AGU PUBLICATIONS

Geophysical Research Letters

RESEARCH LETTER

10.1002/2015GL064002

Key Points:
- Naive (simple mean) and skill-based multi-model ensemble integration are indistinguishable
- This holds at the grid cell level and globally for land sink strength and reference variables
- Carbon metabolism has predictability limits and/or models/references are misspecified

Supporting Information:
- Figures S1–S13 and Table S1

Correspondence to:
C. R. Schwalm,
christopher.schwalm@nau.edu

Citation:
Schwalm, C. R., et al. (2015), Toward "optimal" integration of terrestrial biosphere models, *Geophys. Res. Lett.*, *42*, 4418–4428, doi:10.1002/2015GL064002.

Received 30 MAR 2015
Accepted 19 MAY 2015
Accepted article online 25 MAY 2015
Published online 10 JUN 2015

Toward "optimal" integration of terrestrial biosphere models

Christopher R. Schwalm[1,2], Deborah N. Huntzinger[2,3], Joshua B. Fisher[4], Anna M. Michalak[5], Kevin Bowman[4], Philippe Ciais[6], Robert Cook[7], Bassil El-Masri[8], Daniel Hayes[7], Maoyi Huang[9], Akihiko Ito[10], Atul Jain[8], Anthony W. King[7], Huimin Lei[11], Junjie Liu[4], Chaoqun Lu[12], Jiafu Mao[7], Shushi Peng[6], Benjamin Poulter[13], Daniel Ricciuto[7], Kevin Schaefer[14], Xiaoying Shi[7], Bo Tao[6], Hanqin Tian[12], Weile Wang[15], Yaxing Wei[7], Jia Yang[12], and Ning Zeng[16]

[1]Center for Ecosystem Science and Society, Northern Arizona University, Flagstaff, Arizona, USA, [2]School of Earth Sciences and Environmental Sustainability, Northern Arizona University, Flagstaff, Arizona, USA, [3]Department of Civil Engineering, Construction Management, and Environmental Engineering, Northern Arizona University, Flagstaff, Arizona, USA, [4]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California, USA, [5]Department of Global Ecology, Carnegie Institution for Science, Stanford, California, USA, [6]Laboratoire des Sciences du Climat et de l'Environnement, Gif sur Yvette, France, [7]Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA, [8]Department of Atmospheric Sciences, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA, [9]Atmospheric Sciences and Global Change Division, Pacific Northwest National Laboratory, Richland, Washington, USA, [10]National Institute for Environmental Studies, Tsukuba, Japan, [11]Department of Hydraulic Engineering, Tsinghua University, Beijing, China, [12]International Center for Climate and Global Change Research and School of Forestry and Wildlife Sciences, Auburn University, Auburn, Alabama, USA, [13]Department of Ecology, Montana State University, Bozeman, Montana, USA, [14]National Snow and Ice Data Center, Boulder, Colorado, USA, [15]Ames Research Center, National Aeronautics and Space Administration, Moffett Field, California, USA, [16]Department of Atmospheric and Oceanic Science, University of Maryland, College Park, Maryland, USA

Abstract Multimodel ensembles (MME) are commonplace in Earth system modeling. Here we perform MME integration using a 10-member ensemble of terrestrial biosphere models (TBMs) from the Multiscale synthesis and Terrestrial Model Intercomparison Project (MsTMIP). We contrast optimal (skill based for present-day carbon cycling) versus naïve ("one model-one vote") integration. MsTMIP optimal and naïve mean land sink strength estimates (−1.16 versus −1.15 Pg C per annum respectively) are statistically indistinguishable. This holds also for grid cell values and extends to gross uptake, biomass, and net ecosystem productivity. TBM skill is similarly indistinguishable. The added complexity of skill-based integration does not materially change MME values. This suggests that carbon metabolism has predictability limits and/or that all models and references are misspecified. Resolving this issue requires addressing specific uncertainty types (initial conditions, structure, and references) and a change in model development paradigms currently dominant in the TBM community.

