Divergence in land surface modeling: linking spread to structure

Environmental Research Communications

LETTER

Divergence in land surface modeling: linking spread to structure

Christopher R Schwalm[1], Kevin Schaefer[2], Joshua B Fisher[3], Deborah Huntzinger[4], Yasin Elshorbany[5], Yuanyuan Fang[6], Daniel Hayes[7], Elchin Jafarov[8], Anna M Michalak[9], Mark Piper[10], Eric Stofferahn[11], Kang Wang[10] and Yaxing Wei[12]

[1] Woods Hole Research Center, Falmouth, Massachusetts 02540, United States of America
[2] National Snow and Ice Data Center, Boulder, Colorado 80309, United States of America
[3] Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, 91109, United States of America
[4] School of Earth and Sustainability, Northern Arizona University, Flagstaff, AZ 86011, United States of America
[5] Sustainability and Climate Change, College of Arts & Sciences, University of South Florida St. Petersburg, St. Petersburg, FL 33701, United States of America
[6] Department of Global Ecology, Carnegie Institution for Science, Stanford, CA 94305, United States of America
[7] School of Forest Resources, University of Maine, Orono, Maine 04469, United States of America
[8] Computational Earth Science, Los Alamos National Laboratory, Los Alamos, NM 87545, United States of America
[9] Department of Global Ecology, Carnegie Institution for Science, Stanford, CA 94305, United States of America
[10] CSDMS, Institute of Arctic and Alpine Research and Department of Geological Sciences, University of Colorado Boulder, Boulder, CO 80309, United States of America
[11] Conservation Science Partners, Truckee, CA 96161, United States of America
[12] Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, United States of America

E-mail: cschwalm@whrc.org

Keywords: global change ecology, carbon cycle modeling, data-driven discovery, inter-model spread

Supplementary material for this article is available online

OPEN ACCESS

RECEIVED
22 May 2019

REVISED
20 September 2019

ACCEPTED FOR PUBLICATION
3 October 2019

PUBLISHED
21 October 2019

Abstract

Divergence in land carbon cycle simulation is persistent and widespread. Regardless of model intercomparison project, results from individual models diverge significantly from each other and, in consequence, from reference datasets. Here we link model spread to structure using a 15-member ensemble of land surface models from the Multi-scale synthesis and Terrestrial Model Intercomparison Project (MsTMIP) as a test case. Our analysis uses functional benchmarks and model structure as predicted by model skill in a machine learning framework to isolate discrete aspects of model structure associated with divergence. We also quantify how initial conditions prejudice present-day model outcomes after centennial-scale transient simulations. Overall, the functional benchmark and machine learning exercises emphasize the importance of ecosystem structure in correctly simulating carbon and water cycling, highlight uncertainties in the structure of carbon pools, and advise against hard parametric limits on ecosystem function. We also find that initial conditions explain 90% of variation in global satellite-era values—initial conditions largely predetermine transient endpoints, historical environmental change notwithstanding. As MsTMIP prescribes forcing data and spin-up protocol, the range in initial conditions and high levels of predetermination are also structural. Our results suggest that methodological tools linking divergence to discrete aspects of model structure would complement current community best practices in model development.

1. Introduction

We define divergence as the spread in output from multiple models or, equivalently, the spread in the difference between model outputs and an observational constraint. Results from offline land surface simulations and fully-coupled Earth system models (ESMs) show persistent divergence in carbon cycling (e.g., Schwalm *et al* 2010, Fisher *et al* 2014, Friedlingstein *et al* 2014, Huntzinger *et al* 2017, Merryfield *et al* 2017, Giuntoli *et al* 2018). Furthermore, the added complexity of including more physical and biological processes in recent model

generations has not acted to reduce divergence or increase skill (Knutti and Sedláček, 2013, Wuebbles *et al* 2014). As models are, by definition, approximations of a set of physical and biogeochemical processes, inter-model spread must reflect choices made in this approximation. Such choices include which processes are represented (e.g., presence versus absence of carbon-nitrogen coupling; Huntzinger *et al* 2014), how they are coded mathematically (e.g., light use efficiency versus enzyme kinetics for photosynthesis; Wang *et al* 2011), and parameterizations used (Mendoza *et al* 2015). Divergence however also depends on spatiotemporal resolution (Schwalm *et al* 2010, 2013), forcing data such as precipitation (Samaniego *et al* 2017) and boundary conditions such as land cover history (Jain and Yang 2005). This complicates isolating useful approximations and therefore more correct model representations (Prentice *et al* 2015).

Reducing divergence across models and identifying appropriate representations are however highly desirable in Earth system modeling—both to ensure potential and realized predictability are commensurate (Luo *et al* 2015) and to improve the quality of predictions and projections under anticipated global environmental change. As model improvement assumes better agreement with observed values, resolving model divergence requires validation. That is, a model formulation is useful if it matches a reference set of observations within some tolerance (Luo *et al* 2012). Here we link model spread to model structure by moving beyond point-based benchmarking, e.g., calculating the distance between simulated and observed values. Instead, we apply three analytical approaches to link model-data mismatch to its source. First, we use functional benchmarks to help localize model subroutines that contribute to mismatch. Second, we use information on model structure to predict model skill in a machine learning framework. Third, we quantify how initial conditions prejudice model outcomes after centennial-scale transient simulations.

We demonstrate these three approaches using the Multi-scale synthesis and Terrestrial Model Intercomparison Project (MsTMIP; Huntzinger *et al* 2013, 2017), a 15-member model ensemble of standardized simulations and their outputs, as our analysis test bed. MsTMIP is focused on carbon and water cycling in land surface models—the land component of ESMs—and is based on a constrained model protocol that prescribes spatiotemporal resolution, forcing data, boundary conditions, and spin-up procedures. In additional, divergence seen in previous MIPs (model-intercomparison projects) is present in MsTMIP. For example, global mean satellite-era gross primary productivity (GPP) varies 2-fold from 91 to 185 PgC per annum across the ensemble relative to a benchmark value based on upscaled FLUXNET data of 117 PgC per annum. Using individual eddy covariance towers and corresponding model grid cells reveals a similar 2-fold range (2.2 to 4.4 $gC/m^2/d$) in GPP, relative to the benchmark value of 3.3 $gC/m^2/d$. Metrics of ecosystem structure vary wider still—global mean satellite-era leaf area index (LAI) ranges from 1.4 to 4.1 $m^2\ m^{-2}$ relative to a AVHRR benchmark value of 1.5 $m^2\ m^{-2}$. Overall, the MsMTIP ensemble provides the correct 'model space' to link skill to structure as it excludes confounding factors while preserving inter-model spread.

