



THE UNIVERSITY *of* EDINBURGH

Edinburgh Research Explorer

Evaluating two land surface models for Brazil using a full carbon cycle benchmark with uncertainties

Citation for published version:

Caen, A, Smallman, TL, Anderson de Castro, A, Robertson, E, von Randow, C, Cardoso, M & Williams, M 2021, 'Evaluating two land surface models for Brazil using a full carbon cycle benchmark with uncertainties', *Climate Resilience and Sustainability*. <https://doi.org/10.1002/cli2.10>

Digital Object Identifier (DOI):

[10.1002/cli2.10](https://doi.org/10.1002/cli2.10)

Link:

[Link to publication record in Edinburgh Research Explorer](#)

Document Version:

Publisher's PDF, also known as Version of record

Published In:

Climate Resilience and Sustainability

Publisher Rights Statement:

© 2021 The Authors. Climate Resilience and Sustainability published by John Wiley & Sons Ltd on behalf of Royal Meteorological Society

General rights


Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy

The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.



Evaluating two land surface models for Brazil using a full carbon cycle benchmark with uncertainties

Auguste Caen¹ | T. Luke Smallman¹ | Aline Anderson de Castro² |
Eddy Robertson³ | Celso von Randow² | Manoel Cardoso² |
Mathew Williams¹ 

¹ School of GeoSciences and NCEO,
University of Edinburgh, Edinburgh, UK

² Impacts, Adaptation and Vulnerability
Division, INPE, São José dos Campos, SP,
Brazil

³ Met Office Hadley Centre, Exeter, UK

Correspondence

Mathew Williams, School of GeoSciences
and NCEO, University of Edinburgh,
Edinburgh, EH9 3FF, UK.

Email: mat.williams@ed.ac.uk

Funding information

UK Met Office, Grant/Award Number:
CSSP Brazil; Royal Society, Grant/Award
Number: Wolfson Award; Fundação de
Amparo à Pesquisa do Estado de São
Paulo, Grant/Award Number: 501220;
UKSA, Grant/Award Number: Forests
2020; NCEO; Newton Fund, Grant/Award
Number: CSSP Brazil

Abstract

Forecasts of tropical ecosystem C cycling diverge among models due to differences in simulation of internal processes such as turnover, or transit times, of carbon pools. Estimates of these processes for the recent past are needed to test model representations, and so build confidence in model forecasts within and across biomes. Here, we evaluate carbon cycle process representation in two land surface models [Joint UK Land Environment Simulator (JULES) and Integrated Model of Land Surface Processes (INLAND)] for the period 2001–10 across Brazilian biomes. Model outputs are evaluated using the ILAMB system. Probabilistic benchmarking data were created using the carbon data model framework that assimilates observational times series of leaf area index and maps of woody biomass and soil C. New custom uncertainty metrics assess if models are within benchmark uncertainties. Simulations are better in homogeneous areas of vegetation type, and are less robust at ecotones between biomes, likely due to disturbance effects and parameter errors. Gross biosphere-atmosphere fluxes are robustly modelled across Brazil. However, benchmark uncertainty is too high on net ecosystem exchange to provide an accurate evaluation of the models. The LSMs have significant differences in internal carbon allocation and the dynamics of the different C pools. JULES models dead C stocks more accurately while living C stocks are best resolved for INLAND. JULES' over-estimate of the C wood pool results from over-estimation of both inputs to wood and the transit time of wood. INLAND's under-estimate of dead C stocks arises from an under-estimate of the transit time of dead organic matter. The models are better at simulating annual averages than seasonal variation of fluxes. Analyses of monthly net C exchanges show that INLAND correctly simulates seasonality, but over-estimates amplitudes, whereas JULES correctly simulates the annual amplitudes, but is out of phase with the benchmark.

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2021 The Authors. *Climate Resilience and Sustainability* published by John Wiley & Sons Ltd on behalf of Royal Meteorological Society

KEYWORDS

Amazon, biomass, carbon cycle, Cerrado, LAI, land surface model, mean transit time, model-data fusion, net ecosystem exchange

1 | INTRODUCTION

Tropical ecosystems, which cover 90% of Brazil, are major stores of carbon (C) and drivers of C exchanges between land and atmosphere. The response of these tropical ecosystems to increasing atmospheric CO₂ concentrations and climate changes will have important feedbacks on the Earth system. Understanding and predicting the future of tropical C cycling is thus critical for managing efforts to meet international political agreements, such as the Paris Agreement related to the United Nations Framework Convention on Climate Change. Land surface models (LSMs) are used to predict future C cycling based on theorized carbon cycle climate feedbacks (Bonan, 2018). However, these models are complex, with multiple parameterized representations of interacting processes. Model inter-comparison studies show important and poorly understood differences in model forecasts (Jones *et al.*, 2016). Much of the model-to-model variation in projected land sink during the 21st century is linked to biases existing during the observational era. The primary driver of this variability has been linked to model differences in the representation of slowly changing carbon cycle processes linked to transit times of C through the ecosystem (Hoffman *et al.*, 2014).

LSMs, such as Joint UK Land Environment Simulator (JULES) and Integrated Model of Land Surface Processes (INLAND), have been developed to determine global terrestrial C cycling responses to past and future climate change (Koven *et al.*, 2011; Sitch *et al.*, 2003; Woodward *et al.*, 1995). LSMs represent and link vegetation processes (i.e. growth, turnover and competition) and biogeochemistry (i.e. water and nutrient cycling, soil decomposition). LSMs can characterize C dynamics in response to varied forcing, such as weather or human disturbance, and therefore forecast C cycle responses to future environmental conditions. However, C cycle modelling in LSMs typically relies on parameters derived from literature. Parameters are linked to a limited number of prescribed plant functional types (PFTs) which describe process variability between biomes. Also, LSMs tend to use a spin-up process to ensure that the large C pools [biomass and dead organic matter (DOM)] reach steady state. Further, inherent differences of LSM structure contribute strongly to forecast uncertainties (Nishina *et al.*, 2014; Exbrayat *et al.*, 2018), more than do differences in climate projections (Ahlström

et al., 2012). Many model inter-comparison projects have demonstrated a lack of coherence in future projections of terrestrial C cycling (Ahlström *et al.*, 2012; Friedlingstein *et al.*, 2014). Recent studies have used simulations from the first phase of the Inter-Sectoral Impact Model Inter-comparison Project (ISI-MIP) (Warszawski *et al.*, 2014) to evaluate the importance of key elements regulating vegetation C dynamics, but also the estimated magnitude of their associated uncertainties (Friend *et al.*, 2014; Nishina *et al.*, 2015; Thurner *et al.*, 2017; Nishina *et al.*, 2014; Exbrayat *et al.*, 2018).

An important insight is that transit times (or residence times) of C in LSMs are a key uncertain feature of the global C cycle simulation. Further, uncertainties in LSM estimates of stocks and fluxes are unknown, which weakens their analytical value. LSM inter-comparison has highlighted the need to evaluate terrestrial C cycle process representation against independent data. The goal of such an evaluation should be to highlight particularly the validity of cycling of large slow pools such as wood and DOM C, changes to which ultimately determine C sources and sinks. Poorly understood processes include phenological variability, allocation of photosynthate to wood and residence time of wood. Rigorously assessing models during an observation period provides a way to advance understanding and predictability of terrestrial biogeochemical processes, inform model development and identify relevant measurements from field campaigns and satellites for further improved testing.

Over the past decade, data from field networks and satellite observations have improved understanding of global terrestrial C stocks and phenology at finer resolutions. These products range from machine-learning-based upscaling of FLUXNET data (Jung *et al.*, 2017), remotely sensed biomass products (Thurner *et al.*, 2014; Carvalhais *et al.*, 2014) and the creation of global soil databases (FAO/IIASA/ISRIC/ISSCAS/JRC, 2012) (Hengl *et al.*, 2017). Due to a reliance on interpolation and upscaling with other spatial data, it is challenging to evaluate these products for inherent biases and so to link these for a consistent model evaluation. Also, spatial reporting of key internal variables such as allocation of C to wood, or wood residence time has been sparse and lacking clear uncertainty estimates. It is important that these data sets have robust error assessments so that they can be appropriately weighted in model evaluation.

To produce a benchmark of key C cycle variables, existing spatial observational products can be combined into a consistent, error characterized, description of the C cycle through assimilation with an intermediate complexity (IC) model. Such model-data fusion draws information from varied data sources, taking account of the supplied errors, to calibrate an IC model using Bayesian approaches (Williams *et al.*, 2005; Luo *et al.*, 2009; Fox *et al.*, 2009; Bloom and Williams, 2015). The outcome is a probabilistic assessment of the full C cycle at the resolution of the analysis, from C uptake to allocation, transit times and respiration. The C budget is provided by the IC model with parameters adjusted at the analysis resolution to be consistent with the multiple data sources assimilated for that location, weighted by their errors. The model output is generated by local climate forcing and any other endogenous forcing such as burning or land-use change products. Using Bayesian approaches means that error is propagated throughout the data assimilation, to produce ensembles of model parameters, stocks and fluxes. So, fluxes such as allocation to wood are estimated with errors that are linked to related observational products, such as wood biomass maps. The analysis output provides a means to test processes and dynamics of LSMs, and particularly to investigate those model components implicated in divergent predictions that differ significantly from the benchmark.

Here, we use the carbon data model framework (CARDAMOM) (Bloom and Williams, 2015; Bloom *et al.*, 2016; Smallman *et al.*) to benchmark historical modelling by INLAND and JULES of the Brazilian terrestrial carbon cycle at 1° resolution for the 2001–2010 period. CARDAMOM assimilates gridded observations of leaf area index (LAI) times series, and maps of woody biomass and soil organic carbon (SOC) stocks at these spatial scales into DALEC, an IC C model (i.e. less complex than JULES and INLAND). CARDAMOM finds distributions of parameters and initial conditions for DALEC that are consistent with local atmospheric forcing, observations and a series of ecological and dynamical constraints (EDCs). These EDCs ensure that common sense rules are applied to restrict Monte Carlo searching of the parameter hypervolume of DALEC to realistic regions (i.e. sensible values for root:shoot ratios and relative lifespans for different plant tissues; quasi-steady-state behaviour for C pools). CARDAMOM therefore avoids the use of PTF concepts – instead CARDAMOM produces a continuum of spatially varying parameters, at 1 degree resolution across Brazilian biomes. CARDAMOM avoids steady-state assumptions, instead allowing a large ensemble of parameters for each 1 degree pixel that span a range of quasi-steady-states consistent with observations and forcing.

This paper targets a series of questions to diagnose the models' capabilities to reproduce the benchmark's outputs

in space and time across Brazilian biomes. The analysis focuses on key biogenic stocks and fluxes to develop understanding of ecological process variation in time and space across Brazil. Key questions addressed are: (i) For both LSMs which biomes have the greatest and least consistency with the benchmark? (ii) How do the models compare with the benchmark in terms of internal C processing and the major biosphere-atmosphere C fluxes? (iii) What is the mean transit time for C in each model and how robust are these estimates against the benchmark? (iv) How reliable are the representations of seasonality in the models? The main challenge for this study lies in the complexity of models, which include multiple output rates and pools of C, and the variability of the Brazilian landscape. Within Brazil are found diverse ecosystems, including a large part of the Amazon biome, and the varied seasonal tropical biomes of Cerrado, Caatinga and the Atlantic Forest. Quantitative measures compare the model variables with the benchmark, including scores based on spatial or temporal matching, with a breakdown of results for each biome. The comparison highlights the general matching between each model and the benchmark, and identifies which biomes and which processes have the highest and lowest consistencies. To identify the causes of disagreement between models and the benchmark, a further analysis compares the models in terms of inputs, outputs and internal processing, and isolates the main variables on which total C dynamics are dependent. Knowing these dependent variables provides qualitative targeting for future modelling efforts, with a focus on the transit time of the main pools contributing to carbon storage. Finally, this aggregated analysis helps identify and understand source and sink behaviours of the system, including its variability over seasonal cycles. The study includes metrics for model evaluation against a benchmark that incorporates confidence intervals (CIs). As far as we know, this is the first published evaluation of this kind.

