RESEARCH ARTICLE

Accuracy, realism and general applicability of European forest models

Mats Mahnken1,2 | Maxime Cailleret3,4 | Alessio Collalti5,6,7 | Carlo Trotta6,7 | Corrado Biondo6,7 | Ettore D’Andrea5 | Daniela Dal'monech5 | Gia Marano5,8 | Annikki Mäkelä9 | Francesco Minunno9 | Mikko Peltoniemi10 | Volodymyr Trotsiuk4 | Daniel Nadal-Sala11,12 | Santiago Sabaté12,13 | Patrick Vallet14 | Raphaël Aussac14 | David R. Cameron15 | Friedrich J. Bohn16 | Rüdiger Grote11 | Andrey L. D. Augustynczyk17 | Rasoul Yousefpour18,19 | Nica Huber8,20 | Harald Bugmann8 | Katarína Merganičová21,22 | Jan Merganic23 | Peter Valent23 | Petra Lasch-Born1 | Florian Hartig24 | Iliusi D. Vega del Valle1 | Jan Volkholz1 | Martin Gutsch1 | Giorgio Matteucci5 | Jan Krejza25,26 | Andreas Ibrom27 | Henning Meesenburg28 | Thomas Rötzer29 | Marieke van der Maaten-Theunissen2 | Ernst van der Maaten2 | Christopher P. O. Reyer1

1Potsdam Institute for Climate Impact Research (PIK), Leibniz Association, Potsdam, Germany
2Forest Growth and Woody Biomass Production, TU Dresden, Tharandt, Germany
3UMR RECOVER, INRAE, Aix-Marseille University, Aix-en-Provence, France
4Forest Dynamics Unit, Swiss Federal Research Institute WSL, Birmensdorf, Switzerland
5Forest Modelling Lab, National Research Council of Italy, Institute for Agriculture and Forestry Systems in the Mediterranean (CNR-ISAFOAM), Perugia, Italy
6Department of Innovation in Biological, Agro-Food and Forest Systems (DIBAF), University of Tuscia, Viterbo, Italy
7Division Impacts on Agriculture, Forests and Ecosystem Services (IAFES), Fondazione Centro Euro-Mediterraneo sui Cambiamenti Climatici, Viterbo, Italy
8Department of Environmental Systems Science, Forest Ecology, Institute of Terrestrial Ecosystems, ETH Zurich, Zurich, Switzerland
9Institute of Forest Sciences, Institute for Atmospheric and Earth System Research (INAR) and Faculty of Agriculture and Forestry, University of Helsinki, Helsinki, Finland
10Natural Resources Institute Finland (Luke), Helsinki, Finland
11Institute of Meteorology and Climate Research – Atmospheric Environmental Research (IMK-IFU), Karlsruhe Institute of Technology (KIT), Garmisch-Partenkirchen, Germany
12Ecology Section, Department of Evolutionary Biology, Ecology and Environmental Sciences, University of Barcelona (UB), Barcelona, Spain
13CREAF (Center for Ecological Research and Forestry Applications), Cerdanyola del Vallès, Spain
14LESSEM, INRAE, Univ. Grenoble Alpes, St-Martin-d’Hères, France
15UK Centre for Ecology and Hydrology, Penicuik, Midlothian, UK
16Helmholtz Centre for Environmental Research – UFZ, Leipzig, Germany
17International Institute for Applied Systems Analysis (IIASA), Laxenburg, Austria
18Forestry Economics and Forest Planning, University of Freiburg, Freiburg, Germany
19Institute of Forestry and Conservation, John Daniels Faculty of Architecture, Landscape and Design, University of Toronto, Toronto, Ontario, Canada
20Remote Sensing, Swiss Federal Research Institute WSL, Birmensdorf, Switzerland
21Faculty of Forestry and Wood Sciences, Czech University of Life Sciences Prague, Praha, Czech Republic
22Department of Biodiversity of Ecosystems and Landscape, Institute of Landscape Ecology, Slovak Academy of Sciences, Nitra, Slovakia
23Faculty of Forestry, Technical University in Zvolen, Zvolen, Slovak Republic