2. Methods

2.1. Benchmarking

Even though our emphasis herein is on addressing why models diverge we still need to quantity model-data mismatch to calculate divergence, model skill, and functional benchmarks. Here we use ILAMB (International Land Model Benchmarking) (https://ilamb.ornl.gov/doc/; Collier *et al* 2018), a generic benchmarking framework based on a series of python widgets that allows for the standardized comparison of simulation output and reference datasets. ILAMB can also calculate functional benchmarks relating one variable to another such as GPP as a function of LAI. For this study we use the Permafrost Benchmark System (PBS) version of ILAMB (https://permamodel.github.io/pbs). This version is hosted on the Community Surface Dynamics Modeling System (CSDMS; https://csdms.colorado.edu) and removes the need for individual modelers to install ILAMB locally. Instead, through an ingest tool, simulation output is uploaded to a server host and ILAMB is executed server-side on a high-performance computing cluster controlled through a simple web interface.

2.2. Reference datasets

ILAMB contains a default set of monthly, point-based and gridded reference datasets. In this study, composite skill—used as the target variable in the skill-to-structure mapping exercise—is calculated globally for (1) four fluxes (evapotranspiration [ET] or latent heat, GPP, TER [total ecosystem respiration], and NEP [net ecosystem productivity]) using upscaled and tower-based (model grid cells containing the flux tower are used in the intercomparison) FLUXNET data, (2) LAI from AVHRR and MODIS satellite data, and (3) four mapped biomass reference products (Collier *et al* 2018). Composite skill for each variable ranges from zero (no agreement) to unity (perfect agreement with all reference datasets) and is based on a weighted combination across all skill metrics: correlation, bias, root mean square error, and phase shift (Collier *et al* 2018). For this study the default configuration of weights and score metrics is used.

Table 1. Summary of MsTMIP simulation experiments. Simulation codes reference enabling of historical time-varying climate (SG1), CO_2 concentration (SG2), land cover/land use (SG3) and nitrogen subsidy (BG1). These are sequentially enabled after steady state (RG1) is reached. Capital letters indicate runs used for benchmarking.

Model name	Simulation availability				
	RG1	SG1	SG2	SG3	BG1
BIOME-BGC	y	y			Y
CLASS-CTEM-N	y	y	y	y	Y
CLM	y	y	y	y	Y
CLM4VIC	y	y	y	y	Y
DLEM	y	y	y	y	Y
GTEC	y	y	y	Y	
ISAM	y	y	y	y	Y
LPJ-wsl	y	y	y	Y	
ORCHIDEE-LSCE	y	y	y	Y	
SiB3	y	y	y	Y	
SiBCASA	y	y	y	Y	
TEM6	y	y	y	y	Y
TRIPLEX-GHG	y		y	y	Y
VEGAS2.1	y	y	y	y	
VISIT	y	y	y	y	

2.3. Simulation results

Simulation output is taken from Version 1 of MsTMIP—the Multi-scale synthesis and Terrestrial Model Intercomparison Project (Huntzinger *et al* 2018; data portal: https://doi.org/10.3334/ORNLDAAC/1225). MsTMIP is a 15-member model ensemble that uses a standardized simulation protocol—historical forcing data, boundary conditions, and spin-up procedures are uniform across all models (Huntzinger et al, 2013, Wei *et al* 2014)—to isolate structural differences (table 1). MsTMIP runs are global (0.5° spatial resolution), monthly from 1901 to 2010 and use a semi-factorial set of simulations where historical time-varying climate, CO_2 concentration, land cover/land use, and nitrogen subsidy are sequentially enabled after steady state is reached. For the subset of models with nitrogen cycling (BIOME-BGC, CLASS-CTEM-N, CLM, CLM4VIC, DLEM, ISAM, TEM6, and TRIPLEX-GHG) MsTMIP Version 1 output based on all time-varying factors (simulation BG1) is used. Otherwise, MsTMIP Version 1 model output based on time-varying climate, CO_2 concentration and land cover/land use only (simulation SG3) is used (GTEC, LPJ-wsl, ORCHIDEE-LSCE, SiB3, SiBCASA, VEGAS2.1, and VISIT). Note that not all models simulate all variables (table 2). As an example, SiB3 lacks carbon pools and only 6 models (CLASS-CTEM-N, CLM4, CLM4VIC, ISAM, LPJ-wsl, and SiB3) simulate snow depth.

2.4. Skill-to-structure mapping

Linking skill to a discrete aspect of model structure is a data-driven exercise. Initially, model structure (Huntzinger *et al* 2014) is encoded as a set of indicator variables. These variables span all aspects of model structure and are grouped into four broad themes: carbon cycling, energy exchange, nitrogen cycling, and vegetation dynamics (supplementary tables 1–4). As an example, we use a vegetation dynamics indicator variable based on the presence or absence of a maximum value of LAI beyond which there is no allocation of biomass to leaves. As a second step, ILAMB is used to determine composite skill by variable (Collier *et al* 2018). For this exercise we use GPP, ET and LAI as in the functional benchmarking exercise as well as TER, NEP and total live biomass. Lastly, we predict skill using only structure. Here, the Random Forests algorithm is used to generate 10 000 individual decision trees for each MsTMIP output variable ($n = 6$) separately to simultaneously predict composite skill for all models using the same set of presence/absence indicator variables from the full MsTMIP model ensemble. The variance explained ranges from 69% to 96%: NEP, 69%; GPP, 78%; LAI, 79%; TER, 81%; ET, 94% and biomass, 96%, i.e., the structural determinants of skill are well-captured. We also calculate the gain in skill for each structural indicator variable in the topmost position—the initial splitting variable—across all 10 000 decision trees for each MsTMIP variable separately. Gain quantifies skill improvement based on structural choice. For this we first navigate each decision tree from top to bottom to maximize composite skill at each decision point through to the terminal node. Second, we calculate the difference in skill—the initial composite skill based on using the topmost structural indicator variable only is subtracted from the composite skill from the terminal node—across all 10 000 decision trees. Gain is then the mean skill difference across all decision trees by MsTMIP variable, is always positive, and is expressed in units of standard deviation of skill across the MsTMIP ensemble; analogous to a z-score transformation. Before calculating gain, variable

Table 2. Influence of initial conditions on transient simulation endpoints by variable. Correlation is based on global integrals between initial conditions (see Methods) and the 1981-2010 satellite-era mean. Only vegetated land pixels are used. For snow variables only those pixels with seasonal snow cover are used. There are 15 models in the full MsMTIP Version 1 ensemble (table 1).