2 | METHODS

The evaluation of the two LSMs, JULES and INLAND, across the major Brazilian biomes (Figure 1) was facilitated by the ILAMB (International Land Model Benchmarking (<https://www.ilamb.org/>, (Collier *et al.*, 2018)) evaluation package. ILAMB uses metrics to inter-compare land models and make evaluations against benchmarks. Here, the benchmarking dataset was generated by model-data fusion using the CARDAMOM approach, which describes a full C cycle and its error characteristics (Figure 2). To exploit the uncertainty estimates on the benchmark analysis, we updated ILAMB to test whether model variables lie within

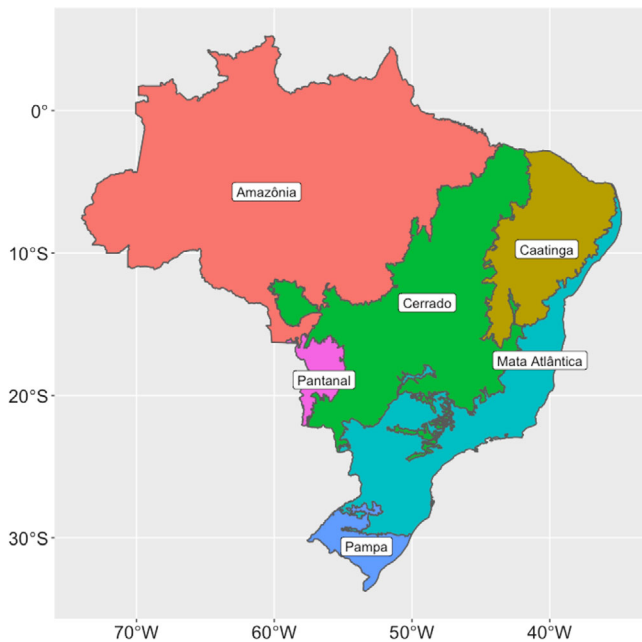


FIGURE 1 The Brazilian biomes used in the benchmarking. The Amazon (Amazônia) and Atlantic forest (Mata Atlântica) correspond to moist tropical forests. The Pantanal biome corresponds to wetlands, and the Cerrado and Caatinga biomes correspond to open woodland and grass savanna vegetation. Map source: Brazilian Institute of Geography and Statistics (IBGE), Biomes and Coastal-Marine System of Brazil map, <https://www.ibge.gov.br/>, accessed 17/11/2020

the benchmark CIs, so that models are not penalized for bias if the benchmark is uncertain.

The analysis focused on the period 2001–2010, when outputs from JULES, INLAND and CARDAMOM were all available across Brazil. Comparison was undertaken at 1° spatial resolution and monthly or annual temporal resolution, depending on the variable analysed. The TRENDY protocol (Sitch et al., 2015) was used to ensure a common climate forcing for models and CARDAMOM reanalyses. Fire was imposed in CARDAMOM using MODIS burned area, whereas the LSMs did not include explicit fire modelling. Forest disturbance in CARDAMOM was imposed using GFW (Hansen et al., 2013), whereas the LSMs used the LUH2 database (Hurtt et al., 2020, 2019a, 2019b). However, the focus of this paper is on evaluating the ecophysiological processes and fluxes, rather than fluxes connected to fire and disturbance.

2.1 | ILAMB – A benchmarking system

The standard metrics for ILAMB are described in Online Appendix C, and summarized in Table 1. Here, new metrics for evaluation of uncertainty and bias are described.

The calibration of the benchmark provides an estimate of statistical uncertainty in terms of a CI with high and low percentiles. ILAMB determines whether a model output in any grid cell (v_{mod}) lies within a given CI of the benchmark (v_{ref}), a pass/fail test:

$$\begin{aligned} \epsilon_{CI}(x) &= 1 \text{ if } v_{ref}^{2.5pc}(x) \leq v_{mod}(x) \leq v_{ref}^{97.5pc} \\ &= 0 \text{ else.} \end{aligned} \quad (1)$$

We define the 95% CI uncertainty score for a biome or region (s_{CI}) by accumulating the tests for all grid cells within the biome to generate a score which varies between 0 and 1:

$$s_{CI} = \frac{1}{area} \int_{x \in area} \epsilon_{CI}(x). \quad (2)$$

A value of 1 means 100% of the model spatial points within the defined area (e.g. biome) fit within the CI of the benchmark, whereas 0 indicates that none of the values of the model are within the CI of the benchmark anywhere within the area or region. As far as we know, this is the first evaluation system capable of comparing model outputs to benchmarks with uncertainties.

The standard suite of ILAMB model-benchmark comparison metrics allows the evaluation of different aspects of the system. For an overall temporal comparison, root mean square error (RMSE) provides the most robust metric as it evaluates misfit for each time step. For the spatial comparison (i.e. averaged over time), the bias scores are more informative, because these scores identify the distance between the model values and the benchmark values averaged over time. Indeed, contrary to the RMSE score, the bias score is not impacted by the time step of the data. However, the bias score does not indicate if the model results are over- or under-estimated given the benchmark values. Thus, we created a modified bias score, called *signed bias score*, to estimate this over- or under-estimation, given its sign:

$$\tilde{s}_{bias}(x) = sgn_{bias}(x) - sgn_{bias}(x)e^{-|bias(x)|}$$

with

$$\begin{aligned} sgn_{bias}(x) &= 1 \quad \text{if } bias(x) > 0 \\ &= -1 \quad \text{if } bias(x) < 0 \\ &= 0 \quad \text{if } bias(x) = 0. \end{aligned} \quad (3)$$

That is, this signed bias score is negative if the studied variable is under-estimated by the model relative to the benchmark, and positive if it over-estimated. Finally, the bias

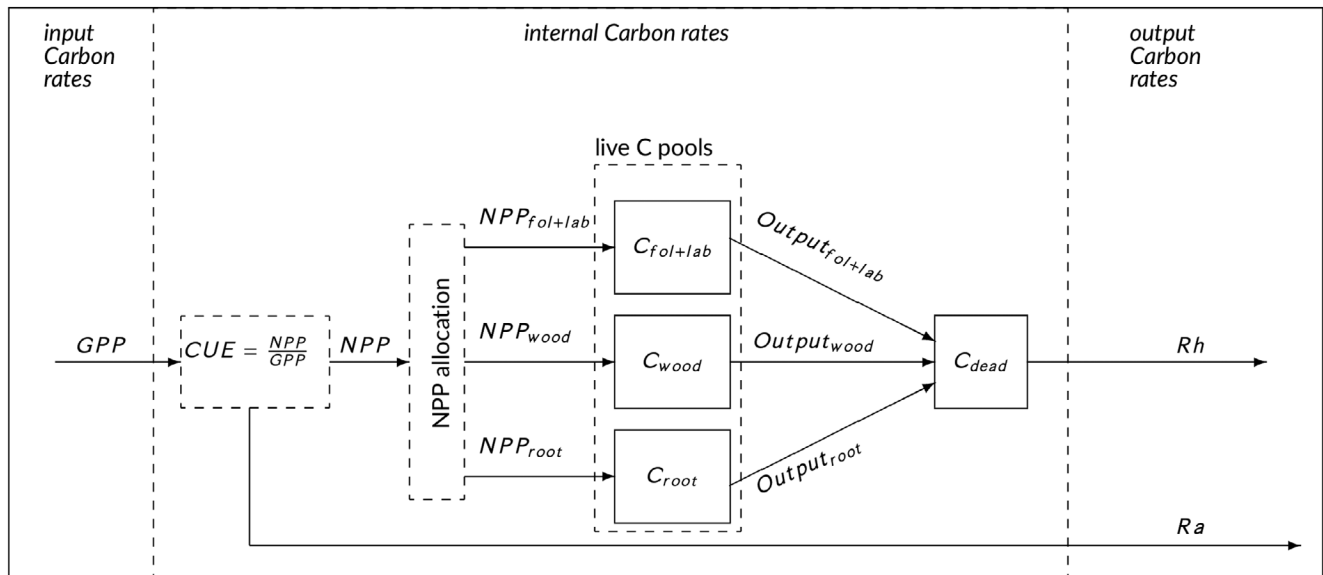


FIGURE 2 The C cycle system as represented in the DALEC model structure, and used in the benchmarking process. The system consists of one incoming rate (photosynthesis, GPP) and two outgoing rates (R_a and R_h). R is respiration (a is autotrophic and h is heterotrophic). NPP is net primary production. Within the system, there are three live carbon pools ($C_{fol+lab}$, C_{wood} and C_{root}), and one dead carbon pool (which corresponds to the sum of C_{lit} and C_{soil} of the DALEC model), whose outgoing and incoming flows correspond to the internal carbon flows of the system

TABLE 1 Summary of the different ILAMB metrics, including the behaviour each metric is designed to evaluate, an explanation of the scoring, and the time steps on which metrics are generated

Scores	Measured behaviour	Values	Compatible time steps
Bias score	Temporal average	1 \Leftrightarrow same temporal average 0 \Leftrightarrow mismatching	Month Annual
Signed bias score	Temporal average	1 \Leftrightarrow over-estimate 0 \Leftrightarrow same temporal average -1 \Leftrightarrow under-estimate	Month Annual
CI score	% of spatial points within the benchmark CI	1 \Leftrightarrow 100% of matching points 0 \Leftrightarrow 0% of matching points	Month Annual
Root mean square error	Temporal and spatial matching	1 \Leftrightarrow same temporal and spatial values 0 \Leftrightarrow mismatching	Month Annual
Seasonal cycle score	Seasonal matching	1 \Leftrightarrow maximum reached in the same month 0 \Leftrightarrow opposite seasonality	Month
Spatial distribution score	Spatial standard deviation	1 \Leftrightarrow same spatial distribution 0 \Leftrightarrow mismatching	Month Annual
Inter-annual variability score	Spatial and inter-annual matching	1 \Leftrightarrow same inter-annual dynamics 0 \Leftrightarrow mismatching	Month

score can be directly derived from the signed bias score, from the following formula:

$$s_{bias}(x) = 1 - |\tilde{s}_{bias}(x)|. \quad (4)$$

2.2 | The land surface models

2.2.1 | JULES

JULES is a UK-based community LSM (Clark et al., 2012) and here we assess the JULES earth system (ES) configuration, which is used for carbon cycle science and as part of the UKESM1.0 earth system model. JULES-ES includes nine natural PFTs and four grass-like agricultural PFTs, the configuration is designed for global simulations and in practice only the tropical broadleaf tree, C3 and C4 grasses have substantial cover in Brazil. The distribution of PFTs is predicted by JULES, using a height-based competition equation, which for Brazil means that grasses can only grow where trees are not viable or where trees are actively prevented from growing by the presence of agriculture. Vegetation carbon is allocated between three plant components: leaf, wood and fine root pools, and each of these pools has a fixed turnover rate and associated litter flux. There are also litter fluxes caused by large-scale disturbance (a constant mortality term representing processes not explicitly modelled, for example, fire, wind-throw, disease, etc.), inter-PFT competition and land-use change. There are four soil carbon pools and these are summed to represent total DOM C.

Following the TRENDY protocol (Sitch et al., 2015), JULES is spun-up to a near steady state under pre-industrial forcing, after which 300 years of transient simulation are performed before the start of the period of study. Fire is not directly simulated in this implementation. During the period of study, JULES will not necessarily be close to a steady state having been perturbed by changes in climate, land-use, nitrogen deposition and atmospheric CO₂ concentration. In particular, changes in land-use can instantaneously push the model far from steady state. Tree cover and soil carbon can take hundreds of years to reach a new equilibrium.