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2022 The Authors. Global Change Biology published by John Wiley & Sons Ltd.
**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.
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; (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).
<table>
<thead>
<tr>
<th>Model</th>
<th>Photosynthesis</th>
<th>Autotrophic respiration</th>
<th>Carbon allocation</th>
<th>Structure development</th>
<th>Mortality</th>
<th>Example applications</th>
<th>Model class</th>
</tr>
</thead>
<tbody>
<tr>
<td>SALEM</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>Allometric equation</td>
<td>Diameter, density, and site index specific stand level-dependent increment model (1, 2)</td>
<td>Diameter-dependent specific self thinning (1)</td>
<td>1, 3, 4</td>
</tr>
<tr>
<td>SIBYLA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>Empirical: based on tree age, site specification, tree vitality and competition</td>
<td>Empirical: based on site specification, tree vitality and competition</td>
<td>Empirical: based on tree dimensions, growth and stand density</td>
<td>5, 6, 7</td>
</tr>
<tr>
<td>ForClim v.3.3</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>Derived from diameter increment under consideration of light availability and climate specific maximum tree height</td>
<td>Modified carbon budget model (8) considering environmental constraints</td>
<td>Age-related, stress-related</td>
<td>v.3.3: 9; for most recent version v.4.0.1 see 10, 11</td>
</tr>
<tr>
<td>FORMIND</td>
<td>Light-use efficiency (12)</td>
<td>Maintenance respiration + dynamic growth respiration</td>
<td>Dynamic allocation based on phenology, temperature, light and water availability</td>
<td>Allometric equations</td>
<td>Dependent on carbon allocation to stem mass and current DBH of the tree</td>
<td>Carbon-based stress mortality</td>
<td>13, 14, 15</td>
</tr>
<tr>
<td>3PG</td>
<td>Light-use efficiency (16)</td>
<td>Constant fraction of GPP</td>
<td>Dynamic allocation based on age, size, soil water, VPD</td>
<td>Allometric equation from DBH, competition, etc.</td>
<td>Dependent on carbon allocation to stem mass and current DBH of the tree</td>
<td>Age-dependent + stress-related + self-thinning</td>
<td>17, 18, 19</td>
</tr>
<tr>
<td>3PGN-BW</td>
<td>Light-use efficiency (16)</td>
<td>Maintenance respiration + dynamic growth respiration</td>
<td>Dynamic allocation based on environmental modifiers</td>
<td>Allometric equations</td>
<td>Dependent on carbon allocation to stem mass and current DBH of the tree</td>
<td>Age-dependent + stress-related + self-thinning with stochastic component</td>
<td>20, 21</td>
</tr>
<tr>
<td>BASFOR</td>
<td>Light-use efficiency</td>
<td>Fixed ratio NPP/GPP</td>
<td>Branch and stem fractions constant, leaf and root fractions functions of water- and nitrogen status</td>
<td>Function of stem dry matter</td>
<td>Function of stem dry matter and height</td>
<td>NA</td>
<td>22, 23</td>
</tr>
</tbody>
</table>
### TABLE 1 (Continued)

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

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 Durský (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: Augustyniczik and Yousefpour (2021); 22: van Oijen et al. (2014); 23: Cameron et al. (2013); 24: Mäkelä (2015); 25: Peltoniemi et al. (2015); 26: Mäkelä (1997); 27: Minunno et al. (2019); 28: Kalliokoski et al. (2019); 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