1. Introduction

Multimodel ensembles (MME) are common in Earth system modeling and are routinely generated for model intercomparison projects (MIPs), e.g., Coupled Model Intercomparison Project Phase 3 (CMIP3) [*Meehl et al.,* 2007], C4MIP [*Friedlingstein et al.,* 2006], CMIP5 [*Taylor et al.,* 2012], and Intersectoral Impact Model Intercomparison Project [*Warszawski et al.,* 2013]. Two central challenges associated with MMEs are integration (how individual ensemble members are combined into a single-ensemble value) and interpretation (how MMEs inform our understanding of Earth system processes and their uncertainties) [*Annan and Hargreaves,* 2010; *Christensen and Boberg,* 2012; *Knutti,* 2010; *Hacker et al.,* 2011; *Stephenson et al.,* 2012; *von Storch and Zwiers,* 2013; *Zhao et al.,* 2013]. Integration methods range from "model democracy" or "one model-one vote" where ensemble integration is the mean across all models [*Zhao et al.,* 2013] to linear combinations of ensemble members informed by model error [*Eckel and Mass,* 2005], degree of independence [*Abramowitz and Gupta,* 2008; *Abramowitz,* 2010; *Masson and Knutti,* 2011], or model skill, e.g., Bayesian model averaging [*Raftery et al.,* 2005], reliability ensemble averaging [*Giorgi and Mearns,* 2002], and "superensembles" [*Stefanova and Krishnamurti,* 2002]. Regardless of approach, integrated ensembles typically show higher skill than all or most of the ensemble members [*Raftery et al.,* 2005] and are often used as the "best estimate" in climate change assessments [*Intergovernmental Panel on Climate Change*, 2007, 2010, 2013].

Ensemble methods may also be used to explore the uncertainty in model simulations that arises from internal variability, boundary conditions, parameter values for a given model structure, or structural uncertainty due to different model formulations [*Fisher et al.,* 2014; *Hawkins and Sutton,* 2009; *Huntzinger et al.,* 2013; *Knutti et al.,* 2010]. Uncertainty is typically quantified as some measure of spread across the ensemble, e.g., standard deviation. An important consideration here is whether the ensemble is broad enough to represent uncertainty [*Annan et al.,* 2011]. "Broadness" relates to how well the ensemble samples representations of a particular process. As an example, an ensemble that does not represent subgrid scale cloud formation or the soil moisture-precipitation feedback will not directly inform uncertainty related to these processes.

Traditionally, MME studies have focused primarily on the atmospheric component of Earth system models. This is related to the legacy of numerical weather prediction (NWP), which serves as the basis for the atmospheric component of climate models [*Sigalotti et al.,* 2014; *Lynch*, 2008], and where leveraging ensemble forecasts has a long tradition [e.g., *Epstein*, 1969]. In contrast, analyses of MME integration and interpretation have received significantly less attention for terrestrial biosphere models (TBMs)—the land component of climate or Earth system models—despite several large-scale model intercomparison projects, e.g., Vegetation/Ecosystem Modeling and Analysis Project [*Melillo*, 1995], Potsdam net primary productivity (NPP) MIP [*Cramer et al.,* 1999], the North American Carbon Program Interim Site [*Schwalm et al.,* 2010] and Regional Syntheses [*Huntzinger et al.,* 2012], the Trends in Net Land-Atmosphere Carbon Exchange [*Piao et al.,* 2013], and the Multiscale synthesis and Terrestrial Model Intercomparison Project (MsTMIP) [*Huntzinger et al.,* 2013].

Apart from equal weighting, MME integration generally requires some basis (e.g., model skill and error) to inform a linear combination of ensemble members. However, uncertainties or model error are not routinely available for TBM outputs, e.g., perturbed-physics ensembles are rare [e.g., *Booth et al.,* 2012; *Huntingford et al.,* 2009; *Zaehle et al.,* 2005]; and "truth" for TBMs, especially at the coarse spatial resolutions that typify TBM output, is not well constrained. Furthermore, total simulation duration for TBMs (years to centuries) is usually much longer than for NWP (days to weeks), resulting in a longer validation cycle. Despite these ongoing challenges for TBM ensemble integration, there is a clear need to better compare TBMs to each other and other independent estimates of land-atmosphere carbon dynamics to better constrain the past and future evolution of the terrestrial carbon land sink.

In this study, we develop a methodology that uses an MME to generate a best estimate of land-atmosphere CO_2 flux and its associated uncertainty. Our approach uses 10 state-of-the-art TBM simulations from a model intercomparison study with a prescribed simulation protocol [*Huntzinger et al.,* 2013; *Wei et al.,* 2014]. The principal goal of this study is to contrast the extent to which an "intelligent" skill-based integration differs from naïve integration. In the following section, we describe the model ensemble and its integration with optimal weights derived using model-reference mismatch or benchmarking [*Luo et al.,* 2012]. In section 3, we contrast the naïve case (one model-one vote) with the optimal case. Lastly, in section 4 we discuss the implications of our findings and suggestions for future research.

