



**HAL**  
open science

## Machine learning for accelerating process-based computation of land biogeochemical cycles

Yan Sun, Daniel S Goll, Yuanyuan Huang, Philippe Ciais, Ying-ping Wang,  
Vladislav Bastrikov, Yilong Wang

► **To cite this version:**

Yan Sun, Daniel S Goll, Yuanyuan Huang, Philippe Ciais, Ying-ping Wang, et al.. Machine learning for accelerating process-based computation of land biogeochemical cycles. *Global Change Biology*, 2023, 29 (11), pp.3221 - 3234. 10.1111/gcb.16623 . hal-04097612

**HAL Id: hal-04097612**

**<https://hal.science/hal-04097612>**






Submitted on 15 May 2023

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

## TECHNICAL ADVANCE

# Machine learning for accelerating process-based computation of land biogeochemical cycles

Yan Sun<sup>1,2</sup>  | Daniel S. Goll<sup>2</sup>  | Yuanyuan Huang<sup>3</sup>  | Philippe Ciais<sup>2</sup> |  
Ying-Ping Wang<sup>3</sup>  | Vladislav Bastrikov<sup>4</sup> | Yilong Wang<sup>5</sup> 

<sup>1</sup>College of Marine Life Sciences, Ocean University of China, Qingdao, China

<sup>2</sup>Laboratoire des Sciences du Climat et de l'Environnement, CEA-CNRS-UVSQ, Gif sur Yvette, France

<sup>3</sup>CSIRO Environment, Aspendale, Australia

<sup>4</sup>Science Partners, Paris, France

<sup>5</sup>State Key Laboratory of Tibetan Plateau Earth System, Resources and Environment (TPESRE), Institute of Tibetan Plateau Research, Chinese Academy of Sciences, Beijing, China

## Correspondence

Daniel S. Goll, Laboratoire des Sciences du Climat et de l'Environnement, CEA-CNRS-UVSQ, Gif sur Yvette, France.  
Email: [dsgoll123@gmail.com](mailto:dsgoll123@gmail.com)

Yuanyuan Huang, CSIRO Environment, Aspendale 3195, Australia.  
Email: [yuanyuan.huang@csiro.au](mailto:yuanyuan.huang@csiro.au)

## Funding information

Agence Nationale de la Recherche, Grant/Award Number: ANR-16-CONV-0003; GENCI-TGCC, Grant/Award Number: A0130106328; National Natural Science Foundation of China, Grant/Award Number: 42201107

## Abstract

Global change ecology nowadays embraces ever-growing large observational datasets (big-data) and complex mathematical models that track hundreds of ecological processes (big-model). The rapid advancement of the big-data-big-model has reached its bottleneck: high computational requirements prevent further development of models that need to be integrated over long time-scales to simulate the distribution of ecosystems carbon and nutrient pools and fluxes. Here, we introduce a machine-learning acceleration (MLA) tool to tackle this grand challenge. We focus on the most resource-consuming step in terrestrial biosphere models (TBMs): the equilibration of biogeochemical cycles (spin-up), a prerequisite that can take up to 98% of the computational time. Through three members of the ORCHIDEE TBM family part of the IPSL Earth System Model, including versions that describe the complex interactions between nitrogen, phosphorus and carbon that do not have any analytical solution for the spin-up, we show that an unoptimized MLA reduced the computation demand by 77%–80% for global studies via interpolating the equilibrated state of biogeochemical variables for a subset of model pixels. Despite small biases in the MLA-derived equilibrium, the resulting impact on the predicted regional carbon balance over recent decades is minor. We expect a one-order of magnitude lower computation demand by optimizing the choices of machine learning algorithms, their settings, and balancing the trade-off between quality of MLA predictions and need for TBM simulations for training data generation and bias reduction. Our tool is agnostic to gridded models (beyond TBMs), compatible with existing spin-up acceleration procedures, and opens the door to a wide variety of future applications, with complex non-linear models benefit most from the computational efficiency.

## KEYWORDS

biogeochemical cycles, computational demand, hybrid modeling, machine learning, terrestrial biosphere model

Daniel S. Goll and Yuanyuan Huang should be considered joint senior authors.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2023 The Authors. *Global Change Biology* published by John Wiley & Sons Ltd.

## 1 | INTRODUCTION

Along with the ever-growing large volumes of heterogeneous observational datasets in ecology (big-data), terrestrial biosphere models (TBMs) have been growing in complexity in order to model ecosystems as realistic as possible (big-model; Fisher & Koven, 2020). This evolution comes with an increasing number of model parameters and computational demand. High computational cost is now the bottleneck in the big-data-big-model era in global change ecology. It hampers the applications of model-data assimilation systems needed to optimize model parameters, the assessment of model uncertainties, the refinement of ecological processes, and model applications at fine spatiotemporal resolution. As a consequence, arguably little progress has been made regarding the reliability and robustness of these models (Prentice et al., 2015).

TBMs are process-based models that resolve ecological and physical processes on a wide range of intrinsic timescales (from hours to millenia), and their interactions. Major developments in the recent past have not only led to additional processes considered but also to a widening of the range of time-scales considered and the degree of non-linear interactions. An example is the refinement of soils (Wang & Goll, 2021). For example, phosphorus cycling links 'slow' chemical rock weathering with 'fast' leaf-level gas exchange (Ellsworth et al., 2022). This coupling increases the numerical computation time needed to bring the modeled biogeochemical cycles into a steady-state (spin-up), which is a common pre-requirement in most model applications and also the most time (computational resource) consuming step in model simulations (Thornton & Rosenbloom, 2005). In the case of the ORCHIDEE model, the version with nitrogen and phosphorus cycles requires up to 10,000 model years to reach a steady-state for carbon and nutrient pools and fluxes, while the version without nutrients requires 2000 model years (Sun et al., 2021). Typical model projections span only a few 100 years (e.g. the historical period 1700–present day), and thus the overall computational demand is dominated by the spin-up (85%–98% for models mentioned above).

Different approaches to accelerate the spin-up have been proposed, but they generally have two major shortcomings. They are usually model specific, and they rely largely on linearity to approximate analytical solutions and become inefficient with increasing non-linearities (e.g. Thornton & Rosenbloom, 2005; Xia et al., 2012). The recent development of TBMs, such as the microbial-explicit soil carbon modules, permafrost dynamics, individual based vegetation processes and nutrient relevant representations, brings in multiple non-linear dynamics. Earlier spin-up methodologies are inadequate in dealing with these increased complexities.

Here, we demonstrate the use of unsupervised and supervised machine learning (ML) to tackle this grand challenge through a combined procedure. ML does not require underlying assumptions on linearity or distributions of data, making them promising for ecological studies. Combining ML with TBMs to advance Earth system studies has been suggested as a research priority (Reichstein et al., 2019) but has proven challenging. Our novel study here takes the most resource consuming yet essential step in TBMs as an example to illustrate how

global change ecology could advance through merging ML with the big-data-big-model. Specifically, we build a ML-based procedure for accelerating the equilibration of biogeochemical coupled C, N and P cycles, which is general enough to be applicable to most TBMs.

## 2 | MATERIALS AND METHODS

The ML-enabled spin-up acceleration procedure (MLA) predicts the steady-state of biogeochemical pools in any land pixel after training on a representative subset of pixels (Figure 1, Section 2.1). As the computational efficiency of current TBMs (e.g. without lateral transfer fluxes) scales linearly with the number of pixels and years simulated, MLA reduces the computation time quasi-linearly with the number of pixels predicted by ML.

