of the model performance framework. In theory, if a model is general in its structure (i.e., more process-based models), it would need less data to be parameterized to different environments or species, if it is less general (i.e., more empirical models), it would need more data. Hence, the effort that is needed for calibrating a model to specific environments is model specific and different calibration efforts would lead to varying levels of improvement of the three dimensions of model performance. But not all three dimensions are dependent on model structure and parameterization to the same extent. The realism of environmental responses is mostly driven by model structure, accuracy of local predictions is affected by both model structure and parameterization while the general applicability is mostly dependent on the model parametrization effort. In summary, the current model performance can be improved not only by development of the model structure itself but also by model parameter calibration.

4.3 | Conclusions and implications for model applications

We performed a large forest model comparison with a wide range of multi-source evaluation data in an innovative model performance framework that complements existing knowledge from model-model and model-data comparisons. We found that the accuracy of local predictions in the historical period is not related to the level of complexity of a model; that is, empirical models do not necessarily provide less accurate predictions than hybrid or process-based models under current climate conditions. Furthermore, accurate predictions of carbon variables at annual scale are more difficult to obtain than accurate predictions of structure variables. The realism of environmental responses in model simulations provides an approximation how well relationships that are crucial to assessing climate impacts are covered. We showed that the model ensemble mean has the most realistic daily GPP responses to environmental variables. General applicability, in terms of the coverage of European tree species is high for most models but less common species that may become more important under climate change are only partly covered by models.

We conclude that, if accuracy is the objective, individual models may provide the best results at single specific locations. Which model will provide optimal results depends on the environmental conditions, structural properties, disturbances, etc. of those locations. Moreover, most individual models cover the most relevant European tree species, but to cover all and particularly the less abundant species, multiple models need to be applied. Finally, we highlight the importance to evaluate several model output variables with a wide range of data, because models struggle to achieve high accuracies for several variables at the same time. Because already multiple models exist to study climate impacts on forests, we expect that our study will provide a common benchmark to test whether new modelling efforts outperform the models presented here to add value to the existing set of tools.

ACKNOWLEDGMENTS

MM, MG, PV and RA acknowledge financial support from I-Maestro (Innovative forest management strategies for a resilient bioeconomy under climate change and disturbances, grant nos. 773324 and 22035418, 2019–2022) funded by the ERA-NET Cofund ForestValue. AC, CB, CT, GM and DD thank Foundation Euro-Mediterranean Centre on Climate Change and the ALForLab (PON03PE_00024_1) project co-funded by the Italian National Operational Program for Research and Competitiveness (PON R&C) 2007–2013, through the European Regional Development Fund (ERDF) and national resource (Revolving Fund—Cohesion Action Plan [PAC]) MIUR, for their support in model development and data analysis. This article is further based upon work from COST Action CA19139 PROCLIAS (PROcess-based models for CLimate Impact Attribution across Sectors), supported by COST (European Cooperation in Science and Technology; https://www.cost.eu), the ERA4CS Joint Call on Researching and Advancing Climate Services (ISIpedia; BMBF grant 01LS1711A) and the German Federal Ministry of Education and Research (BMBF) under the research project ISIAccess (BMBF grant 16QK05). FH acknowledges funding by the Bavarian Ministry of Science and the Arts in the context of Bavarian Climate Research Network

(bayklif). MP acknowledges support from BiodivClim ERA-Net Cofund with Academy of Finland (no. 344722) for the project Funpotential. We are further grateful to all the researchers providing data to the PROFOUND Database and for long-term funding of the stations from national and international sources and networks. In addition, we thank Dario Papale for initiating and continually supporting the post-processing and harmonization of the legacy flux and meteorological data provided by the international flux database FLUXNET. Finally, we thank Dr. Felicitas Suckow and Dr. Chris Kollas for help with 4C development and application, and all data providers for establishing a database that enables thorough model evaluations.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The reference data that support the findings of this study are publicly available in GFZ Data Services at https://doi.org/10.5880/PIK.2020.006. The simulation data that support the findings of this study are publicly available in the ISIMIP repository at https://doi.org/10.48364/ISIMIP.169780.

ORCID

Mats Mahnken https://orcid.org/0000-0002-9755-8814
Maxime Cailleret https://orcid.org/0000-0001-6561-1943
Alessio Collalti https://orcid.org/0000-0002-4980-8487
Carlo Trotta https://orcid.org/0000-0001-6377-0262
Ettore D'Andrea https://orcid.org/0000-0002-5884-210X
Daniela Dalmonech https://orcid.org/0000-0002-1932-5011
Gina Marano https://orcid.org/0000-0003-2600-984X
Annikki Mäkelä https://orcid.org/0000-0001-9633-7350
Francesco Minunno https://orcid.org/0000-0002-7658-6402
Mikko Peltoniemi https://orcid.org/0000-0003-2028-6969
Volodymyr Trotsiuk https://orcid.org/0000-0002-8363-656X
Daniel Nadal-Sala https://orcid.org/0000-0002-0935-6201
Santiago Sabaté https://orcid.org/0000-0003-1854-0761
Patrick Vallet https://orcid.org/0000-0003-2649-9447
Raphaël Aussenac https://orcid.org/0000-0003-1191-4716
David R. Cameron https://orcid.org/0000-0001-8938-0908
Friedrich J. Bohn https://orcid.org/0000-0002-7328-1187
Rüdiger Grote https://orcid.org/0000-0001-6893-6890
Andrey L. D. Augustynczik https://orcid.org/0000-0001-5513-5496
Rasoul Yousefpour https://orcid.org/0000-0003-3604-8279
Nica Huber https://orcid.org/0000-0001-5427-6836
Harald Bugmann https://orcid.org/0000-0003-4233-0094
Katarina Merganičová https://orcid.org/0000-0003-4380-7472
Jan Merganic https://orcid.org/0000-0001-6905-8356
Peter Valent https://orcid.org/0000-0001-7017-6640
Petra Lasch-Born https://orcid.org/0000-0001-6468-4411
Florian Hartig https://orcid.org/0000-0002-6255-9059
Iliusi D. Vega del Valle https://orcid.org/0000-0001-6902-2257
Jan Volkholz https://orcid.org/0000-0002-2533-3739
Martin Gutsch https://orcid.org/0000-0001-7109-273X
Jan Krejza https://orcid.org/0000-0003-2475-2111
Andreas Ibrom https://orcid.org/0000-0002-1341-921X
Henning Meesenburg https://orcid.org/0000-0002-3035-4737
Thomas Rötzer https://orcid.org/0000-0003-3780-7206
Marieke van der Maaten-Theunissen https://orcid.org/0000-0002-2942-9180
Ernst van der Maaten https://orcid.org/0000-0002-5218-6682
Christopher P. O. Reyer https://orcid.org/0000-0003-1067-1492