Variable	Number of models	Correlation
Autotrophic respiration	14	0.95
Evapotranspiration	14	0.35
Fire flux	5	0.98
Gross primary productivity	15	0.87
Heterotrophic respiration	14	0.95
Leaf area index	9	0.95
Net ecosystem productivity	15	0.88
Net primary productivity	13	0.93
Snow depth	6	0.99
Snow water equivalent	6	0.99
Soil carbon	14	0.98
Soil temperature	7	0.99
Soil wetness	4	0.99
Surface runoff	11	0.92
Total ecosystem respiration	15	0.87
Total live biomass	13	0.96

importance is used to select relevant model structural attributes. Variable importance quantifies how much predictive power each predictor variable has, i.e., serves as an indicator for the overall impact of a predictor on composite skill. The permutation variable importance measure used herein quantifies the loss in skill, the algorithm's ability to predict composite score based on structural indicator variables, by randomly permuting the values of a single predictor variable and comparing that to the unpermuted version. We use this metric to filter structural indicator variables not useful in understanding how structure impacts composite skill, i.e., those with negative values are excluded from further analysis (Janitza *et al* 2018). Using variable importance scores as a filter and then calculating gain for topmost splitting variables allows us to identify which structural attributes are most relevant to divergence and to quantify this dependence.

2.5. Initial conditions

To further explore divergence we examine initial conditions ($X_{intital}$) relative to satellite-era conditions ($X_{transient}$), where X refers to a given variable and the subscripts denote initial conditions and those after 110 years of transient forcings, respectively. $X_{intital}$ is taken from the first 30 years (nominally 1901-1930) of the MsTMIP control run (simulation RG1). This is an extension of the steady state run based on randomized meteorological forcing (Huntzinger *et al* 2013, Wei *et al* 2014) and represents initial conditions after steady-state is reached but before transient forcings are applied. In contrast, $X_{transient}$ is taken from the last 30 years (1981-2010, the satellite era) of transient run SG3 or, for models with nitrogen cycling, BG1 (table 1). All variables output by at least 4 MsTMIP models are included (table 2).

3. Results

3.1. Functional benchmarking

For functional benchmarking we highlight GPP, LAI, and latent heat. GPP represents the dominant input of carbon into the terrestrial carbon cycle. If a model does not correctly simulate GPP, it cannot correctly simulate biomass and respiration. All models in this study simulate GPP for a single leaf and scale this to the entire canopy using LAI, defined as the ratio of leaf area to ground area. For GPP as a function of LAI (figure 1), reference data supports a 5 m^2/m^2 upper limit for LAI while modeled values reach 15 $m^2\ m^{-2}$. Both models and observations show a near linear response. All models however underestimate GPP as a function of LAI, i.e., the benchmark functional response serves as an upper limit for the models. This is supported by the ratio of GPP to LAI (ratio of totals) of 1.8 $gC/m^2/d$ for observations relative to the lower modeled values of 0.6 to 1.7 $gC/m^2/d$ (−59 to −2% difference with a mean difference of −31%). This could result from an underestimation in either simulated stomatal conductance or leaf-to-canopy scaling. Both GPP and latent heat flux depend very strongly on simulated stomatal conductance. Looking further, we see that the models—similar to GPP versus LAI—also underestimate latent heat as a function of LAI (differences range from −66 to 0% with a mean of −27%). In

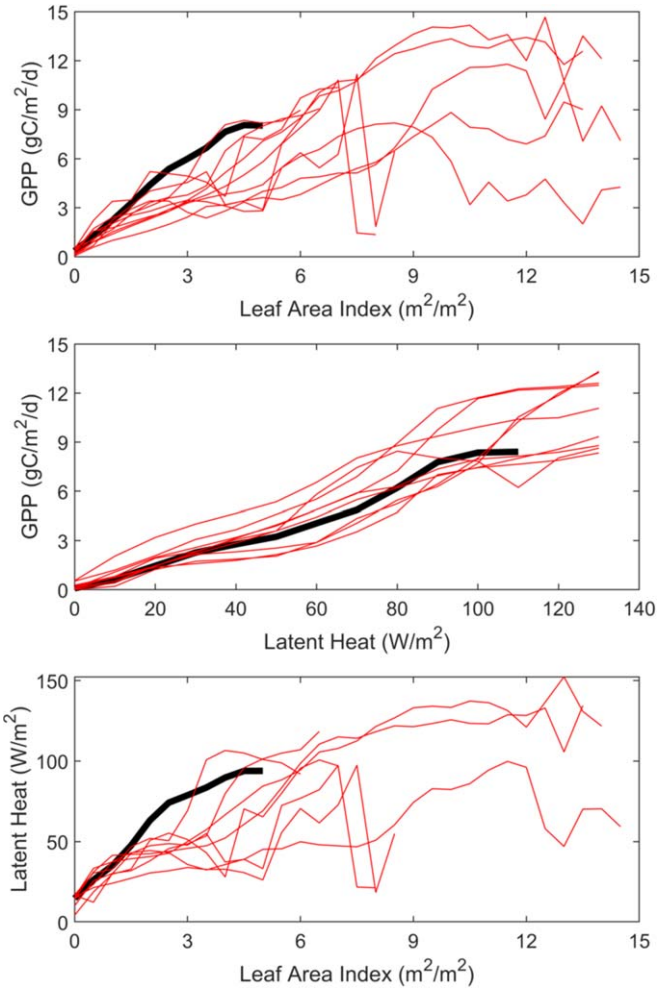

Figure 1. Relationship between GPP, LAI and latent heat from functional benchmarking. Panels show functional relationships between GPP and LAI (top), GPP and latent heat (middle), and latent heat and LAI (bottom). Red lines show individual models. Thick black line is based on covariation between reference datasets; upscaled FLUXNET for GPP and latent heat as well as AVHRR-based LAI. Data values are binned satellite-era long-term averages by pixel, using the *x*-axis, across global vegetated land surface at 0.5 m^2 m^{-2} increments for LAI and at 10 W m^{-2} increments for latent heat. Where red lines continue past the black reference line indicates aphysical simulated values.

contrast, the models all roughly get the right ratio of GPP to latent heat flux; the percent difference ranges from −74 to +26% with 7 of 9 models between −26 and +20% and a mean difference of +13%. This suggests that the models are better able to simulate stomatal conductance, but underestimate leaf-to-canopy scaling and overestimate LAI.