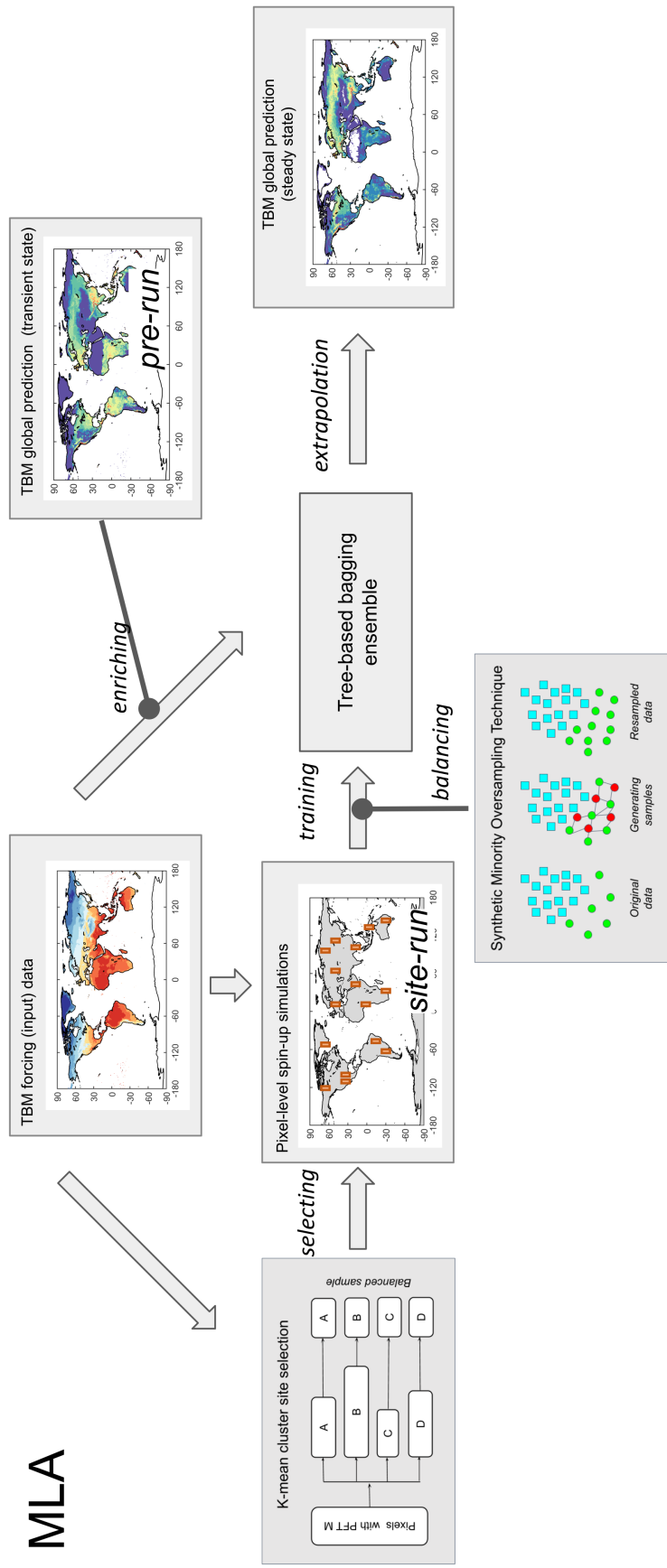
We demonstrate the feasibility of a pragmatic implementation of this approach on three different versions of the ORCHIDEE TBM family (Section 2.2) with varying degrees of complexity and nonlinearity. After evaluating the accuracy of MLA by comparison with the conventional spin-up (Section 2.3), and reducing biases through a short global simulation restarting from MLA equilibrated state (corrected-MLA 2.4), we further illustrated the impact of using corrected-MLA for initialization of TBM simulations for the historic period (Section 2.5). Finally, we estimated how much computational time was saved by the whole procedure (Section 2.6). The implementation of the approach in python for ORCHIDEE TBMs is freely available (<https://doi.org/10.5281/zenodo.7503092>).

### 2.1 | ML-based procedure for accelerated spin-up

The MLA procedure consists of four steps (Figure 1): (1) selecting a subset of representative land pixels as training samples using ML based on prior knowledge (Sections 2.1.1 and 2.1.2); (2) performing conventional spin-up for the selected training pixels; (3) building ML models to link biogeochemical pools at equilibrium (*Response Y*) from (2) with drivers of TBMs (*Predictor X*) for the training pixels (Section 2.1.3); (4) using built ML models to predict the biogeochemical pools at equilibrium for the entire spatial domain.

#### 2.1.1 | Predictors

The predictors consist of up to 27 variables, 20–25 variables depending on the model version characterizing its driving data (forcing; Data S1) and transient states of annual net primary productivity and leaf area index (Table 1). We found that the latter two can substantially improve the ML model performance compared to one trained on only the forcing data (not shown). Both variables are key variables to characterize carbon input to pools and its variation in space containing valuable information in their transient state to predict the equilibrated states. Those two variables are from a *pre-run* simulation with an arbitrary length of 300 years. The length was chosen



**FIGURE 1** Schematic of acceleration procedure for spin-up for terrestrial biosphere models (TBM) which involves a combination of TBM simulation and machine learning steps (machine learning acceleration; MLA).

TABLE 1 Predictors used in prediction model for ORCHIDEE-CNP. The climatic predictors, nutrient deposition predictors and LAI<sub>g</sub> and NPP<sub>g</sub> are yearly variables.

Abbreviations	Variables	Units	Forcing variable
$T_{amp}$	Amplitude of monthly temperature	°C	From 6 hourly data
$T_{max}$	Maximum monthly temperature	°C	From 6 hourly data
$T_{min}$	Minimum monthly temperature	°C	From 6 hourly data
$T_{mean}$	Mean monthly temperature	°C	From 6 hourly data
$T_{std}$	Standard deviation of monthly temperature	°C	From 6 hourly data
$T_{gs}$	Accumulated temperature during growing season (monthly temperature > -4°C)	°C	From 6 hourly data
Rainf	Mean annual precipitation	kg m <sup>-2</sup> year <sup>-1</sup>	From 6 hourly data
Rainf_std	Standard deviation of monthly precipitation	kg m <sup>-2</sup> year <sup>-1</sup>	From 6 hourly data
Rainf_gs	Precipitation during growing season (monthly $T > -4^{\circ}\text{C}$ )	kg m <sup>-2</sup> year <sup>-1</sup>	From 6 hourly data
$Q_{air}$	Near surface specific humidity	kg kg <sup>-1</sup>	From 6 hourly data
$P_{surf}$	Surface pressure	Pa	From 6 hourly data
SW <sub>down</sub>	Shortwave down radiation	W s <sup>-1</sup>	From 6 hourly data
LW <sub>down</sub>	Longwave down radiation	W s <sup>-1</sup>	From 6 hourly data
INT1	Rainf· $T_{mean}$	°C kg m <sup>-2</sup> year <sup>-1</sup>	From Rainf and $T_{mean}$
INT2	Rainf <sub>gs</sub> · $T_{gs}$	°C kg m <sup>-2</sup> year <sup>-1</sup>	From Rainf <sub>gs</sub> and $T_{gs}$
GSL	Growing season length	month	From 6 hourly data (temperature)
Clay	Clay fraction	%	From time invariant data
Silt	Silt fraction	%	From time invariant data
Soil <sub>pH</sub>	Soil pH (only ORCHIDEE-CNP)	—	From time invariant data
$N_{dep\ noy}$	Nitrogen deposition (NO <sub>y</sub> ; only ORCHIDEE-CNP)	gN m <sup>-2</sup> year <sup>-1</sup>	From annual data
$N_{dep\ nhx}$	Nitrogen deposition (NH <sub>x</sub> ; only ORCHIDEE-CNP)	gN m <sup>-2</sup> year <sup>-1</sup>	From annual data
$P_{dep}$	Phosphorus deposition (only ORCHIDEE-CNP)	gP m <sup>-2</sup> year <sup>-1</sup>	From annual data
Soil <sub>bulk</sub>	Soil bulk density	g soil cm <sup>-3</sup>	From time invariant data
Soil <sub>shield</sub>	Soil shield factor [0-1] (only ORCHIDEE-CNP)	—	From time invariant data
Soil <sub>suborder</sub>	Soil suborders (categorical variable; only ORCHIDEE-CNP)	—	From time invariant data
NPP <sub>g</sub>	Annual net primary productivity at the end of <i>pre-run</i>	gC m <sup>2</sup> year <sup>-1</sup>	From daily data
LAI <sub>g</sub>	Annual leaf area index at the end of <i>pre-run</i>	m <sup>2</sup> m <sup>-2</sup>	From daily data

here for demonstration purposes, but should be optimized for each model version based on the trade-off between ML model performance and computation costs of running a conventional TBM simulation, to minimize overall computational demand of the MLA.

## 2.1.2 | ML-enabled training site selection

To ensure a balance dataset, we combine a K-mean cluster method aiming for balanced *Predictor X* (Farquid & Bose, 2012) and Synthetic Minority Oversampling TEchnique (SMOTE) aiming for balanced *Response Y* (see section 2.3.2; Chawla et al., 2002). A detailed description is given in the Data S2.

## 2.1.3 | Building ML models to predict steady states

Bagging (known as bootstrap aggregation) decision trees is an ensemble ML method based on Breiman (1996) to avoid overfitting

issues through a single decision tree. In this study, we grew 100 trees in each ensemble for predicting the equilibrated state variables for each training pixel. Ninety percent of total training pixels were randomly selected to train each tree. Ninety percent was chosen here to incorporate the randomness for training, while not involving too many samples for training to save the overall computation time of the entire procedure. The minimum number of samples of every tree leaf is set as 5. We increased the weights for samples falling out of the 10th–75th quantiles to two to ten times of other samples to reduce the overestimation of high *Response Y* (Table S2). Different from the random forest which randomly selects a subset of the predictors (Breiman, 2001), we used the bagging decision trees method to incorporate all predictors to be in line with forcings of TBMs and traditional spin-up methodology.

Note that our MLA is not limited to the methodology of bagging decision trees. We use this method due to its adequate performance on our training samples and for demonstration purposes, while we acknowledge other ML methods (e.g. gradient boosting and multiple deep learning methods) also fit in our framework.