REFERENCES

Aubinet, M., Hurdebise, Q., Chopin, H., Debacq, A., De Ligne, A., Heinesch, B., Manise, T., & Vincke, C. (2018). Inter-annual variability of net ecosystem productivity for a temperate mixed forest: A predominance of carry-over effects? *Agricultural and Forest Meteorology*, *262*, 340–353. https://doi.org/10.1016/j.agrformet.2018.07.024

Augustynczik, A. L. D., & Yousefpour, R. (2021). Assessing the synergistic value of ecosystem services in European beech forests. *Ecosystem Services*, *49*, 101264. https://EconPapers.repec.org/RePEc:eee:ecoser:v:49:y:2021:i:c:s221204162100022x

Aussenac, R., Pérot, T., Fortin, M., de Coligny, F., Monnet, J., & Vallet, P. (2021). The Salem simulator version 2.0: A tool for predicting the productivity of pure and mixed forest stands and simulating management operations [version 2; peer review: 2 approved]. *Open Research Europe*, *1*(61), 1–40. https://doi.org/10.12688/openreseurope.13671.2

Berzaghi, F., Wright, I. J., Kramer, K., Oddou-Muratorio, S., Bohn, F. J., Reyer, C. P. O., Sabate, S., Sanders, T. G. M., & Hartig, F. (2019). Towards a new generation of trait-flexible vegetation models. *Trends in Ecology & Evolution*, *35*, 191–205. https://doi.org/10.1016/j.tree.2019.11.006

Bohn, F. J., Frank, K., & Huth, A. (2014). Of climate and its resulting tree growth: Simulating the productivity of temperate forests. *Ecological Modelling*, *278*, 9–17. https://doi.org/10.1016/j.ecolmodel.2014.01.021

Bohn, F. J., May, F., & Huth, A. (2018). Species composition and forest structure explain the temperature sensitivity patterns of productivity in temperate forests. *Biogeosciences*, *15*(6), 1795–1813. https://doi.org/10.5194/bg-15-1795-2018

Botkin, D. B., Janak, J., & Wallis, J. R. (1972). Some ecological consequences of a computer model of forest growth. *Journal of Ecology*, *60*, 849–872.

Bravo, F., Fabrika, M., Ammer, C., Barreiro, S., Bielak, K., Coll, L., Fonseca, T., Kangur, A., Löf, M., Merganičová, K., Pach, M., Pretzsch, H., Stojanović, D., Schuler, L., Peric, S., Rötzer, T., Río, M., Dodan, M., & Bravo-Oviedo, A. (2018). Modelling approaches for mixed forests dynamics prognosis research gaps and opportunities. *Forest Systems*, *28*(1), 18. https://doi.org/10.5424/fs/2019281-14342

Brus, D. J., Hengeveld, G. M., Walvoort, D. J. J., Goedhart, P. W., Heidema, A. H., Nabuurs, G. J., & Gunia, K. (2011). Statistical mapping of tree species over Europe. *European Journal of Forest Research*, *131*(1), 145–157. https://doi.org/10.1007/s10342-011-0513-5

Bugmann, H., Seidl, R., Hartig, F., Bohn, F., Bruna, J., Caillaret, M., Francois, L., Heinke, J., Henrot, A.-J., Hickler, T., Hülsmann, L., Huth, A., Jacquemin, I., Kollas, C., Lasch-Born, P., Lexer, M. J., Merganic, J., Merganicová, K., Mette, T., … Reyer, C. P. O. (2019). Tree mortality submodels drive simulated long-term forest dynamics: Assessing 15 models from the stand to global scale. *Ecosphere*, *10*(2), e02616.

Buras, A., & Menzel, A. (2019). Projecting tree species composition changes of European forests for 2061–2090 under RCP 4.5 and

RCP 8.5 scenarios. *Frontiers in Plant Science*, *9*, 1986. https://doi.org/10.3389/fpls.2018.01986

Cameron, D. R., Van Oijen, M., Werner, C., Butterbach-Bahl, K., Grote, R., Haas, E., Heuvelink, G. B. M., Kiese, R., Kros, J., Kuhnert, M., Leip, A., Reinds, G. J., Reuter, H. I., Schelhaas, M. J., De Vries, W., & Yeluripati, J. (2013). Environmental change impacts on the C- and N-cycle of European forests: A model comparison study. *Biogeosciences*, *10*(3), 1751–1773. https://doi.org/10.5194/bg-10-1751-2013

Canell, M. G. R., & Thornley, J. H. M. (2000). Modelling the components of plant respiration: Some guiding principles. *Annals of Botany*, *85*(1), 45–54.

Collalti, A., Ibrom, A., Stockmarr, A., Cescatti, A., Alkama, R., Fernández-Martínez, M., Matteucci, G., Sitch, S., Friedlingstein, P., Ciais, P., Goll, D. S., Nabel, J. E. M. S., Pongratz, J., Arneth, A., Haverd, V., & Prentice, I. C. (2020). Forest production efficiency increases with growth temperature. *Nature Communications*, *11*(1), 5322. https://doi.org/10.1038/s41467-020-19187-w

Collalti, A., Marconi, S., Ibrom, A., Trotta, C., Anav, A., Andrea, E., Matteucci, G., Montagnani, L., Gielen, B., Mammarella, I., Grünwald, T., Knohl, A., Berninger, F., Zhao, Y., Valentini, R., & Santini, M. (2016). Validation of 3D-CMCC Forest ecosystem model (v.5.1) against eddy covariance data for 10 European forest sites. *Geoscientific Model Development*, *9*(2), 479–504. https://doi.org/10.5194/gmd-9-479-2016

Collalti, A., Perugini, L., Santini, M., Chiti, T., Nolè, A., Matteucci, G., & Valentini, R. (2014). A process-based model to simulate growth in forests with complex structure: Evaluation and use of 3D-CMCC Forest ecosystem model in a deciduous forest in Central Italy. *Ecological Modelling*, *272*, 362–378. https://doi.org/10.1016/j.ecolmodel.2013.09.016

Collalti, A., Thornton, P. E., Cescatti, A., Rita, A., Borghetti, M., Nolè, A., Trotta, C., Ciais, P., & Matteucci, G. (2019). The sensitivity of the forest carbon budget shifts across processes along with stand development and climate change. *Ecological Applications*, *29*(2), e01837. https://doi.org/10.1002/eap.1837