2.2.2 | INLAND

INLAND is the Brazilian Dynamic Global Vegetation Model and represents water, energy and carbon fluxes, together with the vegetation dynamics and carbon stocks. The model builds on earlier work (Foley et al., 1996; Kucharik et al., 2000) and represents 12 natural PFTs, including eight upper and four lower canopy PFTs, which includes shrubs (evergreen and deciduous) and grasses

(C3 and C4). The PFT is defined according to the climate restrictions. Competition between species is determined by access to water and light, with shorter stature PFTs more easily accessing water, but being shaded by the higher vegetation (Kucharik et al., 2000). The carbon stocks in the vegetation are calculated as a function of *NPP* and allocation coefficients and residence times for each component, and also considering losses due to land-use changes and fire disturbance. The C transit time (MTT) and allocation coefficients are user-defined model parameters, which can be different for each plant component (leaf, wood and root) and PFT. The dead organic carbon is allocated in four pools with different residence times, from hours in the microbial to thousands of years in the stabilized organic matter. Temperature and soil water content controls the microbial activity and soil texture controls its growth. INLAND has a 400 year spin-up to the pre-industrial state and the follows the TRENDY protocol to produce data for analysis here.

2.3 | CARDAMOM as a C cycle benchmark

CARDAMOM, a model-data fusion framework, provides an error-characterized, complete analysis of the C cycle (Figure 2). CARDAMOM includes a C cycle model, DALEC; this has the advantage of evaluating the observational data for internal consistency (e.g. with mass balance), propagating error across the C cycle and generating internal model variables such as transit time. CARDAMOM/DALEC is independent of the benchmarked models, INLAND and JULES. A full description of the CARDAMOM process and the DALEC model are provided in Online Appendix B. CARDAMOM meteorological drivers are extracted from the CRU-JRAv1.1 dataset, a 6-hourly 0.5 × 0.5 degree reanalysis (Harris, 2019). Fire is imposed using the MODIS burned area product (500 × 500 m) (Giglio et al., 2018). Forest biomass loss is imposed using the global forest watch (GFW) database (30 × 30 m) (Hansen et al., 2013). Time series information on *LAI* is drawn from 1 × 1 km satellite based Earth Observation (EO) estimates (Copernicus Service Information 2020); this directly corresponds to the DALEC *LAI* state variable, which is proportional to its foliar C pool. Prior information on soil carbon stocks is drawn from the SoilGrids database (Hengl et al., 2017), a data-driven interpolation of field inventories (250 × 250 m), and we assume that this corresponds to the DALEC SOM pool. Woody biomass information is drawn from two data sources: (i) across the Brazilian Amazon a map representative of 2014 generated from airborne lidar (50 × 50 m) (Longo et al., 2016) and (ii) for all other areas a 1 × 1 km resolution (Avitabile et al., 2016) map assumed to be nominally

TABLE 2 Studied variables and parameters

Symbol	Definition	Units	Step time
Fluxes			
<i>GPP</i>	Gross primary production	$\text{g.m}^{-2}.\text{d}^{-1}$	month
<i>NPP</i>	Net primary production	$\text{g.m}^{-2}.\text{d}^{-1}$	month
<i>NPP</i> _{leaf+labile}	Net primary production for foliage and labile pools	$\text{g.m}^{-2}.\text{d}^{-1}$	year
<i>NPP</i> _{wood}	Net primary production for wood	$\text{g.m}^{-2}.\text{d}^{-1}$	year
<i>NPP</i> _{root}	Net primary production for fine root	$\text{g.m}^{-2}.\text{d}^{-1}$	year
<i>Output</i> _{leaf+labile}	C losses from foliage and labile pools	$\text{g.m}^{-2}.\text{d}^{-1}$	year
<i>Output</i> _{wood}	C losses from wood	$\text{g.m}^{-2}.\text{d}^{-1}$	year
<i>Output</i> _{root}	C losses from fine roots	$\text{g.m}^{-2}.\text{d}^{-1}$	year
<i>Rh</i>	Heterotrophic respiration	$\text{g.m}^{-2}.\text{d}^{-1}$	month
<i>Ra</i>	Autotrophic respiration	$\text{g.m}^{-2}.\text{d}^{-1}$	month
<i>RECO</i>	Ecosystem respiration	$\text{g.m}^{-2}.\text{d}^{-1}$	month
<i>NEE</i>	Net ecosystem exchange	$\text{g.m}^{-2}.\text{d}^{-1}$	month
C pools			
<i>C</i> _{leaf+labile}	Carbon biomass in foliage and labile	g.m^{-2}	year
<i>C</i> _{wood}	Carbon biomass in wood	g.m^{-2}	year
<i>C</i> _{root}	Carbon biomass in fine roots	g.m^{-2}	year
<i>C</i> _{dead}	Dead carbon mass	g.m^{-2}	year
Parameters			
<i>MTT</i> _{leaf}	Mean transit time in foliar C	year	year
<i>MTT</i> _{wood}	Mean transit time in wood C	year	year
<i>MTT</i> _{root}	Mean transit time in fine root C	year	year
<i>MTT</i> _{dead}	Mean transit time in dead C	year	year
<i>CUE</i>	Carbon use efficiency: $CUE = \frac{NPP}{GPP}$	dimensionless	year

representative of 2007. Both maps provide estimates of aboveground biomass from which is derived total woody biomass, using an allometric relationship (Saatchi et al., 2011), and this is assumed to directly correspond to the DALEC woody C pool. Each data source was aggregated to 1×1 degree spatial resolution. *LAI* and biomass estimates were provided with uncertainty estimates which were used as part of the assimilation process. The SoilGrids database currently lacks an associated estimate of uncertainty and its time period is also poorly defined due to the variation in sampling time of the component studies. For simplicity, we assumed the uncertainty as represented by the standard deviation of the aggregation process.

The quality of DALEC-CARDAMOM outputs over Brazil has been evaluated elsewhere against independent data (Smallman et al., 2021). DALEC-simulated *NEE* was statistically consistent with the CarbonTracker-Europe ensemble (van der Laan-Luijkx et al., 2015) at the 90% CI across >99% of Brazil (2009–2017). The DALEC models are consistent with FLUXCOM *GPP* () at the 90% CI across 94% of Brazil. The DALEC outputs match the calibration information with a high degree of skill: the RMSE is small for *LAI* and the initial soil carbon stock (<5%). RMSE between simulated wood stocks and calibration observations was

16% and is dominated by model-observation mismatch at smaller wood stocks (<50 MgC/ha; 20–28%) with smaller errors (<1%) otherwise.

JULES and INLAND C cycling outputs were matched to the DALEC model structure (Figure 2), including production (*GPP*, *NPP* and their ratio *CUE*), allocation of *NPP* to foliage, wood and fine roots: the biomass in these plant tissues; their annual losses; the C mass in DOM (the sum of all dead C pools); and both autotrophic and heterotrophic respiration, and the mean transit time (*MTT*) of wood and DOM pools. Each C pool has dynamics corresponding to a C mass balance:

$$\frac{dC_{pool}}{dt} = Input_{pool} - \frac{C_{pool}}{MTT_{pool}},$$

$$MTT_{pool} = \frac{C_{pool}}{output_{pool}}, \quad (5)$$

where *MTT*_{pool} is the mean transit time of each pool, calculated for each model and each pool based on the mean value of the pool and its mean output flux. The state variables and parameters used in the ILAMB evaluation are presented in Table 2.

2.4 | Addressing the research questions

Regional evaluation of C cycling metrics across Brazil for both models against the benchmark (Section 3.1.13.1.2) determines model quality for each biome to address question (i). For question (ii), detailed maps are produced showing the spatial variation in signed bias and CI score at pixel scale for all major fluxes and C pools across Brazil. We map the transit times of the large C pools for both LSMs and evaluate their consistency with the benchmark to address question (iii), using the same new metrics. Finally, we analyse time series of mean annual cycles for major C fluxes to evaluate question (iv), taking advantage of the uncertainty on the benchmark to identify where and when LSMs deviate significantly. For brevity, we focus discussion most on the major biomes of the Amazon, Cerrado, Caatinga and Atlantic Forest, although full data are provided for the other biomes.

3 | RESULTS

Overall scores generated by ILAMB for both models at the scale of Brazil provide a starting point for benchmarking (Table 3). The full suite of metrics indicates that some produce clearer indicators of strengths and weaknesses of the models than others. For instance, across Brazil, both models and all variables have similar, low RMSE scores. Thus, direct comparison of RMSE scores provides limited insights. Meanwhile, spatial scores, CI scores and signed bias scores show clearer variation between models and among variables. We therefore focus more on these scores to provide information to differentiate the models and their process representation.

3.1 | Spatial evaluations of C processing and stocks

3.1.1 | Overall evaluation of INLAND

The spatial matching of INLAND with the benchmark identifies clear patterns in bias for fluxes and pools of C (Figure 3). In this figure, the stippling indicates a pixel where INLAND estimates are within the benchmark CI. So, significant biases are in those areas without stipples. The importance of these biases varies by C cycle component. There are some cases where bias scores are large but still the model estimates sit within the 95% CI of the benchmark, such as for *NEE*. There are other cases where bias scores are low, but the model lies outside the benchmark CI, for example, *GPP* in central Cerrado. This


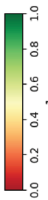
divergence reflects the variation in uncertainty on benchmark data. For instance, INLAND's estimates of *CUE* are significantly different from the benchmark across Brazil, whereas annual *Rh* estimates are not. These results show the value of the new CI metric, which evaluates model bias in the context of benchmark uncertainty.

Biosphere–atmosphere exchanges

INLAND model estimates of *GPP* have a strong match across most of Brazil, fitting within the benchmark CI over 67% of the country (Table 3). The exception is for parts of Cerrado, Pampas and Caatinga where *GPP* values are over-estimated and outside the bounds of CARDAMOM CI (Figure 3). Thus, at the national scale, simulated *GPP* tends to over-predict compared to the benchmark, with a signed bias score of 0.24 (Table 3). Simulated *NPP* is within the benchmark CI across nearly all of Brazil and is generally under-estimated (*NPP* signed bias = -0.15). The *NPP* – *GPP* difference is linked to under-estimation of *CUE* across Brazil and over-estimation of *Ra*, whose signed bias score is equal to 0.63. *Rh* is slightly under-estimated (signed bias score = -0.12) by INLAND across Brazil, but is within benchmark CI for 96% of the spatial points. *NEE* predictions perfectly fit within the benchmark CI (100% of the spatial points) across all of Brazil. However, the *NEE* values do have spatial variability, with some heterogeneous signed bias scores. *NEE* is slightly over-estimated for the Amazon region and over or under-estimated around the borders of Cerrado.

Internal dynamics

Signed bias scores indicate that INLAND tends to over-estimate stocks of $C_{leaf+labile}$ and C_{root} pools, but under-estimate C_{wood} and C_{dead} pools along with their input/output flows (Table 3). Based on CI uncertainty scores INLAND estimates of $C_{leaf+labile}$ lie outside the benchmark CI in around 50% of Brazil, whereas for C_{root} , most estimates fit within the benchmark CI. C_{wood} is significantly underestimated by INLAND particularly in Cerrado, Caatinga and Atlantic Forest, and matches with the benchmark CI and has good bias score in the central Amazon and central Cerrado areas (Figure 3). Wood output flows are under-estimated particularly in Cerrado, Caatinga and Atlantic Forest. The C_{dead} pool is also significantly under-estimated across the whole of Brazil – as shown by the mismatch with the benchmark CI (4%) and a signed bias score equal to -0.50 (Table 3). For each living C pool, the corresponding input and output flow have very similar benchmark matching, as shown by similar signed bias scores distributions (Figure 3). Despite the under-estimation of the dead C pool, the heterotrophic respiration flux closely matches the benchmark.