## 2.2 | Terrestrial biosphere models and simulation setup

### 2.2.1 | TBM ORCHIDEE family

We used three versions of the ORCHIDEE model: ORCHIDEE (v2.2) the main version (trunk) which was used in IPSL-ESM contributing to the coupled model intercomparison project Phase 6 (CMIP6); ORCHIDEE-CNP v1.2 a (branch) version which resolves nitrogen and phosphorus cycles (Goll et al., 2017), and ORCHIDEE-CNP v1.3 an update of the branch version which resolves non-linear microbial dynamics instead of deploying a linear first order decay model like the other two versions (Goll et al., 2022; Table 2). The three versions reflect the general tendency in the TBM community to refine processes operating on various timescales: for example, microbial dynamics on timescales of minutes to months, while soil phosphorus dynamics on timescales up to thousands of years. The number of state variables describing the biogeochemical cycles in soil, litter and plants depends on the model version and ranges between 240 and 825 (Table 2; Data S3). All three versions deploy a tiling approach in which each model pixel contains information on biogeochemical cycles of multiple PFTs (each PFT has a specific vector of parameters) irrespectively of their actual land cover.

### 2.2.2 | Simulation setup

We used protocols from two model intercomparison projects (Table 2) to perform the TBM simulations. Both protocols aim at

reconstruction of historic changes in the land biogeochemistry (Global N<sub>2</sub>O Model Intercomparison Project (NMIP) Phase 2 (Tian et al., 2018), and Global Carbon Budget: Land modeling protocol Trendy version 10 (Friedlingstein et al., 2022)) but are slightly different. The forcing information is described in more detail in Data S4.

Two general types of simulations were performed with each of the ORCHIDEE versions and the corresponding forcing data according to the abovementioned protocols (Table 3). Simulations prescribing constant boundary conditions for the initial year (1700 or 1860), and simulations with varying boundary conditions reflecting their changes over the historic period. The first type of simulations is needed to generate the training information for the ML (Section 2.1), namely a short global simulation (*pre-run*) starting from scratch (low globally uniform values) to derive the transient states of NPP<sub>g</sub>, LAI<sub>g</sub>, as well as pixel level conventional TBM simulations for the selected representative model pixels starting from scratch to steady state (see Section 2.3) in which the biogeochemical cycles were equilibrated (*site-runs*). The second type of simulation is needed for the quantification of the impact of the use of ML for equilibration of biogeochemical cycles on the outcome of typical TBM simulations which are initialized from them (Section 2.5). We performed two simulations with each version of ORCHIDEE for the historic period which either started from (1) the “*true equilibrium*” state from the conventional spin-up (*production-run<sub>conv</sub>*) or from (2) the corrected-MLA (see Section 2.4) generated equilibrated state (*production-run<sub>MLA</sub>*).

TABLE 2 The three different versions of the ORCHIDEE model family and their key features relevant for this study.

Model version	ORCHIDEE v2.2	ORCHIDEE-CNP v1.2	ORCHIDEE-CNP v1.3
References	Krinner et al. (2005)	Sun et al. (2021)	Zhang et al. (in prep)
Nutrient cycles	No	Yes (nitrogen, phosphorus)	Yes (nitrogen, phosphorus)
Microbial dynamics	No	No	Yes (two microbial classes)
Number of PFTs	15	15	15
Number of state variables per pixel	240	825	915
Spatial resolution	2 × 2°	0.5 × 0.5°	2 × 2°
Spin-up acceleration procedure	Yes, (quasi) analytical solution	No	No
Number of years needed to equilibrium criteria	340	~10,000	~6500
Simulation setup	Trendy v10	NMIP v2	Trendy v10

TABLE 3 Specifies the different types of land surface model simulation, and the respective lengths for the different versions of the ORCHIDEE model family.

Acronyms	<i>pre-run</i>	<i>site-run</i>	<i>re-run</i>	<i>spin-up<sub>conv</sub></i>	<i>production-run<sub>mla</sub></i>	<i>production-run<sub>conv</sub></i>
Domain	Global	Pixel	Global	Global	Global	Global
Length in model years (ORCv2.2; ORC-CNPv1.2; ORC-CNPv1.3)	300; 300; 300	340; 10,000; 6500	100; 350; 310	340; 10,000; 6500	320; 167; 320	320; 167; 320
Boundary conditions	Constant	Constant	Constant	Constant	Variable	Variable
Starting point	Scratch	Scratch	ML	Scratch	<i>re-run</i>	<i>spin-up<sub>conv</sub></i>

## 2.3 | Evaluation of MLA

In order to test the accuracy of the MLA-predicted state variables of the TBMs, we compared them with state variables derived from a full spin-up simulation ( $spin-up_{conv}$ ) which reflects the 'true equilibrium' (i.e. annual changes in global land carbon storage are  $<0.05 \text{ Gt year}^{-1}$  when averaged over a 50 year period; Friedlingstein et al., 2022). To assess the accuracies of the ML-based approach in reproducing the pool sizes at steady state we used coefficient of determination ( $R^2$ ), relative bias (rs), normalized root mean squared error by the difference between maximum and minimum (NRMSE), and the regression slope between the results from "true equilibrium" and MLA-based one (slope). rs for a given pool is defined as

$$rs = \frac{I_{ML} - I_{equi}}{I_{equi}} \times 100\%, \quad (1)$$

while NRMSE is defined as

$$NRMSE = \frac{RMSE}{\max(I_{equi}) - \min(I_{equi})}. \quad (2)$$

RMSE is the root mean square error, defined as

$$RMSE = \sqrt{\frac{\sum_{j=1}^N (I_{ML,j} - I_{equi,j})^2}{N}}, \quad (3)$$

where  $I_{ML,j}$  and  $I_{equi,j}$  are values from MLA and  $spin-up_{conv}$  for sample  $j$  of a total sample size of  $N$ , respectively.

## 2.4 | Re-run after ML predictions till equilibrium

ML inevitably yields biases in predictions. Thus, we use a conventional TBM simulation with the same boundary conditions as used for the training starting from the ML-based equilibrated state to reduce biases (*re-run*), with a length such that the requirement for an equilibrium is reached (Section 2.3). After *re-run*, we redo the evaluation as mentioned in Section 2.3. We refer to the equilibrium state after *re-run* as corrected-MLA.

## 2.5 | Impact of bias from MLA on historical C balance

To assess the impact of bias of the corrected-MLA on the outcome of typical TBM simulations, we compared simulation initialized from them ( $production-run_{MLA}$ ) with simulations, which were initialized from a conventional spin-up ( $production-run_{conv}$ ). Both simulations were forced by the same historical forcing (see Section 2.2.2), thus differences in the predicted biogeochemical cycles are solely caused by differences in the initial state. We focused on the global spatiotemporal pattern of net biome productivity (NBP), a key variable in global carbon studies (e.g. Friedlingstein et al., 2022).

## 2.6 | Computational time savings

The computational demand of the current generation of TBMs scales quasi-linearly with the number of simulated model pixels and years. This is due to the fact that the pixels in current TBMs are relatively independent from each other. Operationally, there are minor deviations from this, for example due to reading and writing of data, which requires information from multiple model pixels at a time. The computational demand of the site selection, training ML-models, and extrapolation using ML is negligible compared to global TBM simulations. Based on these assumptions, we approximate the computational demand ( $D$  in %) for each type of simulation as the product of the number of model pixels ( $p$ ) and the number of years relative to a  $spin-up_{conv}$  simulation ( $y$ ),

$$D_i = \frac{y_i \times p_i}{y_j \times p_j} \times 100\%, \quad (4)$$

where  $i$  refers to *pre-run*, *site-run*, *re-run*, and *production-run* and  $j = spin-up_{conv}$ . Note that the actual time savings depend on the machine infrastructure and software and might deviate from the theoretical one (Data S5).