Collalti, A., Tjoelker, M. G., Hoch, G., Mäkelä, A., Guidolotti, G., Heskel, M., Petit, G., Ryan, M. G., Battipaglia, G., Matteucci, G., & Prentice, I. C. (2020). Plant respiration: Controlled by photosynthesis or biomass? *Global Change Biology*, *26*(3), 1739–1753. https://doi.org/10.1111/gcb.14857

Collalti, A., Trotta, C., Keenan, F. K., Ibrom, A., Bond-Lamberty, B., Grote, R., Vicca, S., Reyer, C. P. O., Migliavacca, M., Veroustraete, F., Anav, A., Campioli, M., Scoccimarro, E., Sigut, L., Gieco, E., Cescatti, A., & Matteucci, G. (2018). Thinning can reduce losses in carbon use efficiency and carbon stocks in managed forests under warmer climate. *Journal of Advances in Modeling Earth Systems*, *10*, 2427–2452. https://doi.org/10.1002/2018MS001275

Cysneiros, V. C., de Souza, F. C., Gaui, T. D., Pelissari, A. L., Orso, G. A., Machado, S. D., de Carvalho, D. C., & Silveira, T. B. (2021). Integrating climate, soil and stand structure into allometric models: An approach of site-effects on tree allometry in Atlantic Forest. *Ecological Indicators*, *127*, 107794. https://doi.org/10.1016/j.ecolind.2021.107794

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., … Norby, R. J. (2013). Forest water use and water use efficiency at elevated CO_2: A model-data intercomparison at two contrasting temperate forest FACE sites. *Global Change Biology*, *19*(6), 1759–1779. https://doi.org/10.1111/gcb.12164

De Kauwe, M. G., Medlyn, B. E., Zaehle, S., Walker, A. P., Dietze, M. C., Wang, Y. P., Luo, Y., Jain, A. K., El-Masri, B., Hickler, T., Wårlind, D., Weng, E., Parton, W. J., Thornton, P. E., Wang, S., Prentice, I. C., Asao, S., Smith, B., McCarthy, H. R., … Norby, R. J. (2014). Where does the carbon go? A model-data intercomparison of vegetation carbon allocation and turnover processes at two temperate forest free-air CO_2 enrichment sites. *New Phytologist*, *203*(3), 883–899. https://doi.org/10.1111/nph.12847

de Pury, D. G. G., & Farquhar, G. D. (1997). Simple scaling of photosynthesis from leaves to canopies without the errors of big-leaf models. *Plant, Cell & Environment*, *20*(5), 537–557. https://doi.org/10.1111/j.1365-3040.1997.00094.x

de Wergifosse, L., André, F., Goosse, H., Boczon, A., Cecchini, S., Ciceu, A., Collalti, A., Cools, N., D'Andrea, E., De Vos, B., Hamdi, R., Ingerslev, M., Knudsen, M. A., Kowalska, A., Leca, S., Matteucci, G., Nord-Larsen, T., Sanders, T. G. M., Schmitz, A., … Jonard, M. (2022). Simulating tree growth response to climate change in structurally diverse oak and beech forests. *Science of the Total Environment*, *806*, 150422. https://doi.org/10.1016/j.scitotenv.2021.150422

Didion, M., Kupferschmid, A. D., Zingg, A., Fahse, L., & Bugmann, H. (2009). Gaining local accuracy while not losing generality—Extending the range of gap model applications. *Canadian Journal of Forest Research*, *39*(6), 1092–1107. https://doi.org/10.1139/x09-041

Dietze, M. C., Vargas, R., Richardson, A. D., Stoy, P. C., Barr, A. G., Anderson, R. S., Arain, M. A., Baker, I. T., Black, T. A., Chen, J. M., Ciais, P., Flanagan, L. B., Gough, C. M., Grant, R. F., Hollinger, D., Izaurralde, R. C., Kucharik, C. J., Lafleur, P., Liu, S., … Weng, E. (2011). Characterizing the performance of ecosystem models across time scales: A spectral analysis of the north American carbon program site-level synthesis. *Journal of Geophysical Research: Biogeosciences*, *116*(G4), G04029. https://doi.org/10.1029/2011JG001661

Dirnböck, T., Kraus, D., Grote, R., Klatt, S., Kobler, J., Schindlbacher, A., Seidl, R., Thom, D., & Kiese, R. (2020). Substantial understory contribution to the C sink of a European temperate mountain forest landscape. *Landscape Ecology*, *35*(2), 483–499. https://doi.org/10.1007/s10980-019-00960-2

Fabrika, M., & Ďurský, J. (2005). Algorithms and software solution of thinning models for SIBYLA growth simulator. *Journal of Forest Science*, *51*(10), 431–445. https://doi.org/10.17221/4577-jfs

Farquhar, G. D., von Caemmerer, S., & Berry, J. A. (1980). A biochemical model of photosynthetic CO_2 assimilation in leaves of C3 species. *Planta*, *149*(1), 78–90. https://doi.org/10.1007/BF00386231

Fasiolo, M., Goude, Y., Nedellec, R., & Wood, S. N. (2017). Fast calibrated additive quantile regression. *arxiv*. https://arxiv.org/abs/1707.03307

Finzi, A. C., Giasson, M.-A., Barker Plotkin, A. A., Aber, J. D., Boose, E. R., Davidson, E. A., Dietze, M. C., Ellison, A. M., Frey, S. D., Goldman, E., Keenan, T. F., Melillo, J. M., Munger, J. W., Nadelhoffer, K. J., Ollinger, S. V., Orwig, D. A., Pederson, N., Richardson, A. D., Savage, K., … Foster, D. R. (2020). Carbon budget of the Harvard Forest long-term ecological research site: Pattern, process, and response to global change. *Ecological Monographs*, *90*(4), e01423. https://doi.org/10.1002/ecm.1423

Foken, T. (2008). *Micrometeorology*. Springer.

Fontes, L., Bontemps, J.-D., Bugmann, H., Van Oijen, M., Gracia, C., Kramer, K., Lindner, M., Rötzer, T., & Skovsgaard, J. P. (2010). Models for supporting forest management in a changing environment. *Forest Systems*, *19*, 8–29.