TABLE 3 Evaluation of benchmarking for JULES and INLAND across Brazil via ILAMB statistical scores (Table 1). The colours of the column 'signed bias score' correspond to the following distribution: , and the other score colours correspond to: 

Variables	Median CAR-DAMOM	Percentiles	Model	Median Value	Signed bias Score	Bias Score	RMSE Score	Seasonal Score	Spatial Score	IAV Score	Overall Score	CI Score
GPP	6.19	2.5pc	INLAND	7.67	0.24	0.68	0.20	0.84	0.87	0.20	0.49	0.67
		97.5pc	JULES	7.43	0.20	0.71	0.39	0.91	0.76	0.41	0.56	0.75
NPP	3.17	2.5pc	INLAND	1.99	-0.15	0.68	0.14	0.79	0.54	0.18	0.39	0.84
		97.5pc	JULES	2.81	0.03	0.70	0.34	0.86	0.40	0.35	0.49	0.90
CUE	0.55	2.5pc	INLAND	0.28	-0.40	0.60	0.00	an.	0.23	an.	0.11	0.22
		97.5pc	JULES	0.40	-0.21	0.78	0.00	an.	0.59	an.	0.34	0.91
NPP _{fol+lab}	0.56	2.5pc	INLAND	0.72	0.30	0.60	0.10	an.	0.93	an.	0.49	0.90
		97.5pc	JULES	0.49	0.05	0.66	0.33	an.	0.82	an.	0.52	0.94
NPP _{wood}	1.27	2.5pc	INLAND	0.43	-0.48	0.50	0.25	an.	0.51	an.	0.27	0.58
		97.5pc	JULES	1.85	0.28	0.60	0.16	an.	0.76	an.	0.47	0.87
NPP _{root}	0.82	2.5pc	INLAND	0.72	0.09	0.57	0.13	an.	0.87	an.	0.45	0.93
		97.5pc	JULES	0.37	-0.27	0.60	0.35	an.	0.39	an.	0.38	0.90
Output _{fol+lab}	1.9e+02	2.5pc	INLAND	269.12	0.32	0.60	0.23	an.	0.94	an.	0.52	0.79
		97.5pc	JULES	179.68	0.07	0.65	0.33	an.	0.81	an.	0.51	0.90
Output _{wood}	3.4e+02	2.5pc	INLAND	36.67	-0.54	0.45	0.37	an.	0.52	an.	0.20	0.05
		97.5pc	JULES	700.92	0.58	0.40	0.33	an.	0.94	an.	0.51	0.49
Output _{root}	3.0e+02	2.5pc	INLAND	290.50	0.14	0.61	0.21	an.	0.91	an.	0.51	0.98
		97.5pc	JULES	134.30	-0.27	0.59	0.35	an.	0.34	an.	0.51	0.91
Rh	2.47	2.5pc	INLAND	1.74	-0.12	0.71	0.21	0.56	0.75	0.45	0.47	0.96
		97.5pc	JULES	2.76	0.15	0.72	0.10	0.47	0.72	0.23	0.43	0.95
Ra	2.72	2.5pc	INLAND	5.54	0.63	0.37	0.21	0.84	0.92	0.21	0.48	0.46
		97.5pc	JULES	4.52	0.45	0.53	0.38	0.90	0.98	0.38	0.60	0.79
C _{fol+lab}	2.5e+02	2.5pc	INLAND	562.75	0.63	0.36	0.14	an.	0.54	an.	0.38	0.48
		97.5pc	JULES	215.63	-0.08	0.71	0.35	an.	0.35	an.	0.44	0.98
C _{wood}	3.4e+03	2.5pc	INLAND	1.6e+03	-0.32	0.61	0.34	an.	0.94	an.	0.41	0.52
		97.5pc	JULES	1.3e+04	0.68	0.31	0.32	an.	0.80	an.	0.46	0.35
C _{root}	1.3e+02	2.5pc	INLAND	462.05	0.87	0.13	0.10	an.	0.58	an.	0.44	0.86
		97.5pc	JULES	215.63	0.43	0.55	0.34	an.	0.84	an.	0.58	0.96
C _{dead}	1.4e+04	2.5pc	INLAND	4.6e+03	-0.50	0.50	0.35	an.	0.37	an.	0.19	0.04
		97.5pc	JULES	1.1e+04	-0.11	0.76	0.33	an.	0.89	an.	0.50	0.75
RECO	5.21	2.5pc	INLAND	7.26	0.35	0.63	0.18	0.73	0.96	0.33	0.54	0.92
		97.5pc	JULES	7.50	0.32	0.65	0.18	0.80	0.93	0.28	0.53	0.92
NEE	-0.62	2.5pc	INLAND	-0.31	0.22	0.45	0.11	0.73	0.36	0.10	0.39	1.00
		97.5pc	JULES	0.05	0.40	0.39	0.16	0.49	0.08	0.13	0.35	1.00
C _{tot}	2.0e+04	2.5pc	INLAND	8.0e+03	-0.43	0.57	0.33	an.	0.75	an.	0.27	0.06
		97.5pc	JULES	2.6e+04	0.21	0.65	0.30	an.	0.66	an.	0.44	0.49

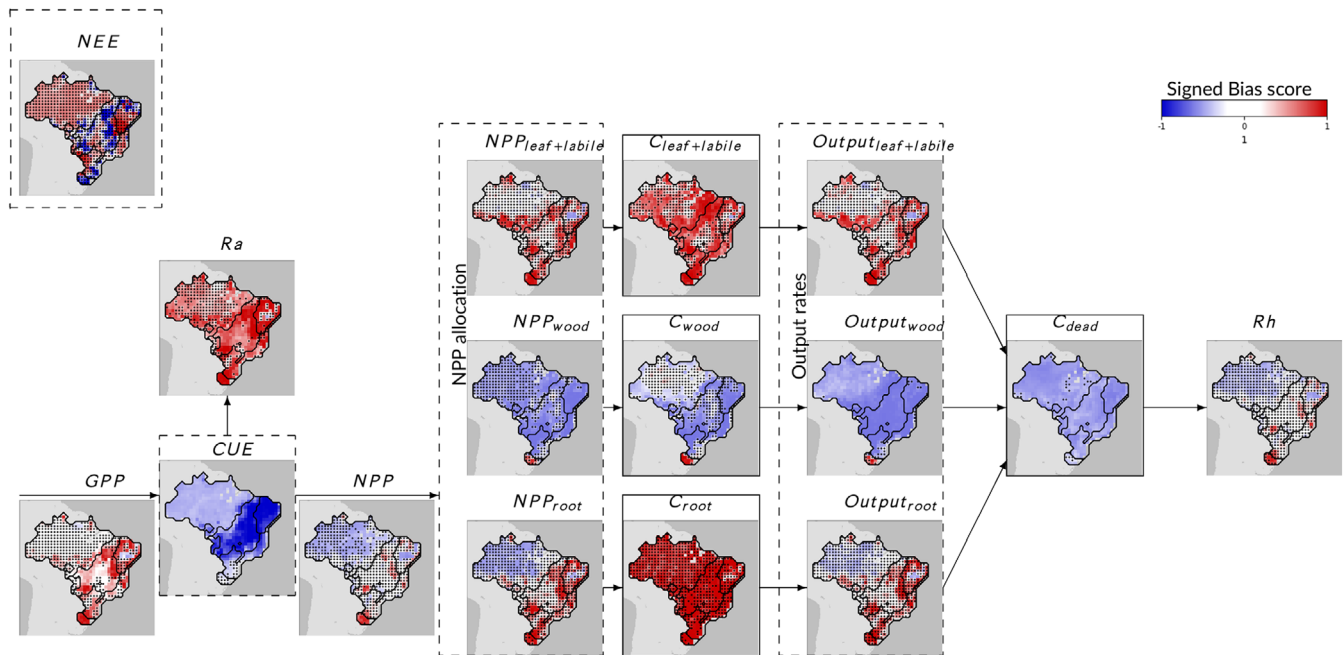


FIGURE 3 Benchmarking for INLAND: Signed bias score (colours) and uncertainty matching (stipples) between CARDAMOM and INLAND. The stipples show if INLAND estimates lie within the benchmark 2.5th and 97.5th percentiles. A negative signed bias score indicates that the model under-estimates the benchmark

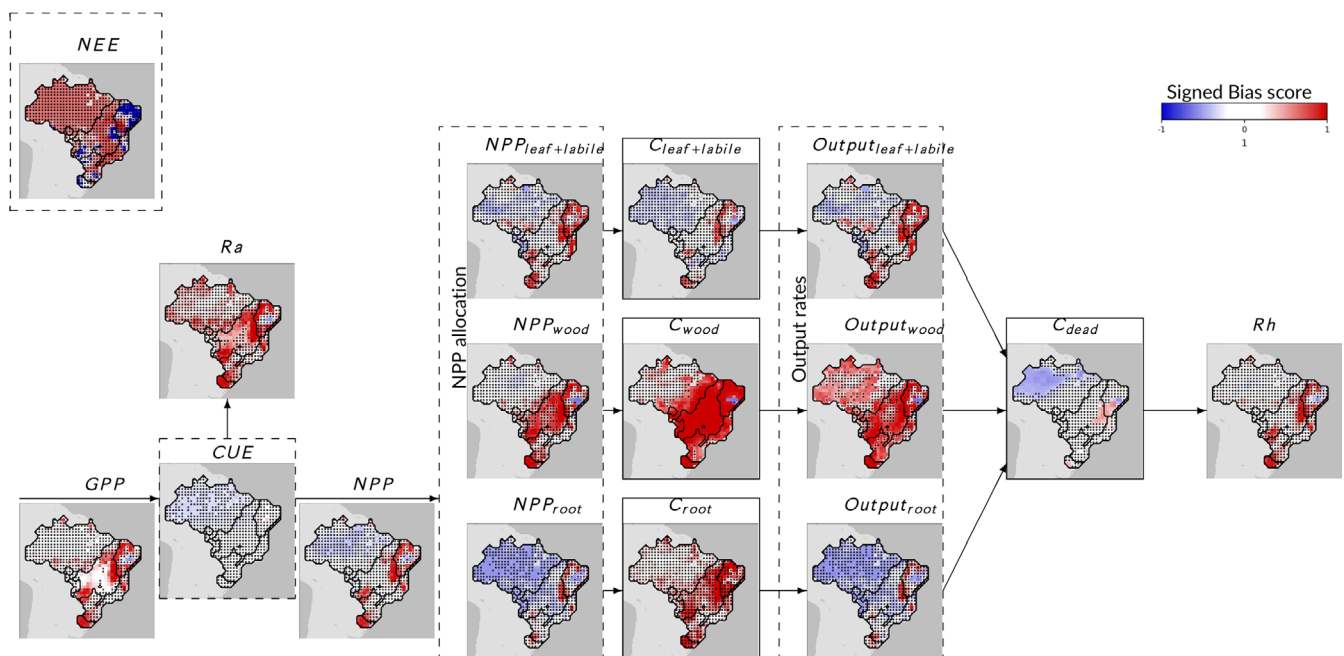


FIGURE 4 Benchmarking for JULES: Signed bias score (colours) and uncertainty matching (stipples) between CARDAMOM and JULES. The stipples show if JULES estimates lie within the benchmark 2.5th and 97.5th percentiles

3.1.2 | Overall evaluation of JULES

Biosphere-atmosphere exchanges

JULES effectively estimates *GPP*, *NPP* and *Rh* flows over the Amazon with relatively low magnitude of signed bias scores and estimates within benchmark CI for > 67% for all

simulated pixels (Figure 4). However, these flows are over-estimated in the Pampas and Cerrado biomes, both in term of larger signed bias and lower CI scores. The *Ra* flux is over-estimated across Brazil (signed bias score = 0.45) and significantly outside the Amazon based on CI scores. *NEE* is broadly over-estimated, except in Cerrado and Pantanal

where it is under-estimated. But JULES *NEE* does not differ significantly from the benchmark, as it still fits with the broad benchmark *NEE* CI across Brazil.