3.2. Model structure to predict model skill

Using the Random Forests machine learning algorithm, we find several model structural characteristics that serve as important controls of skill for multiple MsTMIP variables. For instance, the absence of a carbohydrate reserve pool is associated with higher skill for GPP, LAI, and TER (figure 2) while yielding an average skill gain of 0.75σ, where σ is the standard deviation of skill across the full ensemble. In contrast, while the presence of a below ground litter pool is associated with above average skill for LAI, the absence of this same structural characteristic is linked to skill gains for NEP. A similar pattern is also present for the below ground sapwood carbon pool. More broadly, the analysis is does not reveal an optimal structure of carbon pools bur rather suggests a trade-off between carbon pool structure—or allocation heuristics—and skill by variable within a given model.

A second general tendency is that the absence of thresholds is associated with higher skill. As an example, ET skill increases by 0.88σ when not considering nitrogen limitation on GPP. Similarly, higher skill in GPP and TER is linked to not having autotrophic respiration limited by nitrogen availability. Finally, the absence of a non-zero threshold light level for GPP (figure 2)—GPP may occur at the lowest levels of insolation as opposed to a parameterized threshold—is associated with higher skill for both GPP and TER (0.65σ). The otherwise largest gains in skill are achieved by (1) not using whole canopy stomatal conductance parameters to better simulate LAI

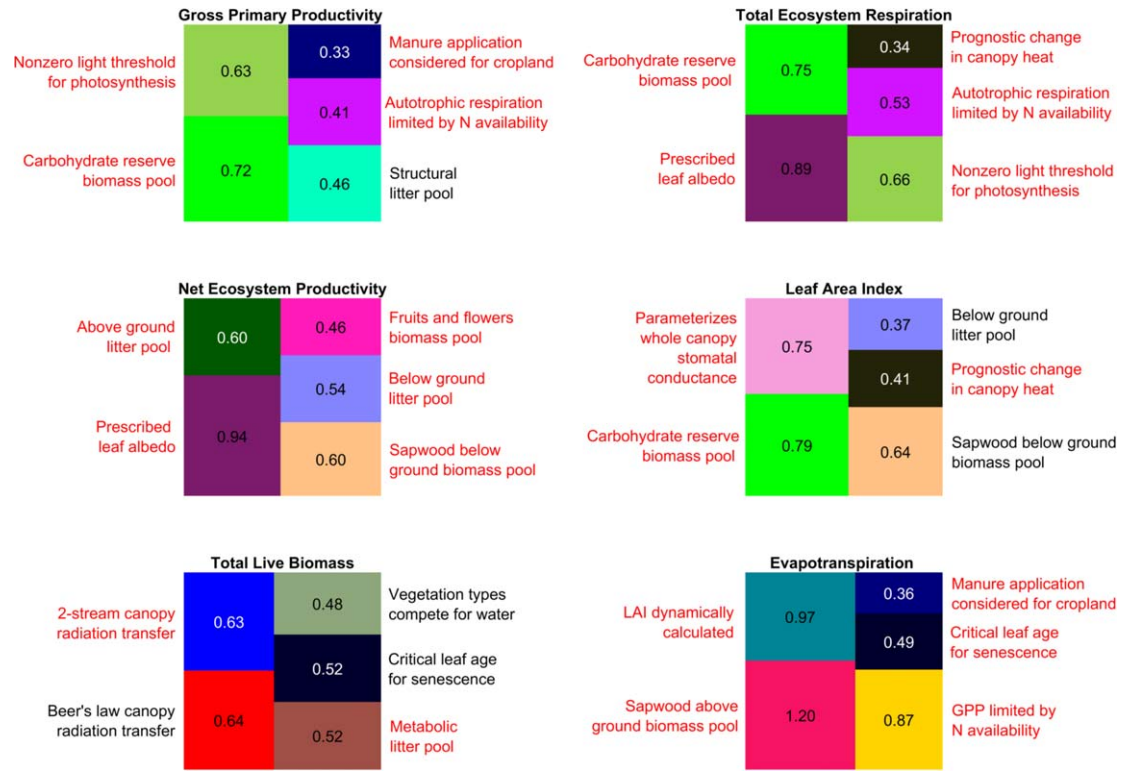

Figure 2. Relevance of model structural characteristics to skill. Treemap shows skill gain (change in skill in units of standard deviation of skill across the ensemble, inset number) by structural attribute (text labels) for each variable. Rectangle colors match attributes, i.e., Critical leaf age for senescence is dark blue in all treemaps. Text label color denotes presence (black) or absence (red) of the relevant model structural attribute.

(0.75σ), (2) not having LAI dynamically calculated for higher skill in ET (0.97σ), and (3) not prescribing leaf albedo for TER and NEP (0.89σ and 0.94σ respectively). These recommendations all highlight the skill-relevance of correctly simulating LAI and leaf-to-canopy scaling as revealed by functional relationships.

Overall, the skill-to-structure mapping exercise emphasizes the importance of ecosystem structure in correctly simulating carbon and water cycling, highlights uncertainties in the structure of carbon pools, and advises against hard parametric limits on ecosystem function. This latter point does not suggest that hard limits, such as cell denaturation at critical temperatures, are wrong, merely that constraining the shape of an idealized response function limits skill and/or that the parameters controlling these hinges are misspecified or unnecessarily invariant across enviroclimatic space (Mendoza *et al* 2014).

3.3. Initial conditions as conditional endpoints

Beyond functional benchmarking and machine learning we find that divergence is embedded *a priori* in all MsTMIP variables due to initial conditions. As an example, soil carbon is known to be particularly 'sticky' and predetermined by initial pool size, i.e., levels of soil carbon before transient forcings persist with minimal change (Exbrayat *et al* 2014; Todd-Brown *et al* 2013). Across MsTMIP (figure 3) initial conditions explain, in a least-squares sense, 90% (median value; table 2) of variation in transient endpoints. Alternatively, 110-years of MsTMIP transient forcings (i.e., dynamic climate, land cover, CO_2 fertilization and nitrogen subsidy; see Methods) explain but one-tenth of variance in the Earth system globally whereas nine-tenths is based on initial conditions.