Forsius, M., Kujala, H., Minunno, F., Holmberg, M., Leikola, N., Mikkonen, N., Autio, I., Paunu, V.-V., Tanhuanpää, T., Hurskainen, P., Mäyrä, J., Kivinen, S., Keski-Saari, S., Kosenius, A.-K., Kuusela, S., Virkkala, R., Viinikka, A., Vihervaara, P., Akujärvi, A., … Heikkinen, R. K. (2021). Developing a spatially explicit modelling and evaluation framework for integrated carbon sequestration and biodiversity conservation: Application in southern Finland. *Science of the Total Environment*, *775*, 145847. https://doi.org/10.1016/j.scitotenv.2021.145847

Franklin, J., Davis, F. W., Ikegami, M., Syphard, A. D., Flint, L. E., Flint, A. L., & Hannah, L. (2013). Modeling plant species distributions under future climates: How fine scale do climate projections need to be? *Global Change Biology*, *19*(2), 473–483. https://doi.org/10.1111/gcb.12051

Friedlingstein, P., Joel, G., Field, C. B., & Fung, I. Y. (1999). Toward an allocation scheme for global terrestrial carbon models. *Global Change Biology*, *5*(7), 755–770. https://doi.org/10.1046/j.1365-2486.1999.00269.x

Frieler, K., Lange, S., Piontek, F., Reyer, C. P. O., Schewe, J., Warszawski, L., Zhao, F., Chini, L., Denvil, S., Emanuel, K., Geiger, T., Halladay, K., Hurtt, G., Mengel, M., Murakami, D., Ostberg, S., Popp, A., Riva, R., Stevanovic, M., … Yamagata, Y. (2017). Assessing the impacts of 1.5°C global warming – Simulation protocol of the inter-sectoral impact model intercomparison project (ISIMIP2b). *Geoscientific Model Development*, *10*(12), 4321–4345. https://doi.org/10.5194/gmd-10-4321-2017

Gauch, H. G., Hwang, J. T. G., & Fick, G. W. (2003). Model evaluation by comparison of model-based predictions and measured values. *Agronomy Journal*, *95*, 1442–1446.

Grote, R. (1998). Integrating dynamic morphological properties into forest growth modelling: II allocation and mortality. *Forest Ecology and Management*, *111*(2), 193–210. https://doi.org/10.1016/S0378-1127(98)00328-4

Grote, R., Korhonen, J., & Mammarella, I. (2011). Challenges for evaluating process-based models of gas exchange at forest sites with fetches of various species. *Forest Systems*, *20*, 389–406. https://doi.org/10.5424/fs/20112003-11084

Grote, R., Kraus, D., Weis, W., Ettl, R., & Göttlein, A. (2020). Dynamic coupling of allometric ratios to a process-based forest growth model for estimating the impacts of stand density changes. *Forestry: An International Journal of Forest Research*, *93*(5), 601–615. https://doi.org/10.1093/forestry/cpaa002

Gupta, R., & Sharma, L. K. (2019). The process-based forest growth model 3-PG for use in forest management: A review. *Ecological Modelling*, *397*, 55–73. https://doi.org/10.1016/j.ecolmodel.2019.01.007

Gutsch, M., Lasch-Born, P., Kollas, C., Suckow, F., & Reyer, C. P. O. (2018). Balancing trade-offs between ecosystem services in Germany's forests under climate change. *Environmental Research Letters*, *13*(4), 045012. https://doi.org/10.1088/1748-9326/aab4e5

Haxeltine, A., & Prentice, I. C. (1996a). A general model for the light-use efficiency of primary production. *Functional Ecology*, *10*(5), 551–561.

Haxeltine, A., & Prentice, I. C. (1996b). BIOME3: An equilibrium terrestrial biosphere model based on ecophysiological constraints, resource availability, and competition among plant functional types. *Global Biogeochemical Cycles*, *10*(4), 693–709. https://doi.org/10.1029/96GB02344

Hlásny, T., Barcza, Z., Barka, I., Merganičová, K., Sedmák, R., Kern, A., Pajtík, J., Balázs, B., Fabrika, M., & Churkina, G. (2014). Future carbon cycle in mountain spruce forests of Central Europe: Modelling framework and ecological inferences. *Forest Ecology and Management*, *328*, 55–68. https://doi.org/10.1016/j.foreco.2014.04.038

Holmberg, M., Aalto, T., Akujärvi, A., Arslan, A. N., Bergström, I., Böttcher, K., Lahtinen, I., Mäkelä, A., Markkanen, T., Minunno, F., Peltoniemi, M., Rankinen, K., Vihervaara, P., & Forsius, M. (2019). Ecosystem services related to carbon cycling - modeling present and future impacts in boreal forests. *Frontiers in Plant Science*, *10*, 343–351.

Horemans, J. A., Henrot, A., Delire, C., Kollas, C., Lasch-Born, P., Reyer, C., Suckow, F., François, L., & Ceulemans, R. (2017). Combining multiple statistical methods to evaluate the performance of process-based vegetation models across three forest stands. *Central European Forestry Journal*, *63*(4), 153–172. https://doi.org/10.1515/forj-2017-0025

Huber, N., Bugmann, H., Cailleret, M., Bircher, N., & Lafond, V. (2021). Stand-scale climate change impacts on forests over large areas: Transient responses and projection uncertainties. *Ecological Applications*, *31*(4), e02313.

Huber, N., Bugmann, H., & Lafond, V. (2020). Capturing ecological processes in dynamic forest models: Why there is no silver bullet to cope with complexity. *Ecosphere*, *11*(5), e03109.

Huntzinger, D. N., Schwalm, C., Michalak, A. M., Schaefer, K., King, A. W., Wei, Y., Jacobson, A., Liu, S., Cook, R. B., Post, W. M., Berthier, G., Hayes, D., Huang, M., Ito, A., Lei, H., Lu, C., Mao, J., Peng, C. H., Peng, S., … Zhu, Q. (2013). The north American carbon program multi-scale synthesis and terrestrial model Intercomparison project – Part 1: Overview and experimental design. *Geoscientific Model Development*, *6*(6), 2121–2133. https://doi.org/10.5194/gmd-6-2121-2013

Irauschek, F., Barka, I., Bugmann, H., Courbaud, B., Elkin, C., Hlásny, T., Klopcic, M., Mina, M., Rammer, W., & Lexer, M. J. (2021). Evaluating five forest models using multi-decadal inventory data from mountain forests. *Ecological Modelling*, *445*, 109493. https://doi.org/10.1016/j.ecolmodel.2021.109493

Kalliokoski, T., Heinonen, T., Holder, J., Lehtonen, A., Mäkelä, A., Minunno, F., Ollikainen, M., Packalen, T., Peltoniemi, M., Pukkala, T., Salminen, O., Schelhaas, M. J., Seppälä, J., Vauhkonen, J., & Kanninen, M. (2019). *Scenario analysis of similarities and differences between forest growth model projections*. Finnish Climate Change Panel.