Internal dynamics

JULES estimates of live C pools show varying consistencies with the benchmark. $C_{leaf+labile}$ are consistent with the benchmark (signed bias close to zero, and a CI score of 98%). C_{wood} is significantly over-estimated across Brazil, except in the Amazon. C_{root} is also over-estimated right across Brazil, but is not significantly different from the benchmark based on CI score. Input flow bias and CI score are very similar to output flows for living carbon pools. According to both scores, input and output flows are well modelled, though the Amazonian region tends to be slightly under-estimated, and the Cerrado area is generally over-estimated, especially at its frontier with Pampas and Caatinga biomes. Estimation of $C_{leaf+labile}$ is very similar to the estimation of its input and output flows, for both bias and CI scores. The C_{dead} pool is only slightly under-estimated (with a signed bias score equal to -0.11), and falls within the benchmark percentiles across 75% of Brazil – there are significant underestimates in the western Amazon and the border of Cerrado and Caatinga.

3.2 | The dynamics of the major C pools

JULES and INLAND have similar behaviours regarding the main flows (*GPP*, *NPP*, *Ra* and *Rh*), but show different internal distribution of the carbon among the pools. In this section, we study the impact of this difference in carbon distribution on the modelling of total carbon dynamics. For evaluation of JULES and INLAND, we track total C:

$$C_{tot} = C_{root} + C_{wood} + C_{leaf+labile} + C_{dead}, \quad (6)$$

and then total live C (C_{root} , C_{wood} , $C_{leaf+labile}$) and dead C against the benchmark values. The input (*GPP*) and output ($RECO = Ra + Rh$) fluxes for JULES (Figure 5a) and INLAND (Figure 5b) have very similar spatial distribution of signed bias scores (Figure 5c). Indeed, both models provide an accurate estimation of these fluxes in the Amazon region (shown by stippling), but both over-estimate the flows in the border areas of the Cerrado region and the Pampas. For both models, predictions of *GPP* are significantly different for much of the Cerrado biome. *Reco* estimates also differ in similar areas of the Cerrado for both models, but these biases are less significant than those of *GPP*.

The important difference between JULES and INLAND lies in their estimates of the total C stock and its partition-

ing in live and dead components. JULES over-estimates total C stocks in some areas and INLAND under-estimates total C across most of Brazil. JULES robustly simulates dead C stocks, except for significant deviations in the western Amazon and the Cerrado-Caatinga boundary. JULES significantly over-estimates live C stocks over most of Brazil, with the exception of the western Amazon where live C is well estimated (simulations are within the benchmark CI). INLAND makes robust estimates of live C within the CI of the benchmark over most of Brazil, so its total C bias is attributed largely to its significant underestimate of dead C across nearly all of Brazil. However, INLAND and JULES make robust estimate of the C input into C_{dead} ($Input^{dead} = Output_{wood} + Output_{leaf+labile} + Output_{root}$), in particular, in terms of CI scores (96% for both models). The patterns of bias in $Input^{dead}$ are not correlated with the patterns of bias in C_{tot}^{dead} , suggesting compensating errors in MTT_{dead} exist.

For both the models and the benchmark, the C_{wood} pool is the main contributor to the living C pool – it accounts for 90% of the living C for CARDAMOM, 57% for INLAND and 97% for JULES (Online Appendix Figure A.2). Relative to the benchmark, INLAND under-estimates the contribution of the C_{wood} pool to live C in the Cerrado region and its surroundings. Thus, the dynamics of total C depend on the dynamics of the C_{wood} and C_{dead} pools. However, the dynamics of the pools depend on their inflows $Input_{pool}$ and mean transit times MTT_{pool} (Equation 5). Analysis of transit times of the large C pools indicates that both models under-estimates MTT_{dead} compared to the benchmark (Figure 6). Overall JULES presents better results on MTT_{dead} than INLAND (smaller signed bias scores), especially in the Cerrado and its surroundings. MTT_{wood} is over-estimated by JULES and INLAND, but JULES has smaller signed biases than INLAND, especially on the Amazon region (Online Appendix Figure A.3). For C inputs to pools (i.e. *NPP* to wood, and total litter inputs to dead C), JULES over-estimates fluxes both into wood and dead pools. INLAND, however, more strongly under-estimates inputs to wood, and slightly under-estimates fluxes into dead C.

3.3 | Analysis of seasonal cycles

For the Amazon, the benchmark estimates a very small seasonality and broad CIs for *GPP*, *NEE* and *RECO*, which contain the estimates of JULES and INLAND (Figure 7). Overall, the models are consistent with, and not significantly different from, CARDAMOM across the Amazon. Close examination shows that JULES and CARDAMOM have very similar seasonality in *GPP*, whereas INLAND

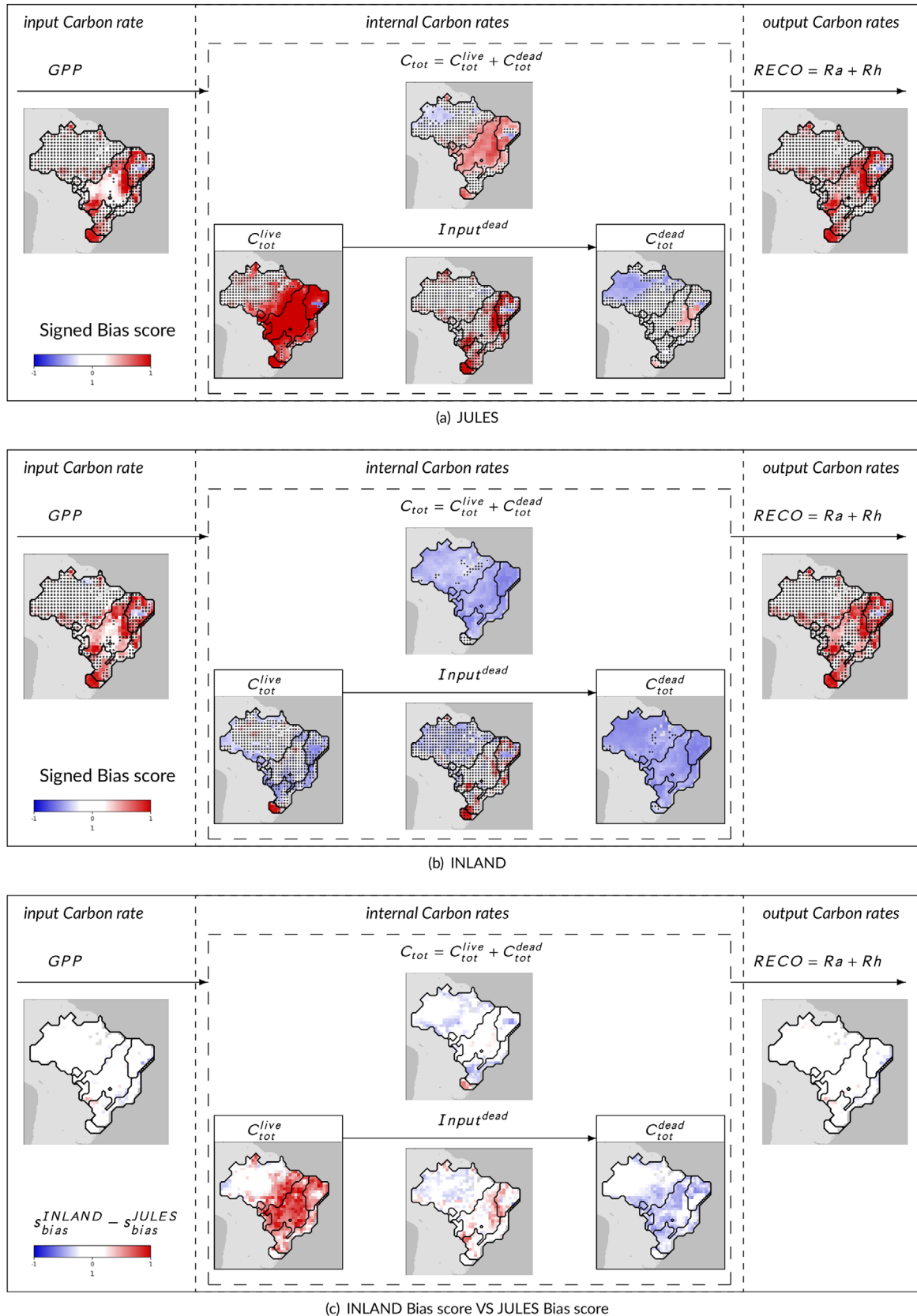


FIGURE 5 Evaluation of *GPP*, *RECO* and internal C cycling pools and flux for JULES and INLAND against the benchmark (a and b), and their comparison (c). Panels a and b show signed bias scores (coloured) and uncertainty matching (stippling, black dots) between CARDAMOM and JULES (a) and between CARDAMOM and INLAND (b). The differences in signed bias scores between the two models are shown in (c)

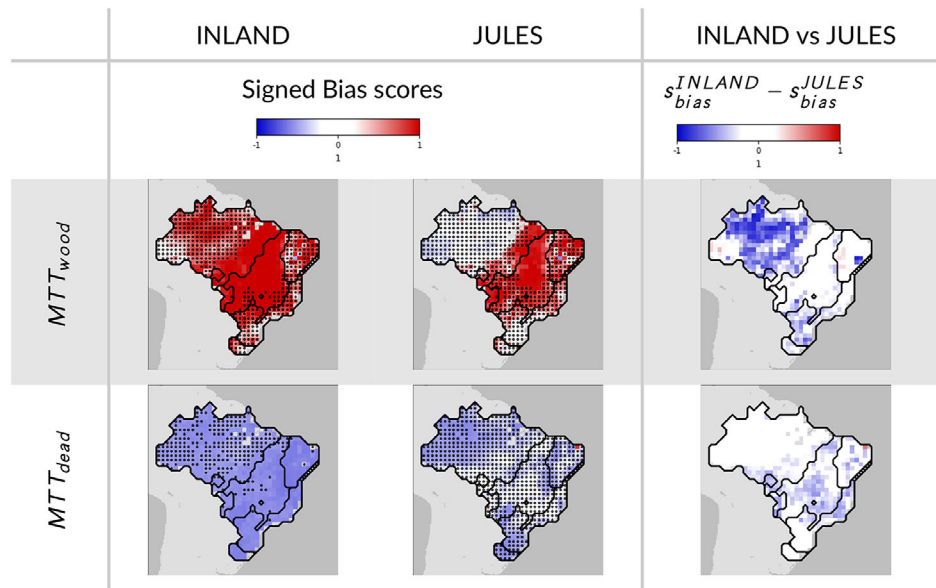


FIGURE 6 Evaluation of the models' mean transit times for wood and dead C pools. The *INLAND* and *JULES* columns show the signed bias scores corresponding to MTT_{wood} and MTT_{dead} for each model; the stipples show that the predictions are within the 2.5th and 97.5th percentiles of the benchmark. The column *INLAND versus JULES* shows the difference between the bias scores of *INLAND* and *JULES* for MTT_{wood} and MTT_{dead}

has almost an opposite cycle, but the differences are small and insignificant.

In Cerrado, *INLAND* *GPP* amplitude is more than twice the magnitude of the benchmark. *INLAND* and *CARDAMOM* peaks and troughs are largely synchronized although *INLAND* has a slightly earlier trough. During October–May, *INLAND* estimates of *GPP* are significantly larger than *CARDAMOM*, whereas during June–September (dry season), they are similar. *JULES* *GPP* has a similar seasonal cycle to *CARDAMOM*, but is consistently biased larger. This bias is largest in the dry season, so it is during May–October that *JULES* *GPP* is significantly different from the benchmark *GPP*. Cerrado *RECO* for both models is similar in amplitude, magnitude and timing. Both models have *RECO* estimate always greater than the benchmark. Cerrado *NEE* seasonality for the LSMs is very different, with a dry season peak source for *INLAND* (and the benchmark) and a dry season sink for *JULES*, which is significantly different from the benchmark, during June–September.