<table>
<thead>
<tr>
<th>Site</th>
<th>Dominant species</th>
<th>Forest type</th>
<th>MAP (mm/year)</th>
<th>MAT (°C)</th>
<th>Elevation (m a.s.l)</th>
<th>Country</th>
<th>Lat.</th>
<th>Long.</th>
<th>Structure</th>
<th>Flux</th>
</tr>
</thead>
<tbody>
<tr>
<td>Solling beech</td>
<td>Fagus sylvatica</td>
<td>Even-aged</td>
<td>1113</td>
<td>6.8</td>
<td>500</td>
<td>DE</td>
<td>51.77</td>
<td>9.57</td>
<td>1967–2014</td>
<td>NA</td>
</tr>
<tr>
<td>Le Bray</td>
<td>Pinus pinaster</td>
<td>Even-aged</td>
<td>920</td>
<td>13.4</td>
<td>61</td>
<td>FR</td>
<td>44.72</td>
<td>−0.77</td>
<td>1986–2009</td>
<td>1996–2008</td>
</tr>
</tbody>
</table>

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.
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). DBH 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. DBH 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):

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

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

$$\text{SB} = (\bar{X} - \bar{Y})^2,$$ (2)

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

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

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

with $r^2 = \frac{(\sum_{n=1}^{N} X_n Y_n)^2}{\sum_{n=1}^{N} x_n^2 \sum_{n=1}^{N} y_n^2}$ 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:

$$\text{norm. MSD} = \frac{\text{MSD}}{\frac{1}{N} \sum_{n=1}^{N} (Y_n - \bar{Y})^2}.$$ (5)

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 $R^2 = 1 - \text{norm. MSD}$ 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 ($R^2_{\text{structure}}$ and $R^2_{\text{carbon and water}}$) and re-projected the resulting values to the range from 0.1 to 1 to derive $A_{\text{structure}}$ and $A_{\text{carbon and water}}$. Overall $A$ was then derived analogous to $A_{\text{structure}}$ and $A_{\text{carbon and water}}$, 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 gam (Fasiolo et al., 2017). The quantile GAMs have the form

$$\text{norm. SB} = \frac{\text{SB}}{\frac{1}{N} \sum_{n=1}^{N} (Y_n - \bar{Y})^2}.\quad (6)$$

$$\text{norm. LC} = \frac{\text{LC}}{\frac{1}{N} \sum_{n=1}^{N} (Y_n - \bar{Y})^2}.\quad (7)$$

$$\text{norm. NU} = \frac{\text{NU}}{\frac{1}{N} \sum_{n=1}^{N} (Y_n - \bar{Y})^2}.\quad (8)$$
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

$$GPP = f_1(\text{temp}) + f_2(\text{rad}) + f_3(\text{vpd}) + f_4(\text{temp}, \text{rad}) + f_5(\text{temp}, \text{vpd}) + f_6(\text{rad}, \text{vpd}) + f_7(\text{temp}, \text{rad}, \text{vpd}),$$  \hspace{1cm} (9)

For each environmental variable, we re-projected the mean absolute slope across all models and sites $|\alpha_i|$ to the range between 0 and 1 (1 to 1) to account for differences in the magnitude of the variable units (temp: °C; rad: J/cm²; vpd: kPa). Then, we derived the realism of environmental responses for each model as the mean of the re-projected slope $\left(\frac{|\alpha_1| + |\alpha_2| + |\alpha_3|}{3}\right)$ 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 3PNG-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...
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.
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 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 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, PREBAS, and SIBYLA for BA, 3PGN-BW for GPP, and Hyytiälä indicated unrealistic productivity responses.