2. Model Ensemble and Integration

The model ensemble is drawn from the Multiscale synthesis and Terrestrial Model Intercomparison Project (MsTMIP) [*Huntzinger et al.,* 2013]. MsTMIP uses a prescribed simulation protocol to isolate structural differences in model output, with driving data, land cover, and steady state spin-up all standardized across models [*Wei et al.,* 2014]. MsTMIP global monthly model runs span a 110 year period (1901–2010) and use a semifactorial set of simulations where time-varying climate, CO_2 concentration, land cover, and nitrogen deposition are sequentially "turned on" after steady state is achieved [*Huntzinger et al.,* 2013]. For this study, we use the simulation results from 10 TBMs (Table 1) released under MsTMIP version 1 (http://nacp.ornl.gov/mstmipdata/mstmip_simulation_results_global_v1.jsp). Here simulations have all factors enabled (MsTMIP simulation BG1). For the subset of models that do not include a nitrogen cycle, SG3 runs (which exclude nitrogen deposition but are otherwise identical to BG1) are used.

For model integration, i.e., combining ensemble members to a single-integrated value, we contrast two use cases (i) the ensemble mean where each model is weighted equally (hereafter: naïve case), and (ii) an optimal case where weights are derived using reliability ensemble averaging (REA) [*Giorgi and Mearns,* 2002]. We apply these two use cases to four variables: net ecosystem exchange (NEE, i.e., land sink strength), gross primary productivity (GPP), vegetation biomass, and net ecosystem productivity (NEP). MsTMIP definitions

Table 1. Characteristics of Terrestrial Biosphere Models and Reference Data Sets[a]

Model	Run	NEE Components	NEE (Pg C yr^{-1})	NEP (Pg C yr^{-1})	GPP (Pg C yr^{-1})	Vegetation Biomass (Gt C)	Reference
BIOME-BGC	BG1	*F*	−0.38	6.46	138	1138	*Thornton et al.* [2002]
CLM	BG1	$D/F/E_{LUC}/P$	0.16	4.46	142	668	*Mao et al.* [2012]
CLM4VIC	BG1	$D/F/E_{LUC}/P$	−0.15	3.57	112	550	*Lei et al.* [2014]
DLEM	BG1	E_{LUC}/P	−1.51	2.18	105	475	*Tian et al.* [2012]
GTEC	SG3	*P*	−2.79	9.67	187	986	*King et al.* [1997] and *Ricciuto et al.* [2011]
ISAM	BG1	E_{LUC}	0.24	1.49	99	642	*Jain and Yang* [2005]
LPJ	SG3	F/E_{LUC}	−0.53	10.55	138	536	*Sitch et al.* [2003]
ORCHIDEE-LSCE	SG3	E_{LUC}/P	−1.84	6.68	118	460	*Krinner et al.* [2005]
VEGAS2.1	SG3	$F/E_{LUC}/P$	−1.11	4.48	117	597	*Zeng et al.* [2005]
VISIT	SG3	–	−3.63	3.63	122	763	*Ito* [2010]
MsTMIP median	–	–	–	–	120	620	This study
FLUXNET-based GPP	–	–	–	–	119	–	*Jung et al.* [2011]
IPCC vegetation Biomass	–	–	–	–	–	491	*Ruesch and Gibbs* [2008]
Naïve integration	–	–	−1.15	5.32	128	681	This study
Optimal integration	–	–	−1.16	5.76	136	699	This study

[a]Native 0.5° spatial resolution for all TBMs. NEE components refer to aspects of biosphere-atmosphere exchange included in NEE: D, maintenance respiration deficit; F, fire emissions; E_{LUC}, land use change emissions; P, product decay emissions. VISIT does not include any of these components. The MsTMIP median model is used for convergence-based reference factors. Carbon fluxes and biomass model values are 1982–2008 global means.

for NEP and NEE are NEP = GPP − R_h − R_a and NEE = $R_h + R_a + E_{LUC} + P$ − GPP, respectively, where R_h is heterotrophic respiration, R_a autotrophic respiration, E_{LUC} emissions from anthropogenic activities (e.g., deforestation, shifting agriculture, and biomass burning) that cause land use change [*Le Quéré et al.,* 2013], and *P* is emissions due to harvested wood product decay.