This prejudice is not solely limited to global mean values but is also broadly uniformly present across the land surface at pixel scale as well. Spatially, initial conditions predetermine transient endpoints to the largest degree in the Northern Hemisphere (80% variance explained on average by pixel; figure 4, supplementary figure 1 is available online at stacks.iop.org/ERC/1/111004/mmedia). There is a weak tendency, especially for ET, for marginal productivity areas to show the lowest level of predetermination. Despite this, the proportion of vegetated land area where initial conditions predetermine transient endpoints by 50% or greater (inset values; figure 4) is at least 0.43 (for ET) and rises to 0.91 (soil carbon). Variability (quantified as standard deviation; figure 4 and supplementary figure 2) and extreme behavior (quantified as the 95th%ile; figure 4 and supplementary figure 3) show similar degrees of prejudice and spatial patterns. Here, the proportion of vegetated land area where initial conditions predetermine transient endpoints by 50% or greater is at least 0.47 for

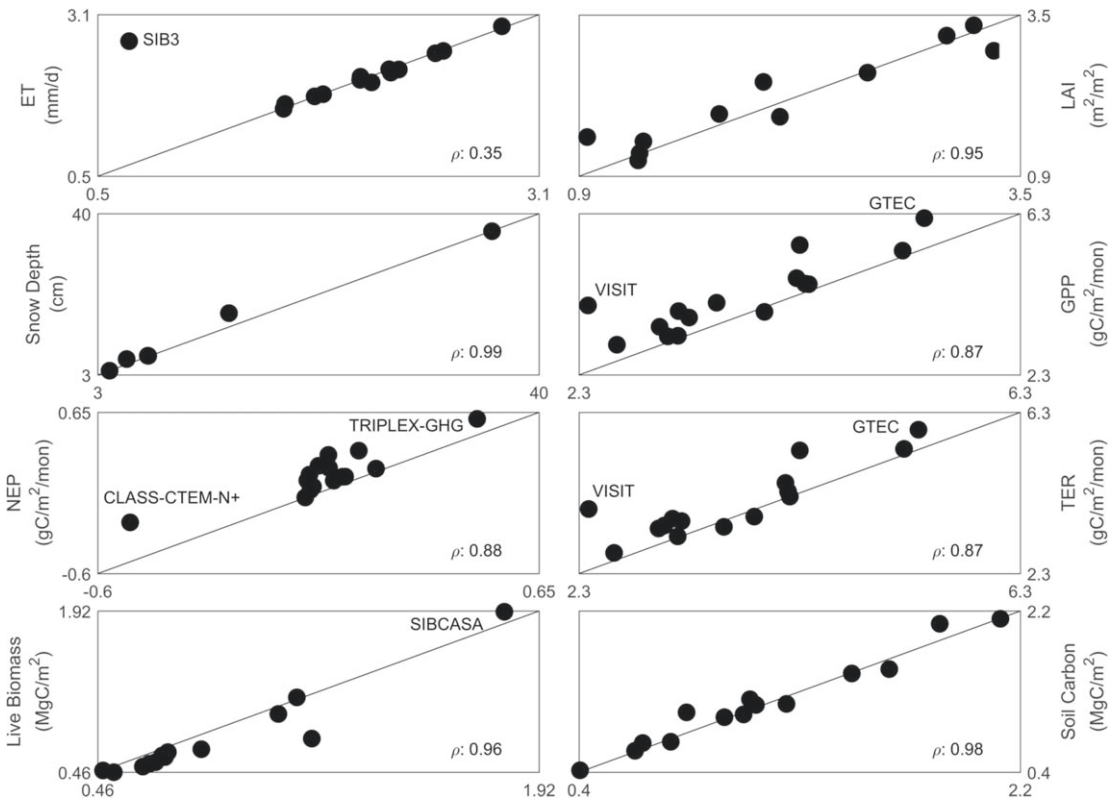

Figure 3. Influence of initial conditions on transient simulation endpoints. Each panel shows, for a given variable (row-wise labels), satellite-era transient simulation (y-axis) versus initial condition (x-axis) global aggregate across all models. Individual models (circles; outlying models are labeled), 1:1 reference (line), and correlation (inset values) shown for each variable. Correlation excluding labeled outliers is 0.99 for ET and 0.56 for NEP. Variance explained, in a least-squares sense, is the square of correlation and ranges from 0.1% for ET with the SiB3 outlier to 96% for soil carbon.

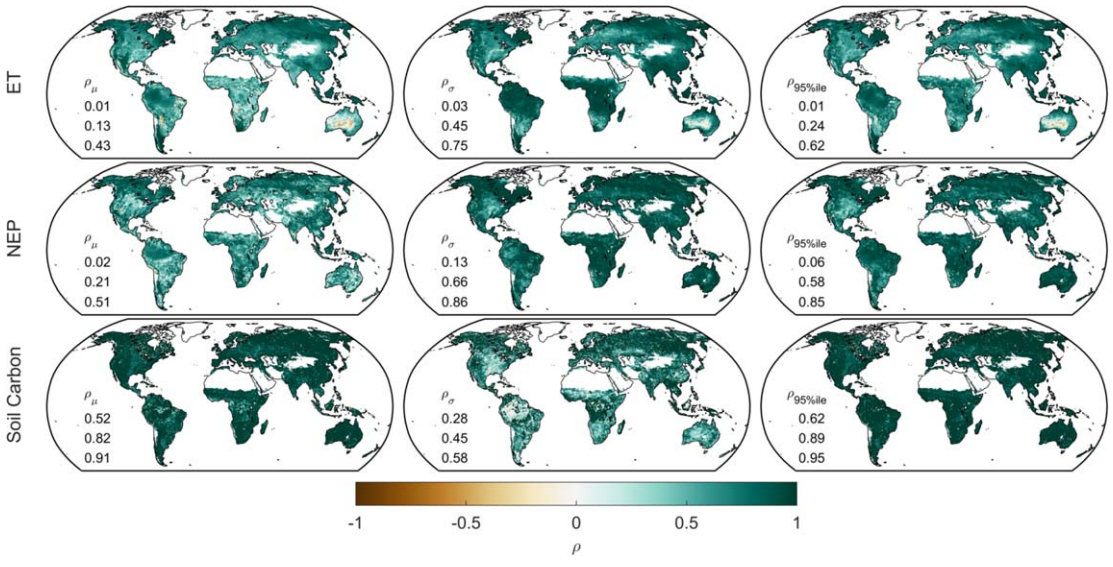

Figure 4. Relationship between initial conditions and transient simulation endpoints. Panels show pixelwise linear correlation of initial condition and satellite-era mean (ρ_{μ}, left column), standard deviation (ρ_{σ}, middle column), and 95th%ile ($\rho_{95th\%ile}$, right column) values across all models by variable (rows). Vegetated land surface only. Inset values give proportion of land surface where at least 95%, 75%, and 50% variance (from top to bottom respectively) in satellite-era values is explained by initial conditions. ET, NEP and soil carbon are a representative sample of all MsTMIP variables (see table 2 and supplementary information).