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 S3; 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 ($|a_i|$) (re-projected mean absolute slope, $|\bar{a_i}|$) | Realism of environmental responses |
|---------------------|-----------------------------------------------------------------------------|-----------------------------------|
| ensemble mean       | temp: $1.887 \times 10^{-9}$ (0.601), rad: $0.913 \times 10^{-9}$ (0.000), vpd: $4.488 \times 10^{-9}$ (0.511) | 1.00                              |
| Landscape-DNDC      | temp: $2.121 \times 10^{-9}$ (0.716), rad: $1.587 \times 10^{-11}$ (0.677), vpd: $1.686 \times 10^{-8}$ (0.000) | 0.70                              |
| 3D-CMCC-FEM BGC     | temp: $1.376 \times 10^{-9}$ (0.352), rad: $1.396 \times 10^{-11}$ (0.485), vpd: $3.847 \times 10^{-8}$ (0.612) | 0.63                              |
| GOTILWA+            | temp: $0.654 \times 10^{-9}$ (0.000), rad: $1.909 \times 10^{-11}$ (1.000), vpd: $3.856 \times 10^{-8}$ (0.615) | 0.45                              |
| PREBAS              | temp: $1.602 \times 10^{-9}$ (0.462), rad: $1.908 \times 10^{-11}$ (0.998), vpd: $2.631 \times 10^{-8}$ (0.268) | 0.33                              |
| BASFOR              | temp: $1.351 \times 10^{-9}$ (0.340), rad: $1.319 \times 10^{-11}$ (0.408), vpd: $5.215 \times 10^{-8}$ (1.000) | 0.31                              |
| 3D-CMCC-FEM LUE     | temp: $1.865 \times 10^{-9}$ (0.590), rad: $1.412 \times 10^{-11}$ (0.501), vpd: $4.412 \times 10^{-8}$ (0.772) | 0.18                              |
| 4C                  | temp: $2.705 \times 10^{-7}$ (1.000), rad: $1.198 \times 10^{-11}$ (0.286), vpd: $3.995 \times 10^{-8}$ (0.654) | 0.10                              |

Note: The mean absolute slope and re-projected mean absolute slope in brackets (see Equations 10–12 and $|\bar{a_i}|$ as well as $|\bar{a_i}|$ 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 $|\bar{a_i}|$ 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).
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

---

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

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ensemble mean</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>3D-CMCC-FEM BGC</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>3D-CMCC-FEM LUE</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Landscape-DNDC</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>ForClim v.3.3</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>3PG</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>3PGN-BW</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>SALEM</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>4C</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>FORMIND</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>SIBYLA</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>PREBAS</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>BASFOR</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>GOTILWA+</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>cover Europe (%)</td>
<td>3.59</td>
<td>1.05</td>
<td>4.12</td>
<td>0.35</td>
<td>0.97</td>
<td>0.44</td>
<td>10.55</td>
<td>0.45</td>
<td>0.20</td>
<td>3.05</td>
<td>0.28</td>
</tr>
</tbody>
</table>

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).
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 manipulative 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., Paschal 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 3PNC-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 (Okawa 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 (Dillmök 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 Hytiälä, 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 mis-
mismatch in the modeled 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., Collaengo 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 ra-
diation 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), espe-
cially 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 deter-
mine 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 esti-
mates 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 ver-
sion could explain the worse performance at the annual scale. The
issue related to the temporal scale in modeling GPP has already been
discussed by Collalbi 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 individ-
ual models that cancel out when aggregated into an ensemble mean.
Nevertheless, the ensemble mean’s performance relative to individ-
ual 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 rela-
tively 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 condi-
tions 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 stud-
ies covering different European forests. The ensemble covers almost
all European forest tree species because the individual models com-
plement each other especially for the less common tree species.

However, as forests may become more species rich and struc-
turally 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
and Menzel, 2019). Although the most important European species
in projected future abundance are already covered by the models
(Prunus sylvestris, Picea abies, Quercus spp., Fagus sylvatica), shifting dis-
turbance 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 theo-
retically 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 re-
sponses 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 real-
ism 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., Irauschenk 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 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 CO2 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, Ilbom, 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 parameterization 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 16KQ05). FH acknowledges funding by the Bavarian Ministry of Science and the Arts in the context of Bavarian Climate Research Network.
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 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 Dalmonetch 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/0000-0002-6255-9059
Mikko Peltoniemi https://orcid.org/0000-0001-6561-1943
Tommaso Pungetti https://orcid.org/0000-0002-1341-921X
Alessio Collalti https://orcid.org/0000-0002-1932-7550
David Cameron https://orcid.org/0000-0001-6902-2257
Ettore D’Andrea https://orcid.org/0000-0002-5884-210X
Alessio Collalti https://orcid.org/0000-0001-6561-1943

REFERENCES


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