The weights required for the optimal case are derived using REA. This method uses reference data products and model-reference mismatch [*Luo et al.,* 2012] as well as intermodel spread [*Giorgi and Mearns,* 2002] to determine model reliability:

$$R_i = \prod_j f_j^{m_j} \tag{1}$$

where R_i is the model reliability factor for model *i* at a given land grid cell, f_j represents model skill relative to reference factor *j*, and m_j is a weighting factor. The m_j exponent term gives the relative importance of model skill for each reference factor *j* [*Eum et al.,* 2012]. In this study, all m_j are initially assumed equal at unity and we calculate reference factors for gross uptake and biomass. We note that while more directly observable quantities (e.g., evapotranspiration per basin or the global residual carbon sink) are available, we use gridded references to recover the spatial morphology of skill and reliability at the scale at which MsTMIP simulations are executed.

For gross uptake, we use the global GPP MPI-BGC product based on upscaled eddy covariance (FLUXNET) data [*Beer et al*., 2010; *Jung et al.,* 2011]. GPP is the largest global carbon flux [*Beer et al*., 2010], the dominant carbon input source for terrestrial ecosystems [*Chapin et al.,* 2006], and is important in model benchmarking as TBMs simulate carbon dynamics “downstream” of GPP, i.e., errors in GPP propagate to errors in carbon stocks and other fluxes [*Schaefer et al.,* 2012]. The MPI-BGC GPP data set is available monthly at 0.5° spatial resolution from 1982 to 2008 and is routinely used in benchmarking [e.g., *Anav et al.,* 2013; *Piao et al.,* 2013]. While the MPI-BGC product also includes NEE (−17.1 ± 4.7 Pg C per annum), it differs markedly from other estimates, e.g., −2.6 ± 0.8 Pg C per annum from the Global Carbon Project [*Le Quéré et al.,* 2013; http://www.globalcarbonproject.org/]. This bias is also present in upscaled ecosystem respiration and is related to processes not well resolved [*Jung et al.,* 2011] by FLUXNET (e.g., land use change, fire emissions, postdisturbance recovery, export of carbon by biomass harvesting and soil erosion [*Regnier et al.,* 2013], and carbon emissions from reduced carbon species [*Ciais et al.,* 2008]).

The biomass reference is taken from the Intergovernmental Panel on Climate Change (IPCC) Tier-1 vegetation biomass product [*Ruesch and Gibbs,* 2008]. This product is based on specific biomass (above and belowground) values for 124 carbon zones mapped using geospatial data sets of global land cover, continent, ecofloristic zone, and forest age. On multidecadal scales, vegetation biomass contributes to net land-atmosphere exchange of carbon [*Houghton*, 2005] and has direct implications for assessing forest deforestation [*Keith et al.,* 2009], especially reductions in emissions from deforestation and forest degradation in tropical forests [*Gibbs et al.,* 2007]. This data set is available for circa 2000 on a 10 min global grid and is regridded using box averaging to 0.5° spatial resolution.

Using these two reference products, we derive, for each grid cell over the 1982–2008 period, seven reference factors (Table S1 in the supporting information) used to calculate R_i. These factors are bound by zero and unity, and quantify (i) bias in mean long-term GPP ($f_{B,i}$), (ii) bias in the standard deviation of mean long-term GPP ($f_{\sigma,i}$), (iii) convergence [*Giorgi and Mearns,* 2002] in simulated GPP ($f_{C,i}$), (iv) bias in GPP trend ($f_{T,i}$), (v) correlation in GPP ($f_{\rho,i}$), (vi) bias in biomass ($f_{\beta,i}$), and (vii) convergence in simulated biomass ($f_{\gamma,i}$). The convergence factors address intermodel spread whereby higher convergence indicates that simulation output is largely insensitive to TBM, i.e., a robust signal is found across the majority of models [*Giorgi and Mearns,* 2002]. All reference factors (except $f_{\rho,i}$) are based on normalizing uncertainty by the absolute difference between the reference and simulation. Finally, all factors use well-established skill metrics from intercomparison studies [e.g., *Cadule et al.,* 2010; *Exbrayat et al.,* 2013; *Fisher et al.,* 2014; *Luo et al.,* 2012] and address both the distance between simulated and reference values as well as their correlation and variability in time and space.