variability (biomass) and 0.62 for extreme behavior (ET). Maximum values are 0.86 for variability (GPP) and 0.95 for extreme behavior (soil carbon). This occurs over manifold variations in initial conditions, e.g., initial condition ET varies more than 4-fold from 0.68 to 2.99 mm/d. Snow depth and net ecosystem productivity

(NEP) show 12-fold variation (figure 3). Across all variables, the departure from the 1:1 line for initial and satellite-era values is small regardless of the initial value which effectively serves to anchors transient run endpoints.

4. Discussion

Carbon cycle modeling continues to evolve, as does model evaluation and benchmarking (Collier *et al* 2018, Haughton *et al* 2018, Luo *et al* 2012, Medlyn *et al* 2015, Prentice *et al* 2015). Missing from these developments, however, is a set of tools that allow us to address why models diverge by mapping model skill to model structure. In other words, as a community, carbon cycling modeling needs a mechanism to attribute poor skill to identifiable aspects of model structure. Our study suggests a three-pronged approach to achieve this goal:

(1) Emphasize functional benchmarking over comparing individual variables to reference values. Simulated point estimates are invariably different from any observational point estimate. Incorporating uncertainty in the comparison is useful as the best any model can do is to match observations within uncertainty (Schaefer *et al* 2012, Schwalm *et al* 2010, 2015). This also points to the limit of benchmarking because we cannot improve models until we improve observations. However, using confidence bounds does not provide a pathway to assess structure per se. As an example, if GPP at some level of spatiotemporal aggregation is 20% larger than observed (and this 20% is outside the uncertainty envelope of the observation) there is no insight into which model choice produced the mismatch. In the past, we've focused on adding new processes to improve skill, made practical by advances in computational frameworks and improvements in our knowledge base (Maslin and Austin, 2012, Stockmann *et al* 2013, Bailey *et al* 2018). However, the added complexity of including more physical and biological processes does not equate to reduced divergence or increased skill (Knutti and Sedláček, 2013, Wuebbles *et al* 2014). Functional benchmarking allows one to identify a specific process independent of model complexity. For MsTMIP, looking at GPP, LAI, and latent heat in isolation does not highlight leaf-to-canopy scaling.

(2) Encode metadata on model structure to allow data analytics. Apart from MsTMIP no large-scale MIP includes, as published metadata with a controlled vocabulary, a systemic survey or database of model structural characteristics. The MsTMIP case nonetheless offers scope for improvement as the structural metadata does not capture all information about a given model and, subsequently, the full ensemble undersamples the range in process representations (cf Annan *et al* 2011). A 15-model ensemble, as in MsTMIP, allows for 2^{15} or 32,768 unique process representations using presence/absence coding. In this study, only 135 structural variables are inventoried which are then truncated to 69 (only 0.2% of the theoretical potential) after semantic duplicates are removed. Any extension in structure encoding need not be limited to presence/absence but should also include increases in ensemble size to better sample model structural space overall, ordered categorical variables (e.g., to traverse a complexity gradient of radiation transfer schemes) as well as a range in parameter values (e.g., to link within-parameter uncertainty to divergence, cf Zaehle and Friend 2010). Here it's important to note that MsTMIP structural attributes highlighted through functional benchmarking have key parameters, with emphasis on V_{cmax} (unstressed Rubisco catalytic capacity) or J_{max} (the maximum electron transport rate) for limits on GPP, that may potentially compensate leaf-to-canopy scaling (Schaefer *et al* 2012). More broadly and regardless of MIP size, model structural characteristics must be encoded in a form amenable to machine learning and traceability (Zhou *et al* 2018). This allows process representations to be linked to gradients of skill and thus provides a mechanism to isolate 'winners' from 'losers'. Process representations that are repeatedly associated with below-average skill levels across multiple MIPs merit consideration for possible deletion from the catalogue of model structural choice. Our results highlight the need for careful curation of metadata on model structure and ensemble broadness to maximize the discriminatory power of our data-driven approach. Model structure encoding combined with data analytics offers a heretofore underutilized approach to discriminate among thousands of model choice decisions and thus improve model reliability (Prentice *et al* 2015).

(3) Acknowledge and resolve how initial conditions prejudice transient endpoints. Initial conditions explain (median value; table 2) 90% of transient endpoint values. This is despite nontrivial changes in MsTMIP transient forcings over the 110-year simulation period. From 1901 to 2010 air temperature increases from 12.7 °C to 13.8 °C (1901 and 2001 decade global means respectively) and the global mean concentration of CO_2 increases monotonically from 295 ppm (1910) to 388 ppm (2010). Nitrogen subsidy increases by a factor of five; land cover and land use changes result in a 17% net loss in forest cover, a five-fold increase in grasslands, and a doubling of cropland extent (Wei *et al* 2014). The changes in forcing data over the

simulation period do not however translate into large departures from initial conditions across MsTMIP simulation outputs. Superimposed on this is a large range in initial conditions themselves (figure 3), which are in turn solely attributable to structural differences due to the MsTMIP protocol that constrains forcing data, boundary conditions, and steady-state spin-up protocols.