In the Caatinga, the patterns are broadly similar to Cerrado. Models and benchmark have similar *GPP* cycles, but the models tend to over-estimate fluxes, with *INLAND* differing significantly during October–November. For *RECO*, the models have similar patterns and tend to over-estimate seasonality compared to the benchmark median, but are not significantly different. For *NEE*, *INLAND* and the benchmark have similar seasonality and although *INLAND* has greater amplitude it rarely differs significantly from *CARDAMOM*. *JULES* has the opposite

behaviour to *INLAND* and the benchmark, indicating a dry season sink, although there is only a significant difference to the benchmark in August.

For the Atlantic Forest, *GPP* seasonality is similar for LSMs and benchmark, although the models are biased high and *INLAND* is therefore significantly over-estimating during October–November. *RECO* shows a similar seasonal cycle for models and *CARDAMOM*, but the models have greater amplitude and are biased high, but not significantly so. The *NEE* produces conflicting results due to differences in *GPP* and *Reco* seasonality; *INLAND* and *CARDAMOM* benchmarks have similar seasonality in *NEE* but *JULES* has a dry season sink and differs significantly from *CARDAMOM* during August.

4 | DISCUSSION

This study evaluates two LSMs (*INLAND* and *JULES*) by comparing them to a probabilistic benchmarking dataset at different spatial scales (Brazil and its key biomes), temporal scales (inter- and intra-annual) and conceptual scales – biosphere–atmosphere exchanges and internal carbon dynamics. The *ILAMB* framework has allowed clear identification of both strengths and weaknesses of each model, thus identifying targets for improvement. However, it was new metrics (a CI score and a signed bias score) added here to *ILAMB* that provided the most useful information for the model evaluations. The CI score takes advantage of the CIs (CI) provided at 1 ζ resolution in the benchmark. The

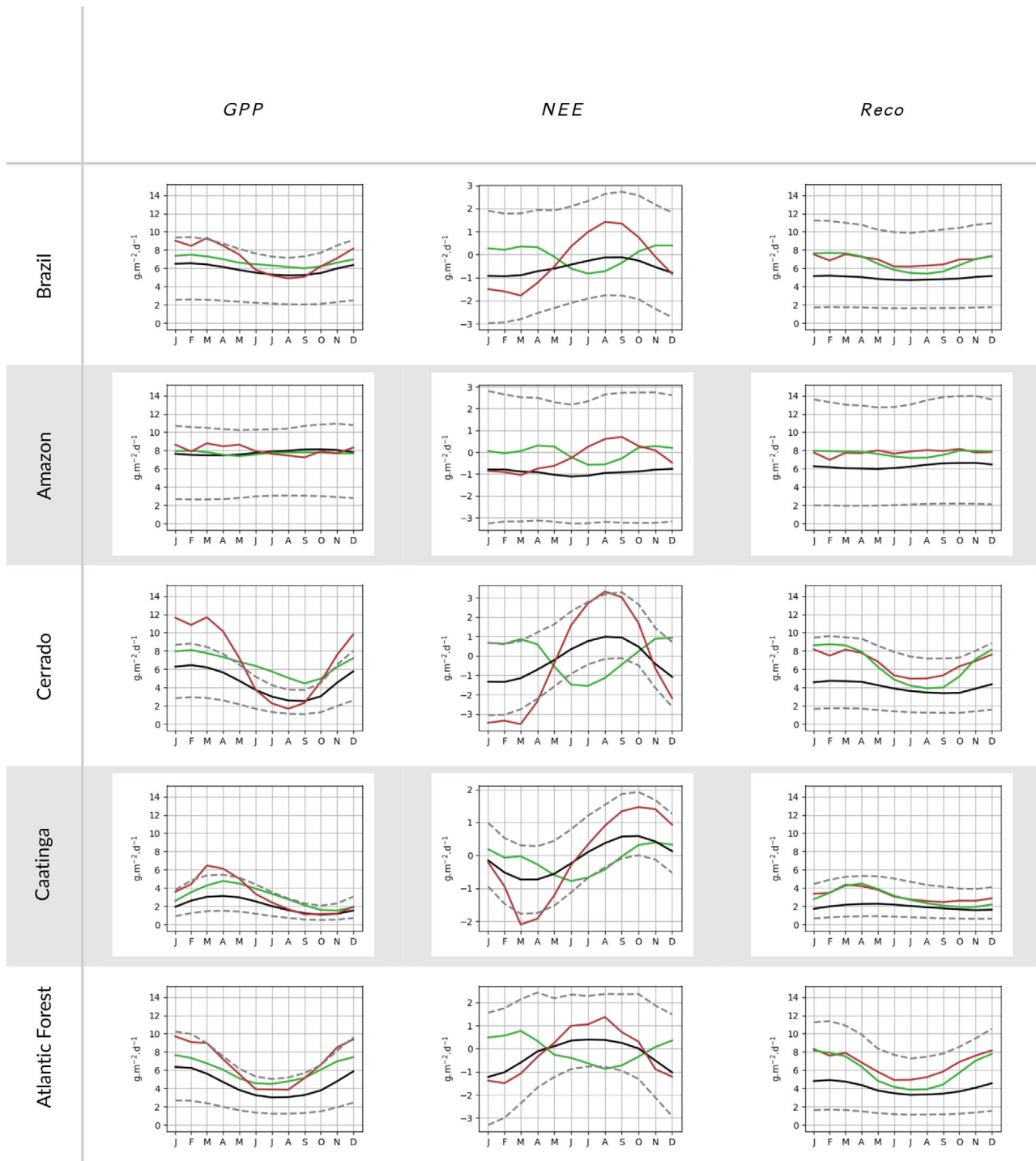


FIGURE 7 Annual cycle by region for daily C fluxes of *GPP*, *NEE* and *RECO*: CARDAMOM (■) and its 2.5 and 97.5 percentiles (■), JULES (■), INLAND (■)

score determines the percentage of pixels within a domain for which the model component is within the CI, allowing a significance test of model-benchmark difference in ILAMB for the first time. The signed bias score allows effective mapping of the degree of model over- or under-estimation, generating deeper insights into how process interactions generate model mismatches.

4.1 | Which biomes have the most consistency with the benchmark?

Independently of the variables and models considered, there was a consistency between models in the geographical variation in their predictive quality. Focusing here on the largest biomes, the Amazon had the most

robust and spatially homogeneous scores for both models (Online Appendix Table A.1). Pools and flux estimates were within the benchmark CI for > 90% of pixels in most cases (Figure 3, Figure 4). The clear exception was an under-estimation of the DOM pool in both INLAND and JULES (the latter for the western Amazon). DOM dynamics in the Amazon had quite different processing between the two LSMs; JULES over-estimated DOM inputs from wood, whereas INLAND under-estimated DOM input. For Atlantic Forest, Cerrado and Caatinga, both LSMs over-estimated GPP in all cases. There were also very different dynamics of the woody pool between models. INLAND under-estimated allocation to wood, wood litter outputs and the size of the woody pool, whereas JULES over-estimated all these. INLAND consistently under-estimated the size of the DOM pool.

Weak scores were often localized around the borders of the regions – especially around the Cerrado and the Pampas. These spatial similarities between both models were confirmed by the INLAND-JULES comparison (Online Appendix Figure A.1). The Brazilian biomes correspond to broadly homogeneous areas in terms of vegetation type and climate. Therefore, the weaker results at biome borders suggest that the models are less effective at simulating ecotones. For the smaller biomes (Pampas, Pantanal), their larger modelling errors likely arise due a greater proportion of ecotone (edges) relative to total biome area. The ecotone inaccuracy for Cerrado may arise for a number of reasons. Both LSMs use a PTF approach, which may have biases. PFTs are likely calibrated to match the dominant vegetation for each biome, rather than functional variations at ecotones. LSMs could be improved by developing alternate PFTs that describe ecotone vegetation. Alternatively, ecotones may be areas of increased disturbance, with disruptions to ecosystem processes challenging the steady-state approach in LSMs. For example, the Amazon–Cerrado boundary is an area where land-use change and fires are stimulated by ongoing human activity, leading to more dynamic C cycling. The CARDAMOM benchmark includes fire disturbance imposed through burned area data driving combustion losses. The two LSMs do not include specific fire disturbance in these simulations. So, the difference with the benchmark at the Cerrado boundary likely reflects the explicit role of fire disturbance in C cycling in the benchmark analysis. The biomass map assimilated into the benchmark will also reflect this zone of disturbance and influence the C cycling of this ecotone in CARDAMOM. An LSM improvement in this case would be to include spatially variable disturbance information in the forcing, for example, from fire. This approach might require finer resolution modelling at ecotones to manage the heterogeneity of these disturbed landscapes. JULES

does not use observed vegetation cover; instead, the distribution of PFTs is estimated by the model. Errors in the PFT distribution will cause errors in the sizes of C pools and fluxes. For example, an overestimation of tree cover in the Cerrado may be responsible for the positive C_{wood} bias. Evaluating the PFT composition in the LSMs at the ecotone would be useful in this regard. Improving the PFT distribution scheme may improve the scores for the JULES model.

4.2 | Evaluation of biosphere–atmosphere flows and internal carbon dynamics

The CIs on the benchmark estimate of NEE are large across Brazil, so that even with large bias scores, the LSM estimates are not significantly different from the benchmark. For this reason, the CI score was 100% across Brazil for NEE for both models (Table 3). However, the evaluation of fluxes driving NEE showed clear patterns in errors. Both LSMs tended to over-estimate GPP , Ra and Rh across Brazil, significantly so in Cerrado and some other biomes (Figure 34). These over-estimates tend to cancel out, leading to a small signed bias in NEE , but clear differences in the rates of the processes which drive NEE . Over-estimates of Ra were high, particularly in INLAND, and correlated to a poorly evaluated CUE .

The exchanges between the biosphere and the atmosphere are spatially similar for both models (Online Appendix Figure A.1) but the internal flows of C to live and dead C pools are different between the two models. Across Brazil, there is a large disparity in the estimates of live and dead C pools, with divergent behaviour between LSMs. For wood C, INLAND has significant under-estimates in 48% of Brazil, whereas JULES has significant over-estimates across 65% of Brazil, and a poorer bias score. For DOM, INLAND under-estimates for nearly all (98%) of Brazil, whereas JULES is consistent across 74% of the country and has a better bias score. Overall, INLAND is better at modelling the dynamics of wood C, whereas JULES models the leaf and root with greater accuracy (Table 3). Although JULES is more robust at estimating DOM C, there is little difference between the two models in the quality of their heterotrophic respiration estimates, that is, turnover of DOM. In JULES, the spatial distribution of Rh and $Output_{leaf}$ biases are very similar, suggesting that the over-estimation of leaf litter is causing an over-estimation in Rh . For JULES, there is no spatial coherence between biases in $Output_{wood}$, C_{dead} and Rh , suggesting a bias in the turnover of wood litter that compensates for the large over-estimation of $Output_{wood}$, particularly in the Cerrado (Figure 4).

4.3 | How reliable are representations of seasonality in the LSMs?

Our analysis highlights that both models have similar seasonal behaviours for gross input (GPP) and output ($RECO$) flows, and these broadly match the benchmark (Figure 7). Seasonality of LAI, however, differs, with JULES showing little to no LAI variation, and INLAND and the benchmark having similar periods (Online Appendix Figure A.4). The benchmark has consistently lower LAI than either LSM. With its lack of LAI seasonality, JULES seasonality in GPP must be driven by physiological rather than phenological factors.