With each reference factor defined and equal importance, equation (1) simplifies to

$$R_i = f_{B,i} \times f_{\sigma,i} \times f_{C,i} \times f_{T,i} \times f_{\rho,i} \times f_{\beta,i} \times f_{\gamma,i} \tag{2}$$

These R_i values are then normalized to composite model reliability $\left(\widetilde{R}_i\right)$ for each model, i.e., R_i is scaled to sum to unity across all n models in the ensemble $\left(\sum_{i=1}^{n} \widetilde{R}_i = 1\right)$ for each grid cell. These reliabilities, $\widetilde{R}_i$, serve as optimal weights for MME integration

$$\widetilde{F} = \sum_{i}^{n} \widetilde{R}_i F_i \tag{3}$$

where F is one of NEE, GPP, vegetation biomass, or NEP for model i, and $\widetilde{F}$, optimally integrated F, is calculated for each vegetated grid cell, i.e., although R_i is derived using GPP and vegetation biomass, it is used for all four variables.

To assess uncertainty of the optimal integration, we generate 1000 bootstrap replicates by randomly varying the relative importance of each reference factor m_j from 0 (i.e., excluded from reliability calculations) to 7 (i.e., only factor considered). Uncertainty is given as either a confidence bound (the 2.5th to 97.5th percentiles) or the standard deviation across all bootstrap replicates where each represents an alternative, albeit plausible, optimal integration.

3. Naïve Versus Optimal Cases

For global aggregates, the naïve and optimal cases are indistinguishable despite strong spatial variability in composite model reliability (Figure S1) and individual reference factors (Figures S2–S11). Naïve case NEE is estimated as −1.15 versus −1.16 Pg C per annum for the optimal case, values reference 1982–2008 means. This difference of −0.01 Pg C per annum is small (Figure 1) relative to the uncertainty of optimal integration (1σ across 1000 replicates: 0.09 Pg C per annum) and relative to interannual variability (1σ across 27 global annual values: 1.13 (naïve) versus 1.02 (optimal) Pg C per annum).

For NEE, the lack of significant difference occurs (i) despite variations in components included in simulated NEE (Table 1), (ii) even though the reference flux GPP does not fully constrain NEE, and (iii) despite smaller ranges in GPP and biomass compared to NEE (Table 1): GPP varies by a factor of ~ 2 (from 99 (Integrated Science Assessment Model (ISAM)) to 187 Global Terrestrial Ecosystem Carbon (GTEC) Pg C per annum), and biomass varies by a factor of ~ 2.5 (from 460 (Organizing Carbon and Hydrology in Dynamic Ecosystems-Laboratoire des Sciences du Climat et de l'Environnement (ORCHIDEE-LSCE) to 1138 (BIOME-BioGeochemical Cycles (BGC)) Gt C), whereas NEE ranges from +0.24 (a weak source; ISAM) to −3.63 (a strong sink; Vegetation Integrative SImulator for Trace gases (VISIT)) Pg C per annum.

The lack of difference between naïve and optimal cases globally is supported by uniformly small grid cell differences. The uncertainty of the optimal integration is greater than the difference between the cases for 84% of the vegetated land surface (Figure 1). Also, the spatial morphology of both cases shows a high