Kalliokoski, T., Mäkelä, A., Fronzek, S., Minunno, F., & Peltoniemi, M. (2018). Decomposing sources of uncertainty in climate change projections of boreal forest primary production. *Agricultural and Forest Meteorology*, *262*, 192–205. https://doi.org/10.1016/j.agrformet.2018.06.030

Keenan, T., Maria Serra, J., Lloret, F., Ninyerola, M., & Sabate, S. (2011). Predicting the future of forests in the Mediterranean under climate change, with niche- and process-based models: CO_2 matters! *Global Change Biology*, *17*(1), 565–579. https://doi.org/10.1111/j.1365-2486.2010.02254.x

Klesse, S., Babst, F., Lienert, S., Spahni, R., Joos, F., Bouriaud, O., Carrer, M., Di Filippo, A., Poulter, B., Trotsiuk, V., Wilson, R., & Frank, D. C. (2018). A combined tree ring and vegetation model assessment of European Forest growth sensitivity to interannual climate variability. *Global Biogeochemical Cycles*, *32*(8), 1226–1240. https://doi.org/10.1029/2017GB005856

Kramer, K., Leinonen, I., Bartelink, H. H., Berbigier, P., Borghetti, M., Bernhofer, C., Cienciala, E., Dolman, A. J., Froer, O., Gracia, C. A., Granier, A., Grünwald, T., Hari, P., Jans, W., Kellomäki, S., Loustau, D., Magnani, F., Markkanen, T., Matteucci, G., … Vesala, T. (2002). Evaluation of six process-based forest growth models using eddy-covariance measurements of CO_2 and H_2O fluxes at six forest sites in Europe. *Global Change Biology*, *8*(3), 213–230. https://doi.org/10.1046/j.1365-2486.2002.00471.x

Landsberg, J. J., & Waring, R. H. (1997). A generalised model of forest productivity using simplified concepts of radiation-use efficiency, carbon balance and partitioning. *Forest Ecology and Management*, *95*(3), 209–228. https://doi.org/10.1016/S0378-1127(97)00026-1

Lasch-Born, P., Suckow, F., Reyer, C. P. O., Gutsch, M., Kollas, C., Badeck, F. W., Bugmann, H. K. M., Grote, R., Fürstenau, C., Lindner, M., & Schaber, J. (2020). Description and evaluation of the process-based forest model 4C v2.2 at four European forest sites. *Geoscientific Model Development*, *13*(11), 5311–5343. https://doi.org/10.5194/gmd-13-5311-2020

Lasslop, G., Reichstein, M., Papale, D., Richardson, A. D., Arneth, A., Barr, A., Stoy, P., & Wohlfahrt, G. (2010). Separation of net ecosystem exchange into assimilation and respiration using a light response curve approach: Critical issues and global evaluation. *Global Change Biology*, *16*(1), 187–208. https://doi.org/10.1111/j.1365-2486.2009.02041.x

Levins, R. (1966). The strategy of model building in population biology. *American Scientist*, *54*(4), 421–431.

Lindauer, M., Schmid, H. P., Grote, R., Mauder, M., Steinbrecher, R., & Wolpert, B. (2014). Net ecosystem exchange over a non-cleared wind-throw-disturbed upland spruce forest—Measurements and simulations. *Agricultural and Forest Meteorology*, *197*, 219–234. https://doi.org/10.1016/j.agrformet.2014.07.005

Lindner, M., Fitzgerald, J. B., Zimmermann, N. E., Reyer, C., Delzon, S., van der Maaten, E., Schelhaas, M. J., Lasch, P., Eggers, J., van der Maaten-Theunissen, M., Suckow, F., Psomas, A., Poulter, B., & Hanewinkel, M. (2014). Climate change and European forests: What do we know, what are the uncertainties, and what are the implications for forest management? *Journal of Environmental Management*, *146*, 69–83. https://doi.org/10.1016/j.jenvman.2014.07.030

Mahnken, M., Collalti, A., Dalmonech, D., Trotta, C., Trotsiuk, V., Augustynczik, A. L. D., Yousefpour, R, Gutsch, M., Cameron, D., Bugmann, H., Huber, N., Thrippleton, T., Bohn, F., Nadal-Sala, D., Sabaté, S., Grote, R., Mäkelä, A., Minunno, F., Peltoniemi, M., … Reyer, C. P. O. (2022) ISIMIP2a simulation data from the regional forests sector (v1.0). *ISIMIP Repository*. https://doi.org/10.48364/ISIMIP.169780

Mäkelä, A. (1997). A carbon balance model of growth and self-pruning in trees based on structural relationships. *Forest Science*, *43*(1), 7–24. https://doi.org/10.1093/forestscience/43.1.7

Mäkelä, A., Pulkkinen, M., Kolari, P., Lagergren, F., Bergibier, P., Lindroth, A., Loustau, D., Nikinmaa, E., Vesala, T., & Hari, P. (2008). Developing an empirical model of stand GPP with the LUE approach: Analysis of eddy covariance data at five contrasting conifer sites in Europe. *Global Change Biology*, *14*(1), 92–108. https://doi.org/10.1111/j.1365-2486.2007.01463.x

Marconi, S., Chiti, T., Nolè, A., Valentini, R., & Collalti, A. (2017). The role of respiration in estimation of net carbon cycle: Coupling soil carbon dynamics and canopy turnover in a novel version of 3D-CMCC Forest ecosystem model. *Forests*, *8*(6), 220. https://www.mdpi.com/1999-4907/8/6/220

Marechaux, I., Langerwisch, F., Huth, A., Bugmann, H., Morin, X., Reyer, C. P. O., Seidl, R., Collalti, A., Dantas de Paula, M., Fischer, R., Gutsch, M., Lexer, M. J., Lischke, H., Rammig, A., Rodig, E., Sakschewski, B., Taubert, F., Thonicke, K., Vacchiano, G., & Bohn, F. J. (2021). Tackling unresolved questions in forest ecology: The past and future role of simulation models. *Ecology and Evolution*, *11*(9), 3746–3770. https://doi.org/10.1002/ece3.7391

McCree, K., & Setlick, I. (1970). Prediction and measurement of photosynthetic productivity. *Proceedings of the IBP/PP technical meeting*, Trebon, 14–21 September 1969.