The assumption of steady state results from a lack of knowledge on ecosystem state, especially for carbon pools (Carvalhais *et al* 2010). The most common approach is to assume human impacts in the preindustrial era (typically before 1750) were nonexistent and thus a steady-state equilibrium condition was the norm. Thus, models are fed thousands of trend-free randomized blocks of forcing data until the net change in, at least, carbon pools is zero over an arbitrary time period within some tolerance. When steady state is achieved the corresponding values for all simulated quantities form the catalogue of initial conditions. While mathematically tractable—and historically computational expensive (Xia *et al* 2012)—the basis for this assumption is inaccurate (Ruddiman 2003, 2007, Kaplan *et al* 2011, Lewis and Maslin, 2015, Ruddiman *et al* 2015). In the preindustrial Holocene alone humans caused a 9 ppm CO_2 increase (Ruddiman 2007). This is 10% of the change in CO_2 seen in MsTMIP from 1901 to 2010 but is at odds with equilibrium conditions prior to industrialization or the existence of a pre-anthropogenic baseline in the mid to late Holocene (Ruddiman, 2007) that underpins the steady-state modeling assumption. Furthermore, there is evidence that skill improves in the absence of steady state (Carvalhais *et al* 2008, 2010, Hashimoto *et al* 2011). Recent developments in semi-analytical solutions for steady state (Huang *et al* 2018 Xia *et al* 2012, Luo *et al* 2017) treat transfers between carbon and nitrogen pools in a single unified matrix solution and greatly reduce spin-up time. This suggests a land carbon MIP centered on steady state, something missing from the landscape of existent and planned MIPs, is both needed and achievable. Such a MIP must include simulation experiments that include varying degrees of steady state relaxation (e.g., Carvalhais *et al* 2008, 2010), random initial or seed values (e.g., Hashimoto *et al* 2011) of carbon pools, and a gradient for the number of grid cells or sites that are required to achieve some threshold (usually no change over an arbitrary time period). Only such an effort can determine how steady state itself leads to divergence.

While these tools provide a means to link divergence to discrete aspects of model code they are not perfect. Using model skill conditions on the truthfulness of the reference datasets. Similarly, the linkage between skill and structure is data-driven and not aware of mechanistic interdependencies in a given model. Lastly, the predetermination of endpoints through initial states requires follow-on research, our analysis is necessarily descriptive. Nonetheless, the persistence of divergence in carbon cycle simulation over several generations of models suggests an urgent need for additional methodological approaches, as described herein, to inform model development.

Acknowledgments

CRS, MP and KS were supported by National Aeronautics and Space Administration (NASA) award 14-CMAC14-NNX16AB19G, A Permafrost Benchmark System to Evaluate Permafrost Models. CRS was also supported by NASA grants N4-TE14-0047-NNH14ZDA001N-TE and 13-CARBON13_2-0036-NNH13ZDA001N-CARBON. KS and KW received support from National Science Foundation (NSF) grant NSF-OPP 1503559. EJ was funded from the Next-Generation Ecosystem Experiments Arctic project, DOE Office of Science. YE acknowledges research support from NSF project 1900795. JBF was supported in part by NASA programs: TE, IDS, CMS, and CARBON.

ORCID iDs

Christopher R Schwalm https://orcid.org/0000-0002-5035-5681
Elchin Jafarov https://orcid.org/0000-0002-8310-3261
Eric Stofferahn https://orcid.org/0000-0001-6960-4193
Kang Wang https://orcid.org/0000-0003-3416-572X

References

Annan J D, Hargreaves J C and Tachiiri K 2011 On the observational assessment of climate model performance *Geophys. Res. Lett.* **38** L24702
Bailey V L, Bond-Lamberty B, DeAngelis K, Grandy A S, Hawkes C V, Heckman K and Wallenstein M D 2018 Soil carbon cycling proxies: understanding their critical role in predicting climate change feedbacks *Global Change Biol.* **24** 895–905
Carvalhais N, Reichstein M, Ciais P, Collatz G J, Mahecha M D, Montagnani L and Seixas J 2010 Identification of vegetation and soil carbon pools out of equilibrium in a process model via eddy covariance and biometric constraints *Global Change Biol.* **16** 2813–29
Carvalhais N, Reichstein M, Seixas J, Collatz G J, Pereira J S, Berbigier P and Rambal S 2008 Implications of the carbon cycle steady state assumption for biogeochemical modeling performance and inverse parameter retrieval *Global Biogeochem. Cycles* **22** GB2007