However, for the balance of these flows (NEE), JULES has an inverted seasonality for several biomes, but a similar amplitude to the benchmark. INLAND NEE displays stronger seasonal variations (amplitude) than the benchmark, but with a similar seasonality (Figure 7). These results underline the phenomena of compensation or accumulation of errors that may occur when one variable is deduced from other variables. For net ecosystem exchanges in both LSMs, there is a clear phenomenon of error compensation in the integration over time with respect to uncertainties. The over-large amplitudes of INLAND's GPP and $RECO$ compensate each other, and the inverse seasonalities of JULES's gross fluxes compensate each other.

The benchmark CIs generally include the seasonal curves of both LSMs. However, seasonally, INLAND is more consistent with the benchmark, because the growth and decay properties of its simulation of NEE match those of the median value and the benchmark CIs. It is interesting that the seasonality of JULES's GPP and $RECO$ across biomes both broadly match the benchmark, but slight differences in their seasonality and amplitudes means that JULES net exchange (NEE) is strongly out of phase with the benchmark, particularly in more seasonal biomes, like Caatinga. JULES' behaviour is likely due to different couplings between input and output process rates from its simulation of internal C cycling compared to INLAND and the benchmark. Both the benchmark and INLAND peak sink strength (minimum NEE) estimate for Caatinga are during April, consistent with independent flux data, whereas JULES peak uptake in June is not (Mendes et al., 2020). However, the CI score shows that INLAND overestimates the strength of the peak sink in Caatinga. Benchmarking against independent estimates of net biome exchange from atmospheric inversions will provide an opportunity to further evaluate model seasonality. A complexity in such evaluations is that CO_2 emissions from fire and land-use change will have to be controlled for also.

TABLE 4 Summary of the impacts of inputs and mean transit times on wood and DOM Carbon pools. In the column *impact on $\frac{dC_{pool}}{dt}$* , a + means that the variable in the corresponding row has an increasing impact on the dynamics of the concerned Carbon pool, and – means a negative impact. These influences come from the qualitative analysis of carbon pool dynamics, Equation 5, and are not specific to a particular model. The columns *JULES* and *INLAND* summarize the signed bias scores for each variable corresponding to the different rows

Variables	impact on $\frac{dC_{pool}}{dt}$	Signed bias scores	
		JULES	INLAND
$Input_{wood}$	+	0.28	–0.48
MTT_{wood}	+	0.37	0.79
C_{wood}	–	0.68	–0.32
$Input_{dead}$	+	0.22	–0.08
MTT_{dead}	+	–0.35	–0.54
C_{dead}	–	–0.10	–0.50

4.4 | Transit times

This analytical analysis at the system level provides insights into the model results. Indeed, the analysis of C dynamics highlights the dependencies between net fluxes (NEE) and aggregated variables (total C) and their underlying variables (component fluxes or pools). Thus, in order to explain the simulation quality of aggregated variables, we must consider the accumulation and compensation of errors of the underlying variables. In that respect, the over-estimation of the total C by JULES is due to the over-estimation of the living C and the correct estimation of the DOM C; conversely, the under-estimation of the total C by INLAND is due to the correct estimation of the living C and the under-estimation of the dead C. At a finer scale, we show the same phenomenon of compensation or accumulation between the inflows and retention times of the wood C and the dead C (Section 3.2).

Our analysis explains why JULES better models dead C stocks, while INLAND better models the living C stocks (Table 4). The basic equation for C pools (Equation 5) highlights that C dynamics is an increasing function of inflows and retention times. Thus, the over-estimation of $INPUT_{wood}$ and MTT_{wood} leads JULES to strongly over-estimate the C_{wood} pool. For INLAND, the under-estimation of $INPUT_{wood}$ is compensated by the over-estimation of MTT_{wood} , so INLAND only slightly under-estimates the C_{wood} pool. For the C_{dead} pool, the behaviour of JULES and INLAND is reversed: JULES under-estimates MTT_{dead} , but over-estimates $INPUT_{dead}$, and thus by compensation of the errors, only slightly under-estimates C_{dead} . On the other hand, INLAND

under-estimates MTT_{dead} and correctly estimates $INPUT_{dead}$, and thus strongly under-estimates C_{dead} .

4.5 | Implications for model forecasts and development priorities

The over-estimation of C_{wood} in JULES outside the Amazon will lead to the model overestimating the carbon budget implications of future land-use change and climate change in these regions. The underestimation of C_{dead} by INLAND will lead to the model underestimating the carbon budget implications of increased temperatures, through decomposition feedbacks.

For JULES, model development should focus on improving (i) GPP simulations for the Cerrado, currently over-productive; (ii) simulation of vegetation C dynamics outside the Amazon, which currently over-estimating live C pools; (iii) seasonality of NEE predictions, which are currently out of phase. For INLAND, model development areas are (i) seasonal variability of biosphere–atmosphere fluxes, which have a larger magnitude of seasonal variation than indicated by the benchmark, perhaps linked to a high sensitivity to soil moisture stress; (ii) adjusting inputs to and mean transit times of dead C stocks to increase their magnitude and (iii) calibration of carbon use efficiency, as currently INLAND allocates too much photosynthate to autotrophic respiration. The photosynthesis and respiration schemes for INLAND need to be evaluated at site scale to test its process representation.

This study highlights the link between the different conceptual scales. Indeed, a strong matching of a variable between a model and a benchmark may be due to the fact that its dependencies are themselves correctly simulated, but alternatively to the fact that the errors of its dependencies compensate each other. Thus, the accuracy of a model can vary according to the considered conceptual scales: JULES and INLAND correctly model the aggregated system, but they fail to model internal pools with the same accuracy. Therefore, improving the quality of the simulation of a variable requires an analytical study of its dependencies, in order to focus on the key processes.

4.6 | Caveats and challenges for model benchmarking

The benchmark itself will be biased as it is based on a single-model structure. Future benchmarks could be developed from range of different model structures linked to observations via model-data fusion to describe structural error (Famiglietti et al., 2021). The focus of this paper was on benchmarking the ecological C cycle, its physiolog-

ical and intrinsic factors. However, both the benchmark and the LSMs included disparate human factors, such as land-use and land-use change forcing, and different fire regimes. Standardized approaches to including management and other extrinsic factors would improve the benchmarking process. The errors and biases within the assimilated information (e.g. soil C maps) remain poorly determined. We use a recent fusion of various biomass maps, but recognize that this has a poor determination of error particularly outside the Amazon. Biomass maps were from particular time periods, and could include biases due to conditions during the measurement periods. Further work should involve using the next generation of biomass products, including time series, and working with data teams to understand the error properties of these products. There is also evidence of changes in leaf traits associated with ageing (Wu et al., 2016) that influence seasonal patterns in GPP and therefore C cycling. The benchmark does not include these ageing properties, and further work should ascertain the uncertainty related to these trait dynamics for both benchmark and models.

5 | CONCLUSIONS

This study has used a benchmark and a comparison framework (ILAMB) to evaluate two land surface models (LSMs) for C dynamics across Brazilian ecosystem. The complexity of such an analysis resides in these models' large number of variables and their spatial and temporal dimensions. ILAMB provides an effective means to summarize and compare the models at biome scale and varied time scales. The benchmark is probabilistic, based on assimilation of multiple independent C cycle observations into an intermediate complexity mass balance model. The benchmark allows LSM evaluation against key pools and fluxes, including internal cycling. For the first time, biases in the LSMs are tested for significance using benchmark confidence intervals. New metrics within ILAMB map significant biases in space and time, and whether the models over- or under-estimate key components of the C cycle.

Spatial analysis indicates that the models give better simulations in homogeneous areas of vegetation type, and are less efficient at ecotones between biomes. Although both models have net ecosystem exchanges between the biosphere and the atmosphere that do not differ significantly from the benchmark, they do have significant differences in internal carbon allocation and the dynamics of the different C pools. JULES models the dead C stocks more accurately, whereas INLAND better resolves living C stocks. The models are more efficient in simulating annual averages than seasonal variations. For some variables, benchmark uncertainty is too high to provide an accurate

evaluation of the models, so it is important to improve the benchmark precision on these specific variables. Finally, while the benchmark takes into account uncertainties of variables, the models, on the other hand, return deterministic evaluations of these variables. A more realistic modelling approach would thus consider and report this parametric uncertainty, through developing probabilistic models.

ACKNOWLEDGEMENTS

We acknowledge support from the Newton Fund through the Met Office Climate Science for Service Partnership Brazil (CSSP Brazil) and UK Space Agency Forests 2020 funding. MW acknowledges funding from the Royal Society. MC acknowledges the support from the São Paulo Research Foundation (FAPESP, Process 2015/50122-0). INLAND developments were supported by Sao Paulo Research Foundation (FAPESP, grant 2017/22269-2). We thank Jean Ometto for access to the Amazon biomass map. We thank A.A. Bloom for inputs on the CARDAMOM MCMC routines used in this study. Darren Slevin and Declan Valters provided advice and support in setting up the ILAMB system. LAI information assimilated by CARDAMOM was generated by the Global Land Service of Copernicus, the Earth Observation programme of the European Commission. The LAI product is based on SPOT-VEGETATION 1km data (copyright CNES and distribution by VITO). We thank the ILAMB developers for access to their code (<https://www.ilamb.org>).

DATA AND CODE AVAILABILITY

The updated ILAMB code and data files used in this study are available on Edinburgh Datashare: <https://doi.org/10.7488/ds/3052>

The benchmark outputs from CARDAMOM are available on Edinburgh Datashare: <https://doi.org/10.7488/ds/2991>