Medlyn, B. E., Duursma, R. A., & Zeppel, M. J. B. (2011). Forest productivity under climate change: A checklist for evaluating model studies. *Wiley Interdisciplinary Reviews: Climate Change*, *2*(3), 332–355. https://doi.org/10.1002/wcc.108

Medlyn, B. E., Zaehle, S., De Kauwe, M. G., Walker, A. P., Dietze, M. C., Hanson, P. J., Hickler, T., Jain, A. K., Luo, Y., Parton, W., Prentice, I. C., Thornton, P. E., Wang, S., Wang, Y.-P., Weng, E., Iversen, C. M., McCarthy, H. R., Warren, J. M., Oren, R., & Norby, R. J. (2015). Using ecosystem experiments to improve vegetation models. *Nature Climate Change*, *5*(6), 528–534. https://doi.org/10.1038/nclimate2621

Merganic, J., Merganicova, K., Vybostok, J., Valent, P., Bahyl, J., & Yousefpour, R. (2020). Searching for pareto fronts for forest stand wind stability by incorporating timber and biodiversity values. *Forests*, *11*(5), 583. https://www.mdpi.com/1999-4907/11/5/583

Mina, M., Bugmann, H., Klopcic, M., & Cailleret, M. (2015). Accurate modeling of harvesting is key for projecting future forest dynamics: A case study in the Slovenian mountains. *Regional Environmental Change*, *17*(1), 49–64. https://doi.org/10.1007/s10113-015-0902-2

Minunno, F., Peltoniemi, M., Härkönen, S., Kalliokoski, T., Makinen, H., & Mäkelä, A. (2019). Bayesian calibration of a carbon balance model PREBAS using data from permanent growth experiments and national forest inventory. *Forest Ecology and Management*, *440*, 208–257. https://doi.org/10.1016/j.foreco.2019.02.041

Moffat, A. M., Beckstein, C., Churkina, G., Mund, M., & Heimann, M. (2010). Characterization of ecosystem responses to climatic controls using artificial neural networks. *Global Change Biology*, *16*(10), 2737–2749. https://doi.org/10.1111/j.1365-2486.2010.02171.x

Monteith, J. L., Moss, C. J., Cooke, G. W., Pirie, N. W., & Bell, G. D. H. (1977). Climate and the efficiency of crop production in Britain. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *281*(980), 277–294. https://doi.org/10.1098/rstb.1977.0140

Moore, A. D. (1989). On the maximum growth equation used in forest gap simulation models. *Ecological Modelling*, *45*, 63–67.

Morales, P., Sykes, M. T., Prentice, I. C., Smith, P., Smith, B., Bugmann, H., Zierl, B., Friedlingstein, P., Viovy, N., Sabate, S., Sanchez, A., Pla, E., Gracia, C. A., Sitch, S., Arneth, A., & Ogee, J. (2005). Comparing and evaluating process-based ecosystem model predictions of carbon and water fluxes in major European forest biomes. *Global Change Biology*, *11*(12), 2211–2233. https://doi.org/10.1111/j.1365-2486.2005.01036.x

Nadal-Sala, D., Grote, R., Birami, B., Lintunen, A., Mammarella, I., Preisler, Y., Rotenberg, E., Salmon, Y., Tatarinov, F., Yakir, D., & Ruehr, N. K. (2021). Assessing model performance via the most limiting environmental driver in two differently stressed pine stands. *Ecological Applications*, *31*(4), e02312. https://doi.org/10.1002/eap.2312

Nadal-Sala, D., Hartig, F., Gracia, C. A., & Sabaté, S. (2019). Global warming likely to enhance black locust (*Robinia pseudoacacia* L.) growth in a Mediterranean riparian forest. *Forest Ecology and Management*, *449*, 117448. https://doi.org/10.1016/j.foreco.2019.117448

Novick, K. A., Ficklin, D. L., Stoy, P. C., Williams, C. A., Bohrer, G., Oishi, A. C., Papuga, S. A., Blanken, P. D., Noormets, A., Sulman, B. N., Scott, R. L., Wang, L., & Phillips, R. P. (2016). The increasing importance of atmospheric demand for ecosystem water and carbon fluxes. *Nature Climate Change*, *6*(11), 1023–1027. https://doi.org/10.1038/nclimate3114

Oberpriller, J., Cameron, D. R., Dietze, M. C., & Hartig, F. (2021). Towards robust statistical inference for complex computer models. *Ecology Letters*, *24*(6), 1251–1261. https://doi.org/10.1111/ele.13728

Oikawa, P. Y., Sturtevant, C., Knox, S. H., Verfaillie, J., Huang, Y. W., & Baldocchi, D. D. (2017). Revisiting the partitioning of net ecosystem exchange of CO_2 into photosynthesis and respiration with simultaneous flux measurements of $^{13}CO_2$ and CO_2, soil respiration and a biophysical model, CANVEG. *Agricultural and Forest Meteorology*, *234–235*, 149–163. https://doi.org/10.1016/j.agrformet.2016.12.016

Pardos, M., del Río, M., Pretzsch, H., Jactel, H., Bielak, K., Bravo, F., Brazaitis, G., Defossez, E., Engel, M., Godvod, K., Jacobs, K., Jansone, L., Jansons, A., Morin, X., Nothdurft, A., Oreti, L., Ponette, Q., Pach, M., Riofrío, J., … Calama, R. (2021). The greater resilience of mixed forests to drought mainly depends on their composition: Analysis along a climate gradient across Europe. *Forest Ecology and Management*, *481*, 118687. https://doi.org/10.1016/j.foreco.2020.118687

Paschalis, A., Fatichi, S., Zscheischler, J., Ciais, P., Bahn, M., Boysen, L., Chang, J., De Kauwe, M., Estiarte, M., Goll, D., Hanson, P. J., Harper, A. B., Hou, E., Kigel, J., Knapp, A. K., Larsen, K. S., Li, W., Lienert, S., Luo, Y., … Zhu, Q. (2020). Rainfall manipulation experiments as simulated by terrestrial biosphere models: Where do we stand? *Global Change Biology*, *26*(6), 3336–3355. https://doi.org/10.1111/gcb.15024

Pastorello, G., Trotta, C., Canfora, E., Chu, H., Christianson, D., Cheah, Y.-W., Poindexter, C., Chen, J., Elbashandy, A., Humphrey, M., Isaac, P., Polidori, D., Reichstein, M., Ribeca, A., van Ingen, C., Vuichard, N., Zhang, L., Amiro, B., Ammann, C., … Papale, D. (2020). The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data. *Scientific Data*, *7*(1), 225. https://doi.org/10.1038/s41597-020-0534-3

Peltoniemi, M., Pulkkinen, M., Aurela, M., Pumpanen, J., Kolari, P., & Mäkelä, A. (2015). A semi-empirical model of boreal forest gross primary production, evapotranspiration, and soil water – Calibration and sensitivity analysis. *Boreal Environment Research*, *20*, 151–171. http://hdl.handle.net/10138/228031

Pretzsch, H., Forrester, D. I., & Rötzer, T. (2015). Representation of species mixing in forest growth models. A review and perspective. *Ecological Modelling*, *313*, 276–292. https://doi.org/10.1016/j.ecolmodel.2015.06.044

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rehfeldt, G. E., Crookston, N. L., Warwell, M. V., & Evans, J. S. (2006). Empirical analyses of plant-climate relationships for the western United States. *International Journal of Plant Sciences*, *167*(6), 1123–1150.