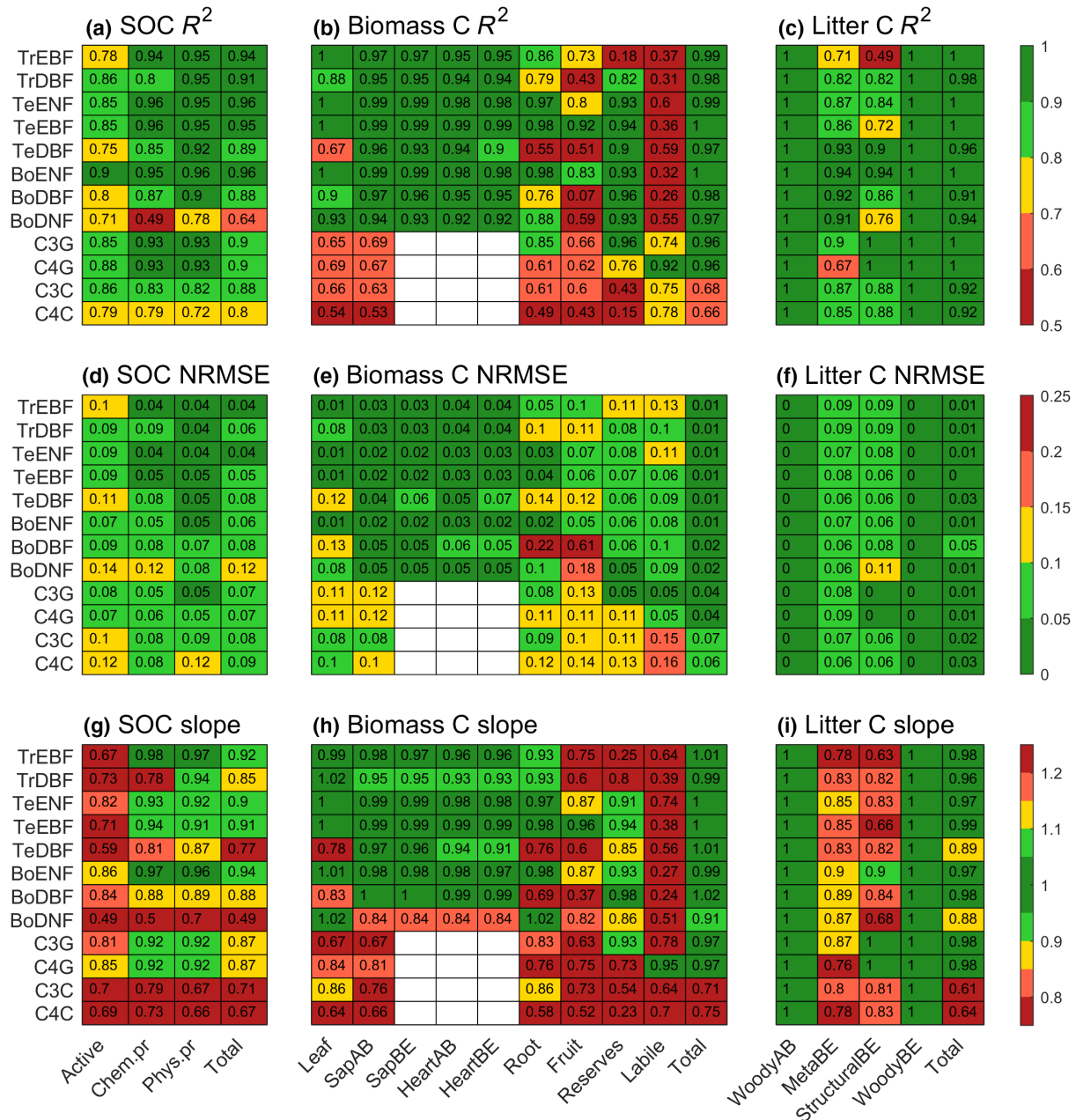
Based on the reported length of a  $spin-up_{conv}$  for a precursor version of ORCHIDEE (Krinner et al., 2005), we approximate the time consumption of a  $spin-up_{conv}$  for ORCHIDEE v2.2 without the use of the version-specific acceleration procedure instead of performing such a simulation.

## 3 | RESULTS AND DISCUSSION

### 3.1 | Evaluation of steady states predicted by MLA

Equilibrated state of model state variables from MLA (before *re-run*) at PFT level over the whole spatial domain compares moderately well with the state from  $spin-up_{conv}$  depending on the model version (Figures 2–4, Figures S3–S6). The slopes are commonly lower than one which is a general issue with ML which tends to overestimate low values but underestimate high values (Belitz & Stackelberg, 2021). The distribution of ML-predicted size of C (N,P) in biomass, litter and soil organic matter on pixel level is however comparable to the one from  $spin-up_{conv}$  for all three model versions (Figure 5, Figures S7 and S8). There is no indication of differences in the performance among PFT which are common to all three model versions. Among all model versions, pools with a small (few years and shorter) residence show the highest biases, namely active soil organic matter, labile, fruits and grass biomass pools (Figures 2–4, Figures S3–S6). We find no systematic biases in the spatial pattern of soil carbon, a key pool that requires long spin up for equilibration (Figures S9–S11), and the two model versions deploying the same soil organic matter module (ORCHIDEE v2.2, and ORCHIDEE-CNP v1.2) show similar bias patterns (Figures S9 and S10).





**FIGURE 2** The performances of machine learning models (ML) on all carbon cycle state variables, with the sum of state variables of major biosphere compartment ((a,d,g) soil organic carbon (SOC), (b,e,h) biomass, and (c,f,i) litter) for each plant functional type compared to the 'true equilibrium' of the conventional spin-up simulation ( $spin-up_{conv}$ ). Three statistics represent the model performance: (a–c) coefficient of determination ( $R^2$ ), (d–f) normalized root mean squared error (NRMSE), and (g–i) the regression slope between the results from ML and  $spin-up_{conv}$  (slope). Shown are results from ORCHIDEE-CNP v1.3. The plant functional types are: Tropical Evergreen Broadleaf Forest (TrEBF), Tropical Deciduous Broadleaf Forest (TrDBF), Temperate Evergreen Needleleaf Forest (TeENF), Temperate Evergreen Broadleaf forest (TeEBF), Temperate Deciduous Broadleaf Forest (TeDBF), Boreal Evergreen Needleleaf Forest (BoENF), Boreal Deciduous Broadleaf Forest (BoDBF), Boreal Deciduous Needleleaf Forest (BoDNF), C3 grassland (C3G), C4 grassland (C4G), C3 cropland (C3C), and C4 cropland (C4C).

The procedure works best for the carbon only model version with the majority of  $R^2$  above 0.9, slopes between 0.9 and 1.0, and NRMSE smaller than 0.1 (Figure S6). The quality of ML predictions is the lowest for ORCHIDEE-CNP v1.2 with  $R^2$  as lower than 0.7 for 46% of all C state variables at PFT-level (Figures S3–S5). This is likely due to the fact a larger number of model pixels has not yet reached a

steady-state in the  $site-runs$  and  $spin-up_{conv}$  than in other two model versions (not shown). This is due to the chosen equilibrium criteria which focus on global C stocks rather than on pixel-level ones (see Section 2) and is thus insufficient to ensure that (1) state variables of all model pixels are at equilibration, (2) the nutrient cycles are in equilibrium.



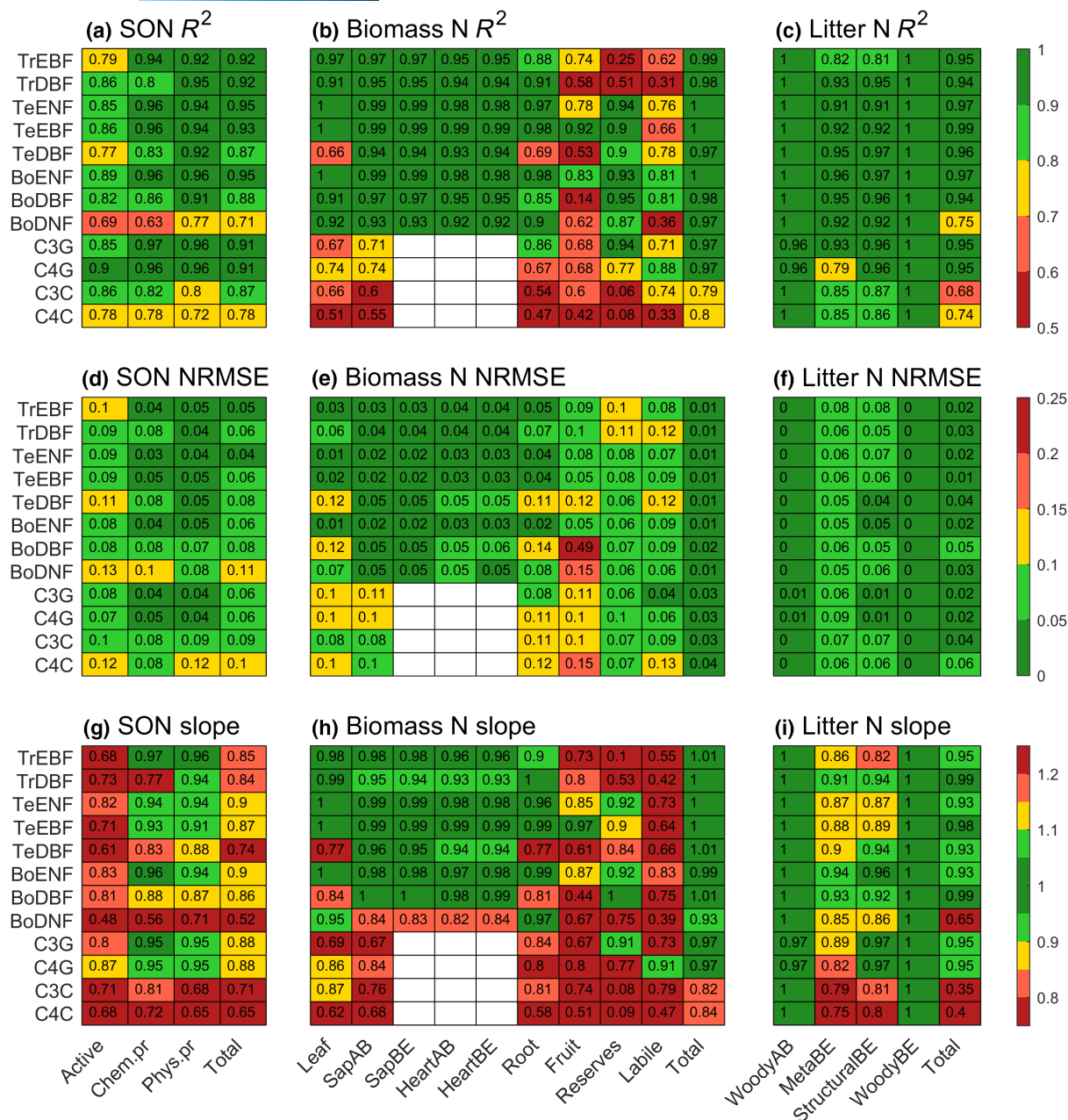


FIGURE 3 Same as Figure 2 but for nitrogen cycle state variables.