Collier N, Hoffman F M, Lawrence D M, Keppel-Aleks G, Koven C D, Riley W J and Randerson J T 2018 The international land model benchmarking (ILAMB) system: design, theory, and implementation *Journal of Advances in Modeling Earth Systems.* **10** 2731–54
Exbrayat J F, Pitman A J and Abramowitz G 2014 Response of microbial decomposition to spin-up explains CMIP5 soil carbon range until 2100 *Geoscientific Model Development* **7** 2683–92
Fisher J B, Sikka M, Oechel W C, Huntzinger D N, Melton J R, Koven C D and Ciais P 2014 Carbon cycle uncertainty in the Alaskan Arctic *Biogeosciences* **11** 4271–88
Friedlingstein P, Meinshausen M, Arora V K, Jones C D, Anav A, Liddicoat S K and Knutti R 2014 Uncertainties in CMIP5 climate projections due to carbon cycle feedbacks *J. Clim.* **27** 511–26
Giuntoli I, Villarini G, Prudhomme C and Hannah D M 2018 Uncertainties in projected runoff over the conterminous United States *Clim. Change* 1–14
Hashimoto S, Wattenbach M and Smith P 2011 A new scheme for initializing process-based ecosystem models by scaling soil carbon pools *Ecol. Modell.* **222** 3598–602
Haughton N, Abramowitz G and Pitman A J 2018 On the predictability of land surface fluxes from meteorological variables *Geoscientific Model Development* **11** 195–212
Huang Y, Lu X, Shi Z, Lawrence D, Koven C D, Xia J and Luo Y 2018 Matrix approach to land carbon cycle modeling: a case study with the community land model *Global Change Biol.* **24** 1394–404
Huntzinger D N, Michalak A M, Schwalm C, Ciais P, King A W, Fang Y and Hayes D 2017 Uncertainty in the response of terrestrial carbon sink to environmental drivers undermines carbon-climate feedback predictions *Sci. Rep.* **7** 4765
Huntzinger D N, Schwalm C, Michalak A M, Schaefer K, King A W, Wei Y and Berthier G 2013 The North American carbon program multi-scale synthesis and terrestrial model intercomparison project: I. Overview and experimental design *Geoscientific Model Development* **6** 2121–33
Huntzinger D N, Schwalm C, Michalak A M, Schaefer K, Wei Y, Cook R B and Jacobson A 2014 NACP MsTMIP summary of model structure and characteristics *Oak Ridge National Laboratory Distributed Active Archive Center for Biogeochemical Dynamics* (https://doi.org/10.3334/ornldaac/1228)
Huntzinger D N *et al* 2018 *NACP MsTMIP: global 0.5-degree model outputs in standard format, Version 1.0.* (Tennessee, USA: ORNL DAAC, Oak Ridge) https://doi.org/10.3334/ORNLDAAC/1225
Jain A K and Yang X 2005 Modeling the effects of two different land cover change data sets on the carbon stocks of plants and soils in concert with CO_2 and climate change *Global Biogeochem. Cycles* **19**
Janitza S, Celik E and Boulesteix A L 2018 A computationally fast variable importance test for random forests for high-dimensional data *Advances in Data Analysis and Classification* **12** 885–915
Kaplan J O, Krumhardt K M, Ellis E C, Ruddiman W F, Lemmen C and Goldewijk K K 2011 Holocene carbon emissions as a result of anthropogenic land cover change *The Holocene* **21** 775–91
Knutti R and Sedláček J 2013 Robustness and uncertainties in the new CMIP5 climate model projections *Nat. Clim. Change* **3** 369–73
Lewis S L and Maslin M A 2015 Defining the anthropocene *Nature* **519** 171
Luo Y, Keenan T F and Smith M 2015 Predictability of the terrestrial carbon cycle *Global Change Biol.* **21** 1737–51
Luo Y, Shi Z, Lu X, Xia J, Liang J, Wang Y and Hararuk O 2017 Transient dynamics of terrestrial carbon storage: mathematical foundation and numeric examples *Biogeosciences* **14** 145–61
Luo Y Q, Randerson J T, Abramowitz G, Bacour C, Blyth E, Carvalhais N and Friedlingstein P 2012 A framework for benchmarking land models. *Biogeosciences* **9** 3857–74
Luo Y Q, Randerson J T, Friedlingstein P, Hibbard K, Hoffman F, Huntzinger D and Mahecha M 2012 A framework for benchmarking land models.
Maslin M and Austin P 2012 Uncertainty: climate models at their limit? *Nature* **486** 183
Medlyn B E, Zaehle S, De Kauwe M G, Walker A P, Dietze M C, Hanson P J and Prentice I C 2015 Using ecosystem experiments to improve vegetation models *Nat. Clim. Change* **5** 528
Mendoza P A, Clark M P, Barlage M, Rajagopalan B, Samaniego L, Abramowitz G and Gupta H 2015 Are we unnecessarily constraining the agility of complex process-based models? *Water Resour. Res.* **51** 716–28
Merryfield W J, Doblas-Reyes F J, Ferranti L, Jeong J H, Orsolini Y J, Saurral R I and Rixen M 2017 Advancing climate forecasting *Eos* **98**
Prentice I C, Liang X, Medlyn B E and Wang Y P 2015 Reliable, robust and realistic: the three R's of next-generation land-surface modelling *Atmos. Chem. Phys.* **15** 5987–6005
Ruddiman W F 2003 The anthropogenic greenhouse era began thousands of years ago *Clim. Change* **61** 261–93
Ruddiman W F 2007 The early anthropogenic hypothesis: challenges and responses *Rev. Geophys.* **45** RG4001
Ruddiman W F, Ellis E C, Kaplan J O and Fuller D Q 2015 Defining the epoch we live in *Science* **348** 38–9
Samaniego L, Kumar R, Breuer L, Chamorro A, Flörke M, Pechlivanidis I G and Zeng X 2017 Propagation of forcing and model uncertainties on to hydrological drought characteristics in a multi-model century-long experiment in large river basins *Clim. Change* **141** 435–49
Schaefer K, Schwalm C R, Williams C, Arain M A, Barr A, Chen J M and Humphreys E 2012 A model-data comparison of gross primary productivity: results from the North American Carbon Program site synthesis *Journal of Geophysical Research: Biogeosciences* **117** G03010
Schwalm C R, Huntinzger D N, Michalak A M, Fisher J B, Kimball J S, Mueller B and Zhang Y 2013 Sensitivity of inferred climate model skill to evaluation decisions: a case study using CMIP5 evapotranspiration *Environ. Res. Lett.* **8** 024028
Schwalm C R, Huntzinger D N, Fisher J B, Michalak A M, Bowman K, Ciais P and Ito A 2015 Toward 'optimal' integration of terrestrial biosphere models *Geophys. Res. Lett.* **42** 4418–28
Schwalm C R, Williams C A, Schaefer K, Anderson R, Arain M A, Baker I and Ciais P 2010 A model-data intercomparison of CO_2 exchange across North America: results from the North American carbon program site synthesis *J. Geophys. Res* **115** G00H05
Stockmann U, Adams M A, Crawford J W, Field D J, Henakaarchchi N, Jenkins M and Wheeler I 2013 The knowns, known unknowns and unknowns of sequestration of soil organic carbon *Agriculture, Ecosystems & Environment* **164** 80–99
Taylor K E, Stouffer R J and Meehl G A 2012 An overview of CMIP5 and the experiment design *Bull. Am. Meteorol. Soc.* **93** 485–98
Todd-Brown K E O, Randerson J T, Post W M, Hoffman F M, Tarnocai C, Schuur E A G and Allison S D 2013 Causes of variation in soil carbon simulations from CMIP5 Earth system models and comparison with observations. *Biogeosciences* **10** 1717–36
Wang W, Dungan J, Hashimoto H, Michaelis A R, Milesi C, Ichii K and Nemani R R 2011 Diagnosing and assessing uncertainties of terrestrial ecosystem models in a multimodel ensemble experiment: 1 *Primary production. Global Change Biology* **17** 1350–66
Wei Y, Liu S, Huntzinger D N, Michalak A M, Viovy N, Post W M and Tian H 2014 The North American carbon program multi-scale synthesis and terrestrial model intercomparison project: II. Environmental driver data *Geosci. Model Dev.* **7** 2875–93

Wuebbles D, Meehl G, Hayhoe K, Karl T R, Kunkel K, Santer B and Goodman A 2014 CMIP5 climate model analyses: climate extremes in the United States *Bull. Am. Meteorol. Soc.* **95** 571–83

Xia J Y, Luo Y Q, Wang Y P, Weng E S and 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** 1259–71

Zaehle S and Friend A D 2010 Carbon and nitrogen cycle dynamics in the O-CN land surface model: I. Model description, site-scale evaluation, and sensitivity to parameter estimates *Global Biogeochem. Cycles* **24** GB1005

Zhou S *et al* 2018 Sources of uncertainty in modeled land carbon storage within and across three MIPs: diagnosis with three new techniques *J. Clim.* **31** 2833–51