Reichstein, M., Falge, E., Baldocchi, D., Papale, D., Aubinet, M., Berbigier, P., Bernhofer, C., Buchmann, N., Gilmanov, T., Granier, A., Grünwald, T., Havránková, K., Ilvesniemi, H., Janous, D., Knohl, A., Laurila, T., Lohila, A., Loustau, D., Matteucci, G., … Valentini, R. (2005). On the separation of net ecosystem exchange into assimilation and ecosystem respiration: Review and improved algorithm. *Global Change Biology*, *11*(9), 1424–1439. https://doi.org/10.1111/j.1365-2486.2005.001002.x

Reyer, C., Silveyra Gonzalez, R., Dolos, K., Hartig, F., Hauf, Y., Noack, M., Lasch-Born, P., Rötzer, T., Pretzsch, H., Meesenburg, H., Fleck, S., Wagner, M., Bolte, A., Sanders, T., Kolari, P., Mäkelä, A., Vesala, T., Mammarella, I., Pumpanen, J., … Frieler, K. (2020a). The PROFOUND database for evaluating vegetation models and simulating climate impacts on European forests version V. 0.3. *GFZ Data Services*. https://doi.org/10.5880/PIK.2020.006

Reyer, C. P. O., Lasch-Born, P., Suckow, F., Gutsch, M., Murawski, A., & Pilz, T. (2013). Projections of regional changes in forest net primary productivity for different tree species in Europe driven by climate change and carbon dioxide. *Annals of Forest Science*, *71*(2), 211–225. https://doi.org/10.1007/s13595-013-0306-8

Reyer, C. P. O., Silveyra Gonzalez, R., Dolos, K., Hartig, F., Hauf, Y., Noack, M., Lasch-Born, P., Rötzer, T., Pretzsch, H., Meesenburg, H., Fleck, S., Wagner, M., Bolte, A., Sanders, T. G. M., Kolari, P., Mäkelä, A., Vesala, T., Mammarella, I., Pumpanen, J., … Frieler, K. (2020b). The PROFOUND database for evaluating vegetation models and simulating climate impacts on European forests. *Earth System Science Data*, *12*(2), 1295–1320. https://doi.org/10.5194/essd-12-1295-2020

Richardson, A. D., Anderson, R. S., Arain, M. A., Barr, A. G., Bohrer, G., Chen, G., Chen, J. M., Ciais, P., Davis, K. J., Desai, A. R., Dietze, M. C., Dragoni, D., Garrity, S. R., Gough, C. M., Grant, R., Hollinger, D. Y., Margolis, H. A., McCaughey, H., Migliavacca, M., … Xue, Y. (2012). Terrestrial biosphere models need better representation of vegetation phenology: Results from the north American carbon program site synthesis. *Global Change Biology*, *18*(2), 566–584. https://doi.org/10.1111/j.1365-2486.2011.02562.x

Rödig, E., Huth, A., Bohn, F., Rebmann, C., & Cuntz, M. (2017). Estimating the carbon fluxes of forests with an individual-based forest model. *Forest Ecosystems*, *4*(1), 4. https://doi.org/10.1186/s40663-017-0091-1

Rollinson, C. R., Dawson, A., Raiho, A. M., Williams, J. W., Dietze, M. C., Hickler, T., Jackson, S. T., McLachlan, J., Moore, J. P., Poulter, B., Quaife, T., Steinkamp, J., & Trachsel, M. (2021). Forest responses to last-millennium hydroclimate variability are governed by spatial variations in ecosystem sensitivity. *Ecology Letters*, *24*(3), 498–508. https://doi.org/10.1111/ele.13667

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., & Dietze, M. C. (2017). 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*, *23*(7), 2755–2767. https://doi.org/10.1111/gcb.13626

Sabaté, S., Gracia, C. A., & Sánchez, A. (2002). Likely effects of climate change on growth of *Quercus ilex*, *Pinus halepensis*, *Pinus pinaster*, *Pinus sylvestris* and *Fagus sylvatica* forests in the Mediterranean region. *Forest Ecology and Management*, *162*, 23–37.

Schaefer, K., Schwalm, C. R., Williams, C., Arain, M. A., Barr, A., Chen, J. M., Davis, K. J., Dimitrov, D., Hilton, T. W., Hollinger, D. Y., Humphreys, E., Poulter, B., Raczka, B. M., Richardson, A. D., Sahoo, A., Thornton, P., Vargas, R., Verbeeck, H., Anderson, R., … Zhou, X. (2012). A model-data comparison of gross primary productivity: Results from the north American carbon program site synthesis. *Journal of Geophysical Research: Biogeosciences*, *117*(G3), G03010. https://doi.org/10.1029/2012JG001960

Schweier, J., Molina-Herrera, S., Ghirardo, A., Grote, R., Díaz-Pinés, E., Kreuzwieser, J., Haas, E., Butterbach-Bahl, K., Rennenberg, H., Schnitzler, J.-P., & Becker, G. (2017). Environmental impacts of bioenergy wood production from poplar short-rotation coppice grown at a marginal agricultural site in Germany. *GCB Bioenergy*, *9*(7), 1207–1221. https://doi.org/10.1111/gcbb.12423

Stoy, P. C., Dietze, M. C., Richardson, A. D., Vargas, R., Barr, A. G., Anderson, R. S., Arain, M. A., Baker, I. T., Black, T. A., Chen, J. M., Cook, R. B., Gough, C. M., Grant, R. F., Hollinger, D. Y., Izaurralde, R. C., Kucharik, C. J., Lafleur, P., Law, B. E., Liu, S., … Weng, E. (2013). Evaluating the agreement between measurements and models of net ecosystem exchange at different times and timescales using wavelet coherence: An example using data from the north American carbon program site-level interim synthesis. *Biogeosciences*, *10*(11), 6893–6909. https://doi.org/10.5194/bg-10-6893-2013

Thornley, J. H. M. (1970). Respiration, growth and maintenance in plants. *Nature*, *227*(5255), 304–305. https://doi.org/10.1038/227304b0

Toïgo, M., Perot, T., Courbaud, B., Castagneyrol, B., Gégout, J.-C., Longuetaud, F., Jactel, H., & Vallet, P. (2018). Difference in shade tolerance drives the mixture effect on oak productivity. *Journal of Ecology*, *106*(3), 1073–1082. https://doi.org/10.1111/1365-2745.12811