### 3.2 | Predictor importance and representativeness of training sites

Considering the biases in ML predicted state variables (Figures 2–4), it is necessary to understand the factors influencing the ML performances. To do so, we ranked the predictor importance for predictions at the example of the passive SOC at the PFT-level for ORCHIDEE-CNP v1.2 which shows the highest biases (Figure S3). The size of this pool is mainly affected by plant productivity ( $LAI_g$  and  $NPP_g$ ) and climate predictors for all PFTs (Figure 6), whereas edaphic predictors are among the top predictors for some PFTs. The ranking of factors is in line with the theoretical understanding of

drivers of SOC stocks of the underlying type of soil organic matter model (Huang et al., 2018).

To test the representativeness of training sites, we perform an additional ranking of predictor importance this time from a training using all pixels of the global domain (Figure S12). We found that top five key predictors in this ranking are generally the same as in the training using a subset of pixels (Figure 6) indicating a sufficient representativeness of the selected sites. However, for boreal forests and grasslands we found differences (i.e. edaphic factors ranked highest), indicating a low representativeness of the selected training sites, which coincides with a low performance of the MLA. At the example of BoENF PFT we show that the biases in ML-predicted SOC

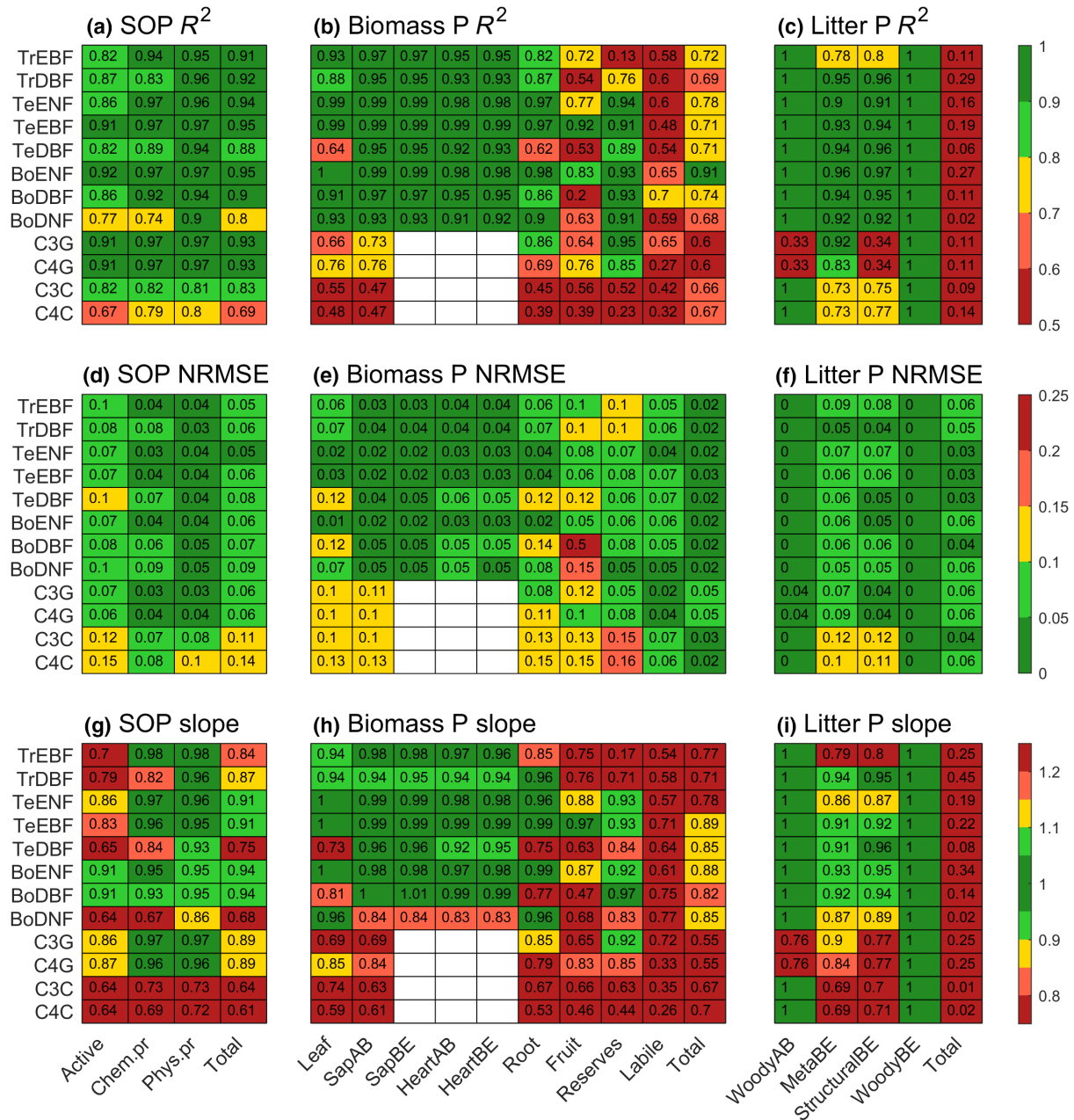


FIGURE 4 Same as Figure 2 but for phosphorus cycle state variables.

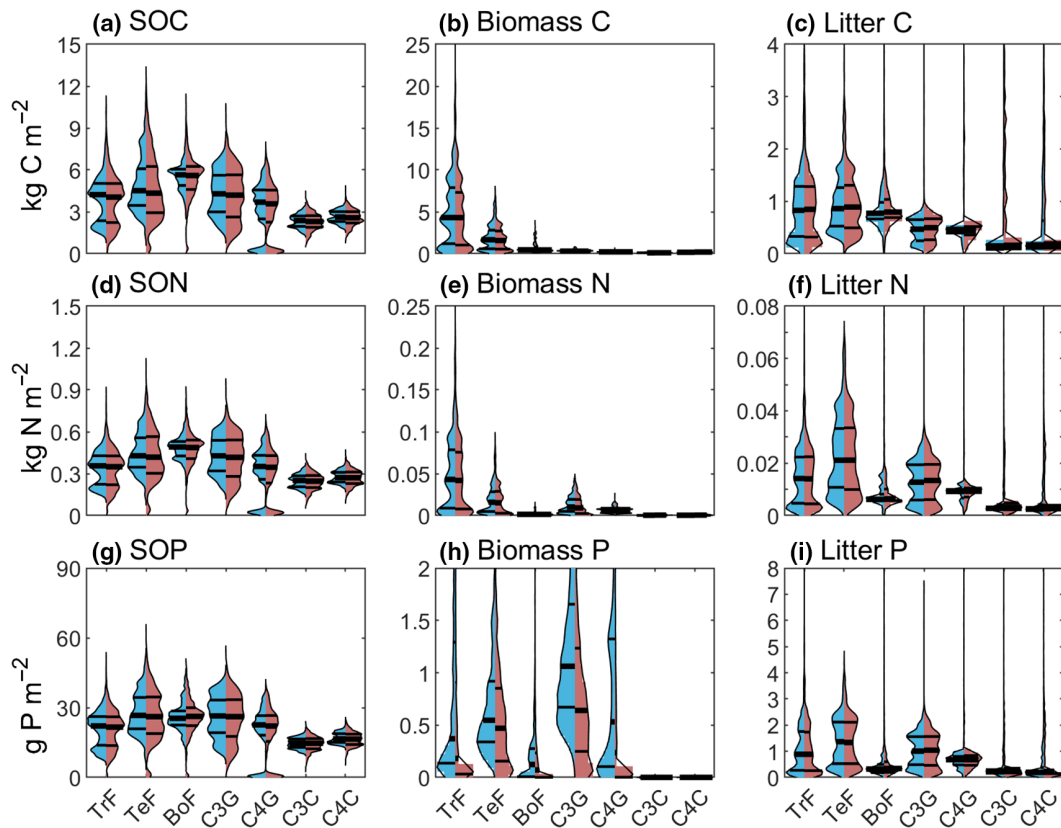
pools are highest for pixels with rare soil suborders (i.e. Andisols, Entisols, Gelisols, and Oxisols; Figure S13) in the training dataset, which provides additional evidence that the selected sites insufficiently cover the variation in edaphic conditions.