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ORCID

Mathew Williams  <https://orcid.org/0000-0001-6117-5208>

REFERENCES

- Ahlström, A., Schurgers, G., Arneeth, A. & Smith, B. (2012) Robustness and uncertainty in terrestrial ecosystem carbon response to CMIP5 climate change projections. *Environmental Research Letters*, 7(4), 044008.
- Avitabile, V., Herold, M., Heuvelink, G.B.M., Lewis, S.L., Phillips, O.L., Asner, G.P. et al. (2016) An integrated pan-tropical biomass map using multiple reference datasets. *Global Change Biology*, 22(4), 1406–1420. <https://onlinelibrary.wiley.com/doi/abs/10.1111/gcb.13139>.
- Bloom, A.A., Exbrayat, J.F., van der Velde, I.R., Feng, L. & Williams, M. (2016) The decadal state of the terrestrial carbon cycle: global retrievals of terrestrial carbon allocation, pools, and residence times. *Proceedings of the National Academy of Sciences*, 113(5), 1285–1290. <https://www.pnas.org/content/113/5/1285>.
- Bloom, A.A. & Williams, M. (2015) Constraining ecosystem carbon dynamics in a data-limited world: integrating ecological “common sense” in a model–data fusion framework. *Biogeosciences*, 12(5), 1299–1315. <https://bg.copernicus.org/articles/12/1299/2015/>.
- Bonan, G.B. (2008) Forests and climate change: forcings, feedbacks, and the climate benefits of forests. *Science*, 320(5882), 1444–1449. <http://www.sciencemag.org/cgi/content/abstract/320/5882/1444>.
- Carvalho, N., Forkel, M., Khomik, M., Bellarby, J., Jung, M., Migliavacca, M. et al. (2014) Global covariation of carbon turnover times with climate in terrestrial ecosystems. *Nature*, 514(7521), 213–217. <https://doi.org/10.1038/nature13731>.
- Clark, D., Mercado, L., Sitch, S., Jones, C., Gedney, N., Best, M. et al. (2012) The joint UK land environment simulator (JULES), model description—Part 2: carbon fluxes and vegetation. *Geoscientific Model Development*, 4, 701–722.
- Collier, N., Hoffman, F.M., Lawrence, D.M., Keppel-Aleks, G., Koven, C.D., Riley, W.J. et al. (2018) The international land model benchmarking (ILAMB) system: design, theory, and implementation. *Journal of Advances in Modeling Earth Systems*, 10(11), 2731–2754. <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018MS001354>.
- Exbrayat, J.F., Bloom, A.A., Falloon, P., Ito, A., Smallman, T.L. & Williams, M. (2018) Reliability ensemble averaging of 21st century projections of terrestrial net primary productivity reduces global and regional uncertainties. *Earth System Dynamics*, 9(1), 153–165. <https://esd.copernicus.org/articles/9/153/2018/>.
- Famiglietti, C.A., Smallman, T.L., Levine, P.A., Flack-Prain, S., Quetin, G.R., Meyer, V. et al. (2021) Optimal model complexity for terrestrial carbon cycle prediction. *Biogeosciences*, 18(8), 2727–2754. <https://bg.copernicus.org/articles/18/2727/2021/>.
- Foley, J.A., Prentice, I.C., Ramankutty, N., Levis, S., Pollard, D., Sitch, S. et al. (1996) An integrated biosphere model of land surface processes, terrestrial carbon balance, and vegetation dynamics. *Global biogeochemical cycles*, 10(4), 603–628.
- Fox, A., Williams, M., Richardson, A.D., Cameron, D., Gove, J.H., Quaife, T. et al. (2009) The REFLEX project: comparing different algorithms and implementations for the inversion of a terrestrial ecosystem model against eddy covariance data. *Agricultural and Forest Meteorology*, 149(10), 1597–1615. <http://www.sciencedirect.com/science/article/pii/S0168192309001014>.
- Friedlingstein, P., Meinshausen, M., Arora, V.K., Jones, C.D., Anav, A., Liddicoat, S.K. et al. (2014) Uncertainties in CMIP5 climate projections due to carbon cycle feedbacks. *Journal of Climate*, 27(2), 511–526. <https://doi.org/10.1175/JCLI-D-12-00579.1>.
- Friend, A.D., Lucht, W., Rademacher, T.T., Keribin, R., Betts, R., Cadule, P. et al. (2014) Carbon residence time dominates uncertainty in terrestrial vegetation responses to future climate and atmospheric CO₂. *Proceedings of the National Academy of Sciences*, 111(9), 3280–3285. <https://www.pnas.org/content/111/9/3280>.
- Giglio, L., Boschetti, L., Roy, D.P., Humber, M.L. & Justice, C.O. (2018) The collection 6 MODIS burned area mapping algorithm

- and product. *Remote Sensing of Environment*, 217, 72–85. <http://www.sciencedirect.com/science/article/pii/S0034425718303705>.
- Hansen, M.C., Potapov, P.V., Moore, R., Hancher, M., Turubanova, S.A., Tyukavina, A. et al. (2013) High-resolution global maps of 21st-century forest cover change. *Science*, 342(6160), 850–853. <https://science.sciencemag.org/content/342/6160/850>.
- Harris I.C. (2019) CRU JRA v1.1: a forcings dataset of gridded land surface blend of Climatic Research Unit (CRU) and Japanese reanalysis (JRA) data; Jan.1901 - Dec.2017. *Centre for Environmental Data Analysis*, <https://doi.org/10.5285/13f3635174794bb98cf8ac4b0ee8f4ed>.
- Hengl, T., Mendes de Jesus, J., Heuvelink, G.B.M., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A. et al. (2017) SoilGrids250m: global gridded soil information based on machine learning. *PLoS One*, 12(2), 1–40. <https://doi.org/10.1371/journal.pone.0169748>.
- Hoffman, F.M., Randerson, J.T., Arora, V.K., Bao, Q., Cadule, P., Ji, D. et al. (2014) Causes and implications of persistent atmospheric carbon dioxide biases in Earth System Models. *Journal of Geophysical Research: Biogeosciences*, 119(2), 141–162. <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2013JG002381>.
- Hurt G., Chini L., Sahajpal R., Froking S., Bodirsky B.L., Calvin K. et al. (2019a) Harmonization of global land use change and management for the period 2015–2300. *Earth System Grid Federation*, <https://doi.org/10.22033/ESGF/input4MIPs.10468>.
- Hurt G., Chini L., Sahajpal R., Froking S., Bodirsky B.L., Calvin K. et al. (2019b) Harmonization of global land use change and management for the period 850–2015. *Earth System Grid Federation*, <https://doi.org/10.22033/ESGF/input4MIPs.10454>.
- Hurt G.C., Chini, L., Sahajpal, R., Froking, S., Bodirsky, B.L., Calvin, K. et al. (2020) Harmonization of global land use change and management for the period 850–2100 (LUH2) for CMIP6. *Geoscientific Model Development*, 13(11), 5425–5464.
- Jones, C.D., Arora, V., Friedlingstein, P., Bopp, L., Brovkin, V., Dunne, J. et al. (2016) C4MIP - the coupled climate-carbon cycle model intercomparison project: experimental protocol for CMIP6. *Geoscientific Model Development*, 9(8), 2853–2880. <https://www.geosci-model-dev.net/9/2853/2016/>.
- Jung, M., Reichstein, M., Schwalm, C.R., Huntingford, C., Sitch, S., Ahlström, A. et al. (2017) Compensatory water effects link yearly global land CO₂ sink changes to temperature. *Nature*, 541(7638), 516–520. <https://doi.org/10.1038/nature20780>.
- Koven, C.D., Ringeval, B., Friedlingstein, P., Ciais, P., Cadule, P., Khvorostyanov, D. et al. (2011) Permafrost carbon-climate feedbacks accelerate global warming. *Proceedings of the National Academy of Sciences*, 108(36), 14769–14774. <https://www.pnas.org/content/108/36/14769>.
- Kucharik, C.J., Foley, J.A., Delire, C., Fisher, V.A., Coe, M.T., Lenters, J.D. et al. (2000) Testing the performance of a dynamic global ecosystem model: water balance, carbon balance, and vegetation structure. *Global Biogeochemical Cycles*, 14(3), 795–825.
- van der Laan-Luijkx, I.T., van der Velde, I.R., Krol, M.C., Gatti, L.V., Domingues, L.G., Correia, C.S.C. et al. (2015) Response of the Amazon carbon balance to the 2010 drought derived with CarbonTracker South America. *Global Biogeochemical Cycles*, 29(7), 1092–1108. <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2014GB005082>.
- Longo, M., Keller, M., dos Santos, M.N., Leitold, V., Pinagé, E.R., Bacchini, A. et al. (2016) Aboveground biomass variability across intact and degraded forests in the Brazilian Amazon. *Global Biogeochemical Cycles*, 30(11), 1639–1660. <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2016GB005465>.
- Luo, Y., Weng, E., Wu, X., Gao, C., Zhou, X. & Zhang, L. (2009) Parameter identifiability, constraint, and equifinality in data assimilation with ecosystem models. *Ecological Applications*, 19(3), 571–574. <https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1890/08-0561.1>.
- Mendes, K.R., Campos, S., da Silva, L.L., Mutti, P.R., Ferreira, R.R., Medeiros, S.S. et al. (2020) Seasonal variation in net ecosystem CO₂ exchange of a Brazilian seasonally dry tropical forest. *Scientific Reports*, 10(1), 9454. <https://doi.org/10.1038/s41598-020-66415-w>.
- Nishina, K., Ito, A., Beerling, D.J., Cadule, P., Ciais, P., Clark, D.B. et al. (2014) Quantifying uncertainties in soil carbon responses to changes in global mean temperature and precipitation. *Earth System Dynamics*, 5(1), 197–209. <https://esd.copernicus.org/articles/5/197/2014/>.
- Nishina, K., Ito, A., Falloon, P., Friend, A.D., Beerling, D.J., Ciais, P. et al. (2015) Decomposing uncertainties in the future terrestrial carbon budget associated with emission scenarios, climate projections, and ecosystem simulations using the ISI-MIP results. *Earth System Dynamics*, 6(2), 435–445. <https://esd.copernicus.org/articles/6/435/2015/>.
- Saatchi, S.S., Harris, N.L., Brown, S., Lefsky, M., Mitchard, E.T.A., Salas, W. et al. (2011) Benchmark map of forest carbon stocks in tropical regions across three continents. *Proceedings of the National Academy of Sciences*, 108(24), 9899–9904. <https://www.pnas.org/content/108/24/9899>.
- Sitch, S., Friedlingstein, P., Gruber, N., Jones, S.D., Murray-Tortarolo, G., Ahlström, A. et al. (2015) Recent trends and drivers of regional sources and sinks of carbon dioxide. *Biogeosciences*, 12(3), 653–679.
- Sitch, S., Smith, B., Prentice, I.C., Arneth, A., Bondeau, A., Cramer, W. et al. (2003) Evaluation of ecosystem dynamics, plant geography and terrestrial carbon cycling in the LPJ dynamic global vegetation model. *Global Change Biology*, 9(2), 161–185. <https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1365-2486.2003.00569.x>.
- Smallman T.L., Exbrayat J.F., Mencuccini M., Bloom A.A. & Williams M. Assimilation of repeated woody biomass observations constrains decadal ecosystem carbon cycle uncertainty in aggrading forests. *Journal of Geophysical Research: Biogeosciences*, 122(3), 528–545. <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2016JG003520>.
- Smallman, T.L., Milodowski, D.T., Neto, E.S., Koren, G., Ometto, J. & Williams, M. (2021) Parameter uncertainty dominates C cycle forecast errors over most of Brazil for the 21st Century. *Earth System Dynamics Discuss*, 2021, 1–52. <https://esd.copernicus.org/preprints/esd-2021-17/>.
- Thurner, M., Beer, C., Ciais, P., Friend, A.D., Ito, A., Kleidon, A. et al. (2017) Evaluation of climate-related carbon turnover processes in global vegetation models for boreal and temperate forests. *Global Change Biology*, 23(8), 3076–3091. <https://onlinelibrary.wiley.com/doi/abs/10.1111/gcb.13660>.
- Thurner, M., Beer, C., Santoro, M., Carvalhais, N., Wutzler, T., Schepaschenko, D. et al. (2014) Carbon stock and density of northern boreal and temperate forests. *Global Ecology and Biogeography*, 23(3), 297–310. <https://onlinelibrary.wiley.com/doi/abs/10.1111/gcb.12125>.

- Warszawski, L., Frieler, K., Huber, V., Piontek, F., Serdeczny, O. & Schewe, J. (2014) The inter-sectoral impact model intercomparison project (ISI-MIP): project framework. *Proceedings of the National Academy of Sciences*, 111(9), 3228–3232. <https://www.pnas.org/content/111/9/3228>.
- Williams, M., Schwarz, P.A., Law, B.E., Irvine, J. & Kurpius, M.R. (2015) An improved analysis of forest carbon dynamics using data assimilation. *Global Change Biology*, 11(1), 89–105. <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2486.2004.00891.x>.
- Woodward, F.I., Smith, T.M. & Emanuel, W.R. (1995) A global land primary productivity and phytogeography model. *Global Biogeochemical Cycles*, 9(4), 471–490. <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/95GB02432>.
- Wu, J., Albert, L.P., Lopes, A.P., Restrepo-Coupe, N., Hayek, M., Wiedemann, K.T. et al. (2016) Leaf development and demography explain photosynthetic seasonality in Amazon evergreen forests. *Science*, 351(6276), 972–976. <http://science.sciencemag.org/content/sci/351/6276/972.full.pdf>.

SUPPORTING INFORMATION

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

How to cite this article: Caen, A., Smallman, T.L., de Castro, A.A., Robertson, E., von Randow, C., Cardoso, M., Williams, M. (2021) Evaluating two land surface models for Brazil using a full carbon cycle benchmark with uncertainties. *Climate Resilience and Sustainability*. e10. <https://doi.org/10.1002/cli2.10>