Toïgo, M., Vallet, P., Perot, T., Bontemps, J.-D., Piedallu, C., Courbaud, B., & Canham, C. (2015). Overyielding in mixed forests decreases with site productivity. *Journal of Ecology*, *103*(2), 502–512. https://doi.org/10.1111/1365-2745.12353

Trotsiuk, V., Hartig, F., Cailleret, M., Babst, F., Forrester, D. I., Baltensweiler, A., Buchmann, N., Bugmann, H., Gessler, A., Gharun, M., Minunno, F., Rigling, A., Rohner, B., Stillhard, J., Thurig, E., Waldner, P., Ferretti, M., Eugster, W., & Schaub, M. (2020). Assessing the response of forest productivity to climate extremes in Switzerland using model-data fusion. *Global Change Biology*, *26*, 2463–2476. https://doi.org/10.1111/gcb.15011

Trugman, A. T., Anderegg, L. D. L., Anderegg, W. R. L., Das, A. J., & Stephenson, N. L. (2021). Why is tree drought mortality so hard to predict? *Trends in Ecology & Evolution*, *36*(6), 520–532. https://doi.org/10.1016/j.tree.2021.02.001

Vallet, P., & Pérot, T. (2018). Coupling transversal and longitudinal models to better predict *Quercus petraea* and *Pinus sylvestris* stand growth under climate change. *Agricultural and Forest Meteorology*, *263*, 258–266.

van Oijen, M., Balkovi, J., Beer, C., Cameron, D. R., Ciais, P., Cramer, W., Kato, T., Kuhnert, M., Martin, R., Myneni, R., Rammig, A., Rolinski, S., Soussana, J. F., Thonicke, K., Van der Velde, M., & Xu, L. (2014). Impact of droughts on the carbon cycle in European vegetation: A probabilistic risk analysis using six vegetation models. *Biogeosciences*, *11*(22), 6357–6375. https://doi.org/10.5194/bg-11-6357-2014

Wagener, T., Reinecke, R., & Pianosi, F. (2022). On the evaluation of climate change impact models. *WIREs Climate Change*, *13*(3), e772. https://doi.org/10.1002/wcc.772

Walker, A. P., De Kauwe, M. G., Bastos, A., Belmecheri, S., Georgiou, K., Keeling, R. F., McMahon, S. M., Medlyn, B. E., Moore, D. J. P., Norby, R. J., Zaehle, S., Anderson-Teixeira, K. J., Battipaglia, G., Brienen, R. J. W., Cabugao, K. G., Cailleret, M., Campbell, E., Canadell, J. G., Ciais, P., … Zuidema, P. A. (2021). Integrating the evidence for a

terrestrial carbon sink caused by increasing atmospheric CO_2. *New Phytologist*, *229*(5), 2413–2445. https://doi.org/10.1111/nph.16866

Walker, A. P., Zaehle, S., Medlyn, B. E., De Kauwe, M. G., Asao, S., Hickler, T., Parton, W., Ricciuto, D. M., Wang, Y.-P., Wårlind, D., & Norby, R. J. (2015). Predicting long-term carbon sequestration in response to CO_2 enrichment: How and why do current ecosystem models differ? *Global Biogeochemical Cycles*, *29*(4), 476–495. https://doi.org/10.1002/2014GB004995

Wei, Y., Liu, S., Huntzinger, D. N., Michalak, A. M., Viovy, N., Post, W. M., Schwalm, C. R., Schaefer, K., Jacobson, A. R., Lu, C., Tian, H., Ricciuto, D. M., Cook, R. B., Mao, J., & Shi, X. (2014). The north American carbon program multi-scale synthesis and terrestrial model Intercomparison project – Part 2: Environmental driver data. *Geoscientific Model Development*, *7*(6), 2875–2893. https://doi.org/10.5194/gmd-7-2875-2014

Weisberg, M. (2007). Forty years of 'the strategy': Levins on model building and idealization. *Biology and Philosophy*, *21*(5), 623–645. https://doi.org/10.1007/s10539-006-9051-9

Wood, S. N. (2017). *Generalized additive models: An introduction with R* (2nd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781315370279

Xenakis, G., Ray, D., & Mencuccini, M. (2008). Sensitivity and uncertainty analysis from a coupled 3-PG and soil organic matter decomposition model. *Ecological Modelling*, *219*(1), 1–16. https://doi.org/10.1016/j.ecolmodel.2008.07.020

Zaehle, S., Medlyn, B. E., De Kauwe, M. G., Walker, A. P., Dietze, M. C., Hickler, T., Luo, Y., Wang, Y.-P., El-Masri, B., Thornton, P., Jain, A., Wang, S., Warlind, D., Weng, E., Parton, W., Iversen, C. M., Gallet-Budynek, A., McCarthy, H., Finzi, A., … Norby, R. J. (2014). Evaluation of 11 terrestrial carbon–nitrogen cycle models against observations from two temperate free-air CO_2 enrichment studies. *New Phytologist*, *202*(3), 803–822. https://doi.org/10.1111/nph.12697

Zhang, Q., Ficklin, D. L., Manzoni, S., Wang, L., Way, D., Phillips, R. P., & Novick, K. A. (2019). Response of ecosystem intrinsic water use efficiency and gross primary productivity to rising vapor pressure deficit. *Environmental Research Letters*, *14*(7), 074023. https://doi.org/10.1088/1748-9326/ab2603

Zhang, Z., Zhang, R., Cescatti, A., Wohlfahrt, G., Buchmann, N., Zhu, J., Chen, G., Moyano, F., Pumpanen, J., Hirano, T., Takagi, K., & Merbold, L. (2017). Effect of climate warming on the annual terrestrial net ecosystem CO_2 exchange globally in the boreal and temperate regions. *Scientific Reports*, *7*(1), 3108. https://doi.org/10.1038/s41598-017-03386-5

Zhou, H., Yue, X., Lei, Y., Zhang, T., Tian, C., Ma, Y., & Cao, Y. (2021). Responses of gross primary productivity to diffuse radiation at global FLUXNET sites. *Atmospheric Environment*, *244*, 117905. https://doi.org/10.1016/j.atmosenv.2020.117905

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Mahnken, M., Cailleret, M., Collalti, A., Trotta, C., Biondo, C., D'Andrea, E., Dalmonech, D., Marano, G., Mäkelä, A., Minunno, F., Peltoniemi, M., Trotsiuk, V., Nadal-Sala, D., Sabaté, S., Vallet, P., Aussenac, R., Cameron, D. R., Bohn, F. J., Grote, R. … Reyer, C. P. O. (2022). Accuracy, realism and general applicability of European forest models. *Global Change Biology*, *28*, 6921–6943. https://doi.org/10.1111/gcb.16384