### 3.3 | Corrected-MLA and *re-run*

To reduce biases in the ML-predicted states, we performed *re-run* simulations. In order to reduce the drift in the global land carbon stock to be less than 0.05 Gt year<sup>-1</sup> over five decades, 100, 350 and 310 years were needed for ORCHIDEE v2.2., ORCHIDEE-CNPv1.2, and ORCHIDEE v1.3 (Figure S14). As expected, the poorer the performance of the ML

prediction (more information Data S6), the longer the length of the *re-run* simulations. On a pixel-level, drifts (linear trend) during the last 50 years of the *re-run* simulation remain more pronounced than at the end of the spin-up<sub>conv</sub> in particular for soil carbon stocks (Figures S15 and S16), illustrating the limitation of this approach.

We find that only for ORCHIDEE v2.2, biases in the ML predicted model state variables at PFT level are consistently reduced among all pixels (Figure S17), while for the other two versions the biases in biomass increase for boreal needle leaf PFTs (Figures S18 and S19), irrespective of the area cover fraction (Figures S20 and S21). Interestingly, the bias increase during the *re-run* occurs often at points with low initial bias (Figures S22–S25) and for PFTs with reduced representatives of training sites (Section 3.2). This



**FIGURE 5** Distribution of carbon (C), nitrogen (N) and phosphorus (P) in (a,d,g) soil organic matter, (b,e,h) biomass, and (c,f,i) litter for seven biomes for ORCHIDEE-CNP v1.3. Shown are results from the *spin-up<sub>conv</sub>* (blue) and the ML prediction (red). The biomes are tropical forest (TrF; consisting of PFTs 2 & 3), temperate forest (TeF; PFTs 3, 4 & 6), boreal forest (BoF; PFTs 7, 8 & 9), C3 grassland (C3G; PFT 10, 14 & 15), C4 grassland (C4G; PFT 11), C3 cropland (C3C; PFT 12), C4 cropland (C4C; PFT 13). The thick black horizontal lines indicate the median values, while thin ones indicate the 25th and 75th quantiles. For PFT identifications see [Table S1](#).

behavior could be either caused by the issue with training sites or by the allocation scheme which relies on multiple nested thresholds to control plant allometry (Goll et al., 2017), which might set targets which are different from the ones in the *spin-up<sub>conv</sub>* due to the slight deviation in the relative distribution of plant biomass among pools in the MLA. These thresholds may lead to a new steady state biomass.

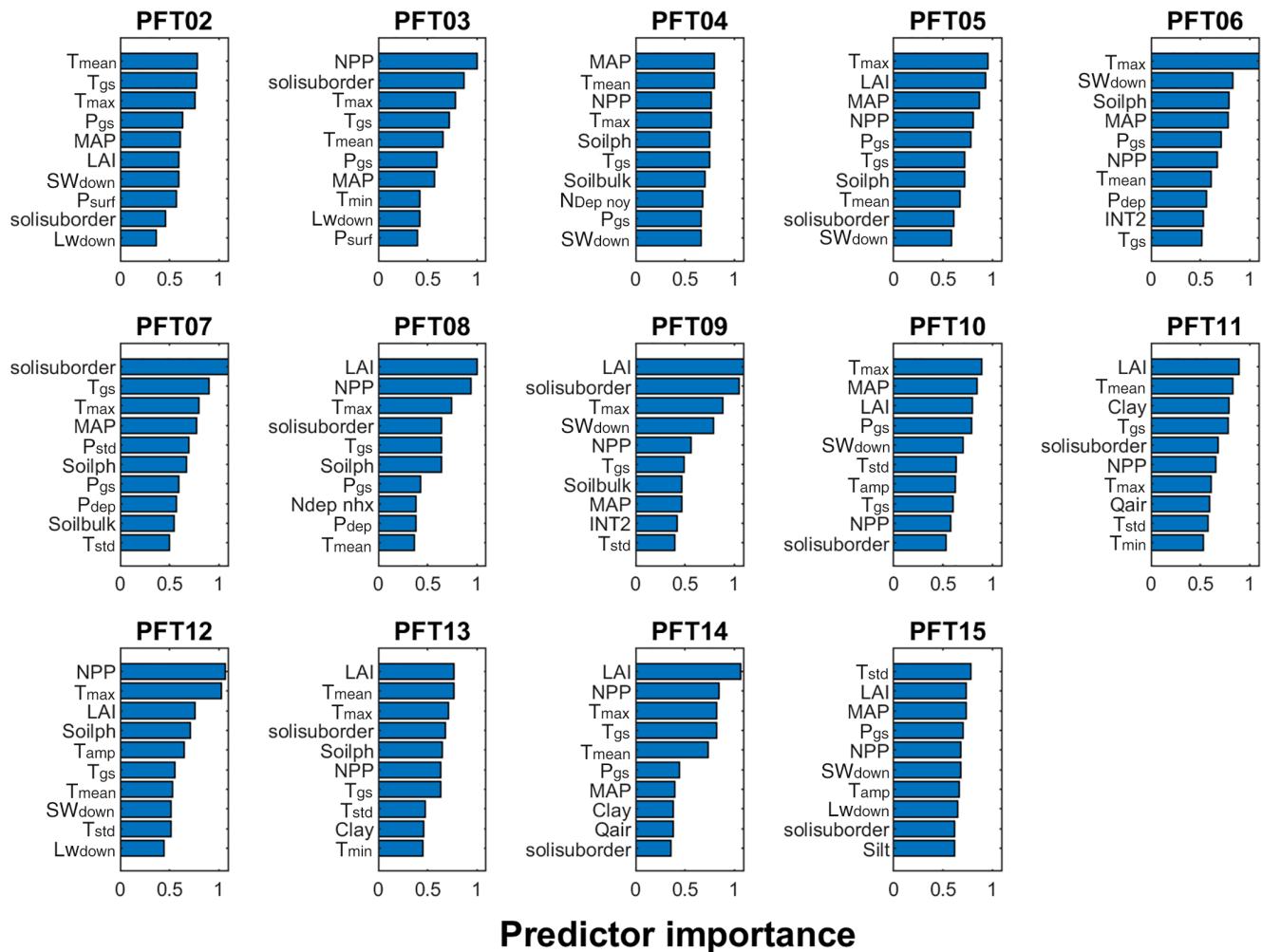
### 3.4 | Computational time savings

The MLA procedure as a whole consumes 78%, 80%, and 78% less computational time compared to a conventional spin-up without any additional acceleration procedure for ORCHIDEE v2.2, ORCHIDEE-CNP v1.2, and ORCHIDEE-CNPv1.3, respectively (Figure 7). The much lower computational demand of the spin-up for model versions of different structural complexity, was achieved with a common procedure and without any optimization of the lengths of the *pre-run*, *re-run* and *site-run* simulations as well as the number of pixels selected for the *site-run* simulation. Therefore, we expect a one-order of magnitude lower computation demand is in reach, for example by optimizing these parameters to specific model versions.

When deployed with an existing version-specific acceleration procedure of ORCHIDEE v2.2, the computational demand of the spin-up is halved compared to the one when using only the version-specific acceleration procedure. This illustrates the versatility of the MLA approach which can be combined with other acceleration procedures to further reduce the computational demand.

### 3.5 | Impact of remaining biases from corrected-MLA on the historical TBM simulation

The use of MLA inevitably introduces errors in the predicted model state variables at steady-state, which affect the results of the *production-run*. The question is whether such errors will lead to significant differences in the *production-run<sub>MLA</sub>* compared to *production-run<sub>conv</sub>* or not. To address this question, we assessed differences in the spatiotemporal patterns of net biome productivity (NBP) which is a key output variable of TBMs. NBP is defined as the net C exchange between the atmosphere and the terrestrial biosphere (a positive sign indicates a net land uptake) and the balance of multiple biologically controlled fluxes and ones caused by disturbances.



**FIGURE 6** Predictor importance for machine learning predicting model trained on a subset of all global pixels (13.5%). Exemplary results are shown for the predicting passive SOC pool at the level of plant functional types (PFT) in the case of ORCHIDEE-CNP v1.2. For detailed information on PFTs and predictors see [Table 1](#) and [Table S1](#) respectively.

We found that global and regional predictions for annual NBP during the last six decades by the three tested versions of the ORCHIDEE family are only marginally affected by the use of the MLA derived steady-state compared to the one of a conventional spin-up ([Figure 8](#)). The deviations in NBP between *production-run<sub>MLA</sub>* and *production-run<sub>conv</sub>* result in negligible or small differences in land carbon stock when accumulated over the period 1950–2010 of 0.0, 3.0, and -1.2 Gt (C) for the ORCHIDEE v2.2, ORCHIDEE-CNP v1.2 and ORCHIDEE-CNP v1.3, respectively.

Global pixel-to-pixel comparisons in NBP for single years (1990, 2000 and 2010) between *production-run<sub>MLA</sub>* and *production-run<sub>conv</sub>* show high spatial  $R^2$  (0.87–0.99) and low RMSE ( $<17 \text{ g C m}^{-2} \text{ year}^{-1}$ ) for all three model versions ([Figures S26–S28](#)). The spatial patterns of NBP are similar between *production-run<sub>MLA</sub>* and *production-run<sub>conv</sub>* without large biases in any regions ([Figures S29–S31](#)) indicating the absence of systematic errors in NBP.

This illustrates that the accuracy of the MLA is sufficiently high for regional–global applications. The impact is substantially smaller than the differences among model versions which reflect the impact of

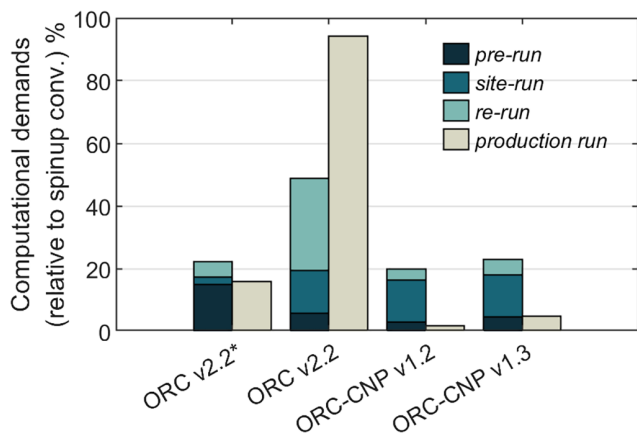
uncertainties in model structure and parameterizations. There are deviations at pixel level, which are due to the biases in the MLA predicted state of certain PFTs. The drifts in carbon stocks at PFT or pixel level during the *re-run* type simulations are indicators where such deviations can occur, and targeted measures should then be used to improve the MLA. It should be noted that the impact of the use of MLA on target variables depends on their sensitivities on the initial pool sizes and errors of ML.

### 3.6 | Implications for the studies of global change biology

#### 3.6.1 | Towards high spatial resolution TBMs

There is an increasing demand to develop and apply TBMs with ever increasing spatiotemporal resolutions. The improved computation efficiency here is sufficient to allow predictions at a spatial resolution of 10 km globally (1 km regionally) with ORCHIDEE family models. This offers advanced ecological and biogeochemical information

to satisfy societal and management needs and allows new areas of scientific applications. It further reduces the gap between resolutions of TBM and satellite-based earth observations.

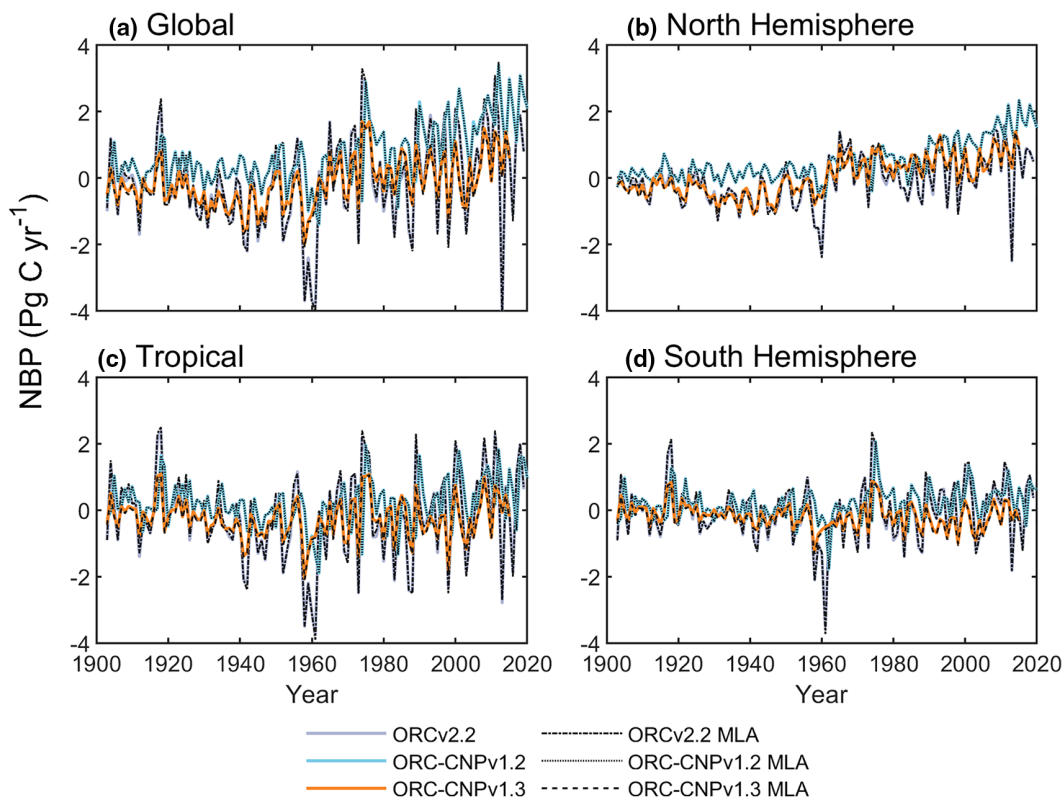


**FIGURE 7** The computational demand of the historical simulation type (*production-run<sub>MLA</sub>* or *production-run<sub>conv</sub>*) and the MLA based spin-up (consisting of *pre-run*, *site-run<sub>MLA</sub>*, and *re-run*) as percentage of the demand of the conventional spin-up type (*spin-up<sub>conv</sub>*) for three versions of the ORCHIDEE family. In the case of ORCHIDEE v2.2 we show the computational demands when the version-specific spin-up acceleration procedure was activated (ORC v2.2, specific + MLA) or not (ORC v2.2\*, MLA-only). ORC v2.2 shows computation time relative to the version-specific spin-up acceleration procedure.

While we envision our approach to play an important role in promoting high spatial resolution TBMs, how to adapt our approach to future TBMs (or other ESM components) possibly with complex horizontal dynamics (e.g. water transfers among neighbouring pixels and soil erosion) needs further improvements. Advanced ML methods (such as recursive or graphic neural networks) might provide a solution, which needs future investigation.

### 3.6.2 | Towards the improved assimilation of observations into TBMs

TBMs have been criticized regarding their reliability, primarily because of a large number of unconstrained model parameters. The applications of model-data fusion (or data assimilation) for TBMs are hampered by the computational bottleneck, with measurements of carbon stocks seldom being systematically assimilated to improve parameterization (MacBean et al., 2022). Instead, parameters that control short-term processes are optimized assuming that long-term processes have limited impacts on short-term dynamics. While this assumption, imposed by the computational cost, facilitates specific applications, it is inadequate to solve the parameterization challenge for TBMs with multiple interactive processes with various timescales in their evolution. The elimination of the computational bottleneck through MLA is an essential step toward systematic model-data fusion.



**FIGURE 8** Changes in total yearly net biome productivity (NBP) during 1901–2016 from and *production-run<sub>conv</sub>* (colored) and *production-run<sub>MLA</sub>* (black) type simulations for three different versions of ORCHIDEE family. Shown are averages over the following spatial domains: (a) global, (b) North Hemisphere (30°N–90°N), (c) Tropics (30°S–30°N) and (d) South Hemisphere (30°S–90°S).



### 3.6.3 | Towards refined realism of TBMs

TBMs have been criticized for a lack of ecological processes or detail which potentially lead to biases in the predicted response of the biosphere to global change drivers (Fisher et al., 2014; Fisher & Koven, 2020; Prentice et al., 2015). Different types of dedicated models were developed to incorporate fundamental ecological processes from ecologists (e.g. Abs et al., 2020). Despite being well tested at site level, their global implementations were much delayed or assessed as impossible because the computational framework is inadequate to efficiently spin-up highly nonlinear models at a high spatial resolution for global applications. Our MLA approach overcomes this big obstacle, and will facilitate the integration of a diversity of ecological processes for global applications.

### 3.6.4 | Towards machine learning enabled TBMs

Here, we creatively merge ML and a TBM for global scale studies to tackle the primary computational bottleneck. The rapid expansion of heterogeneous big datasets and growing complexity of TBMs requires new methodologies to bring TBMs to the next level that adequately accommodates multiple theoretical breakthroughs and novel high dimensional real-world information. Different ML methodologies have evolved to generate inputs for model simulations, optimize model parameters, emulate model behaviors, or substitute model components (Reichstein et al., 2019). The applications of ML for large scale TBMs are typically centered on data (inputs, evaluation, and benchmarking) generation, despite the urgent needs of novel approaches in tackling multiple challenges for ESMs through integrating breakthroughs in ML. Our approach serves as a case study to inspire the machine learning enabled next generation TBMs.

#### AUTHOR CONTRIBUTIONS

The concept was developed by Yan Sun, Daniel S. Goll and Philippe Ciais. The machine learning algorithm was built by Yan Sun based on TBM data produced by Daniel S. Goll. The analysis was performed by Yan Sun and Daniel S. Goll. Yilong Wang and Vladislav Bastrikov provided technical support with the coding. Yan Sun, Yuanyuan Huang, and Daniel S. Goll wrote the manuscript with inputs from all authors.

#### ACKNOWLEDGMENTS

This work was granted access to the HPC resources of GENCI-TGCC under the allocation A0130106328. YS is supported by National Natural Science Foundation of China (NSFC; project number: 42201107). DSG and PC benefited from the French state aid managed by the ANR under the "Investissements d'avenir" programme with the reference ANR-16-CONV-0003.

#### CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

#### DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available at Mendeley Data <https://data.mendeley.com/datasets/8p5xyhv8j9/1>. The raw data from the TBM simulation are archived at TGCC, and is available upon request. The source code of the land surface models versions are openly available in IPLS data catalog at <https://doi.org/10.14768/20200407002.1>, <https://doi.org/10.14768/391825ae-d257-4365-9820-30ea1940914c>, [http://forge.ipsl.jussieu.fr/orchidee/wiki/GroupActivities/CodeAvailabilityPublication/ORCHIDEE\\_2.2\\_gmd\\_2022](http://forge.ipsl.jussieu.fr/orchidee/wiki/GroupActivities/CodeAvailabilityPublication/ORCHIDEE_2.2_gmd_2022). The implementation of the approach in python for ORCHIDEE TBMs is openly available in ZENODO at <https://doi.org/10.5281/zenodo.7503092>.

#### ORCID

Yan Sun  <https://orcid.org/0000-0003-0481-7192>

Daniel S. Goll  <https://orcid.org/0000-0001-9246-9671>

Yuanyuan Huang  <https://orcid.org/0000-0003-4202-8071>

Ying-Ping Wang  <https://orcid.org/0000-0002-4614-6203>

Yilong Wang  <https://orcid.org/0000-0001-7176-2692>

#### REFERENCES

- Abs, E., Leman, H., & Ferrière, R. (2020). A multi-scale eco-evolutionary model of cooperation reveals how microbial adaptation influences soil decomposition. *Communications Biology*, 3(1), 1–13.
- Belitz, K., & Stackelberg, P. E. (2021). Evaluation of six methods for correcting bias in estimates from ensemble tree machine learning regression models. *Environmental Modelling & Software*, 139, 105006.
- Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2), 123–140.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357.
- Ellsworth, D. S., Crous, K. Y., De Kauwe, M. G., Verryckt, L. T., Goll, D., Zaehle, S., Bloomfield, K. J., Ciais, P., Cernusak, L. A., Domingues, T. F., Dusenge, M. E., Garcia, S., Guerrieri, R., Ishida, F. Y., Janssens, I. A., Kenzo, T., Ichie, T., Medlyn, B. E., Meir, P., ... Wright, I. J. (2022). Convergence in phosphorus constraints to photosynthesis in forests around the world. *Nature Communications*, 13(1), 1–12.
- Farquard, M. A. H., & Bose, I. (2012). Preprocessing unbalanced data using support vector machine. *Decision Support Systems*, 53(1), 226–233.
- Fisher, J. B., Huntzinger, D. N., Schwalm, C. R., & Sitch, S. (2014). Modeling the terrestrial biosphere. *Annual Review of Environment and Resources*, 39(1), 91–123.
- Fisher, R. A., & Koven, C. D. (2020). Perspectives on the future of land surface models and the challenges of representing complex terrestrial systems. *Journal of Advances in Modeling Earth Systems*, 12(4), e2018MS001453.
- Friedlingstein, P., Jones, M. W., O'Sullivan, M., Andrew, R. M., Bakker, D. C., Hauck, J., Le Quéré, C., Peters, G. P., Peters, W., Pongratz, J., Sitch, S., Canadell, J. G., Ciais, P., Jackson, R. B., Alin, S. R., Anthoni, P., Bates, N. R., Becker, M., Bellouin, N., ... Zeng, J. (2022). Global carbon budget 2021. *Earth System Science Data*, 14(4), 1917–2005.
- Goll, D., Vuichard, N., Maignan, F., Jornet-Puig, A., Sardans, J., Violette, A., Peng, S., Sun, Y., Kvakic, M., Guimberteau, M., & Guenet, B. (2017). A representation of the phosphorus cycle for ORCHIDEE (revision 4520). *Geoscientific Model Development Discussions*, 10(10), 3745–3770. <https://doi.org/10.5194/gmd-10-3745-2017>



- Goll, D. S., Bauters, M., Zhang, H., Ciais, P., Balkanski, Y., Wang, R., & Verbeeck, H. (2022). Atmospheric phosphorus deposition amplifies carbon sinks in simulations of a tropical forest in Central Africa. *New Phytologist*. <https://doi.org/10.1111/nph.18535>
- Huang, Y., Zhu, D., Ciais, P., Guenet, B., Huang, Y., Goll, D. S., Guimberteau, M., Jornet-Puig, A., Xingjie, L., & Luo, Y. (2018). Matrix-based sensitivity assessment of soil organic carbon storage: A case study from the ORCHIDEE-MICT model. *Journal of Advances in Modeling Earth Systems*, 10(8), 1790–1808.
- Krinner, G., Viovy, N., de Noblet-Ducoudré, N., Ogée, J., Polcher, J., Friedlingstein, P., Ciais, P., Sitch, S., & Prentice, I. C. (2005). A dynamic global vegetation model for studies of the coupled atmosphere-biosphere system. *Global Biogeochemical Cycles*, 19(1). <https://doi.org/10.1029/2003GB002199>
- MacBean, N., Bacour, C., Raoult, N., Bastrikov, V., Koffi, E. N., Kuppel, S., Maignan, F., Ottlé, C., Peaucelle, M., Santaren, D., & Peylin, P. (2022). Quantifying and reducing uncertainty in global carbon cycle predictions: Lessons and perspectives from 15 years of data assimilation studies with the ORCHIDEE terrestrial biosphere model. *Global Biogeochemical Cycles*, 36, e2021GB007177. <https://doi.org/10.1029/2021GB007177>
- Prentice, I. C., Liang, X., Medlyn, B. E., & Wang, Y. P. (2015). Reliable, robust and realistic: The three R's of next-generation land-surface modelling. *Atmospheric Chemistry and Physics*, 15(10), 5987–6005.
- Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., & Carvalhais, N. (2019). Deep learning and process understanding for data-driven earth system science. *Nature*, 566(7743), 195–204. <https://doi.org/10.1038/s41586-019-0912-1>
- Sun, Y., Goll, D. S., Chang, J., Ciais, P., Guenet, B., Helfenstein, J., Huang, Y., Lauerwald, R., Maignan, F., Naipal, V., Wang, Y., Yang, H., & Zhang, H. (2021). Global evaluation of the nutrient-enabled version of the land surface model ORCHIDEE-CNP v1. 2 (r5986). *Geoscientific Model Development*, 14(4), 1987–2010.
- Thornton, P. E., & Rosenbloom, N. A. (2005). Ecosystem model spin-up: Estimating steady state conditions in a coupled terrestrial carbon and nitrogen cycle model. *Ecological Modelling*, 189(1–2), 25–48. <https://doi.org/10.1016/j.ecolmodel.2005.04.008>
- Tian, H., Yang, J., Lu, C., Xu, R., Canadell, J. G., Jackson, R., Arneeth, A., Chang, J., Chen, G., Ciais, P., & Gerber, S. (2018). The global N<sub>2</sub>O model Intercomparison project (NMIP): Objectives, simulation protocol and expected products. *Bulletin of the American Meteorological Society*, 99(6), 1–51. <https://doi.org/10.1175/BAMS-D-17-0212.1>
- Wang, Y. P., & Goll, D. S. (2021). Modelling of land nutrient cycles: Recent progress and future development. *Faculty Reviews*, 10. <https://doi.org/10.12703/r/10-53>
- Xia, J. Y., Luo, Y. Q., Wang, Y. P., Weng, E. S., & Hararuk, O. (2012). A semi-analytical solution to accelerate spin-up of a coupled carbon and nitrogen land model to steady state. *Geoscientific Model Development*, 5(5), 1259–1271. <https://doi.org/10.5194/gmd-5-1259-2012>

## 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:** Sun, Y., Goll, D. S., Huang, Y., Ciais, P., Wang, Y.-P., Bastrikov, V., & Wang, Y. (2023). Machine learning for accelerating process-based computation of land biogeochemical cycles. *Global Change Biology*, 29, 3221–3234. <https://doi.org/10.1111/gcb.16623>