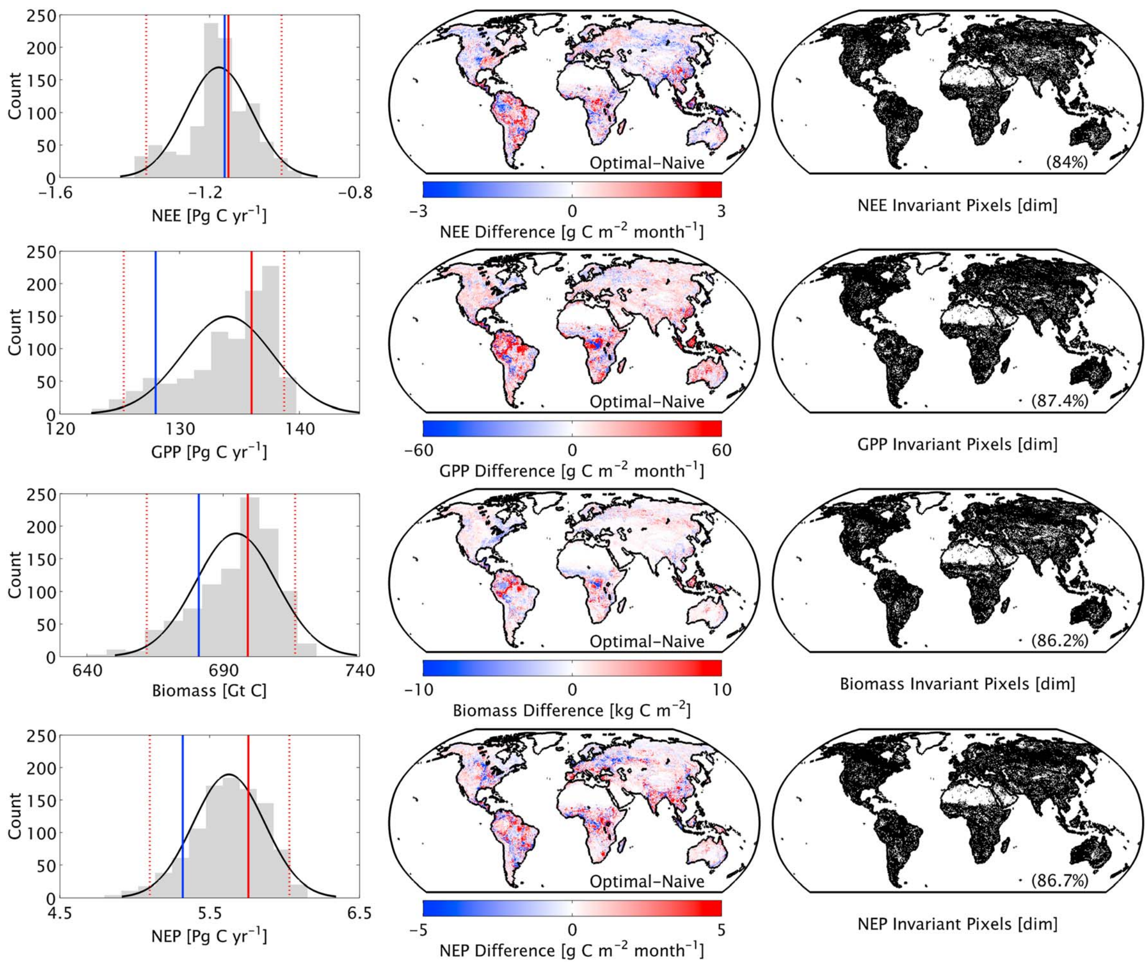

Figure 1. Difference between optimal and naïve cases for NEE, GPP, biomass, and NEP. (left column) Histograms (gray), fitted normal distribution (black line), naïve case (blue line), optimal case (dark red line), and optimal case uncertainty bounds (light dashed red lines) for global values. Distributions of optimal case based on 1000 bootstrap replicates with varying reference factor importance. Uncertainty bounds are given by the 2.5th to 97.5th percentiles. (middle column) Difference map of optimal and naïve cases. (right column) Black grid cells indicate where the naïve is indistinguishable from the optimal case (values in parentheses show percentage of indistinguishable grid cells for the vegetated land surface). All values reference 1982–2008 means.

degree of similarity without any region that skews the global integrals; only a weak tendency for slightly larger (albeit statistically insignificant) differences in tropical forests is present (Figure 2). This holds for composite model reliability as well as considering each reference factor singly (Figure S12).

In using TBM skill for GPP and biomass to estimate reliability for NEE, we assume model skill is transitive, i.e., skill in the former is relevant for a model's ability to simulate the latter. As a test, we evaluate integration differences for GPP and biomass as well. A result in contrast to NEE would violate this assumption. While there are larger magnitude differences between the optimal and naïve case for GPP (128 and 136 Pg C per annum for naïve and optimal, respectively) and biomass (681 and 699 Gt C for naïve and optimal, respectively), these differences are statistically insignificant relative to the uncertainty of the optimal case (Figure 1).

A key concern in the comparison of naïve and optimal values is the semantic differences in NEE [*Hayes and Turner*, 2012]. While all TBMs adhere to the MsTMIP protocol, not all TBMs are able to simulate all components of NEE (Table 1). That is, if NEE is indistinguishable across naïve and optimal integration, this begs the question if the inclusion/exclusion of relevant NEE components acts in a compensatory manner. Thus, as an additional check on the equivalence of naïve and optimal cases, we test the impact of variable NEE semantics directly using NEP. This test is based on using the largest subset of NEE components simulated across the full ensemble. Here only gross uptake and gross loss are simulated by all TBMs. The disequilibrium between these two fluxes is *per definitionem* NEP. As seen with GPP and biomass, which are

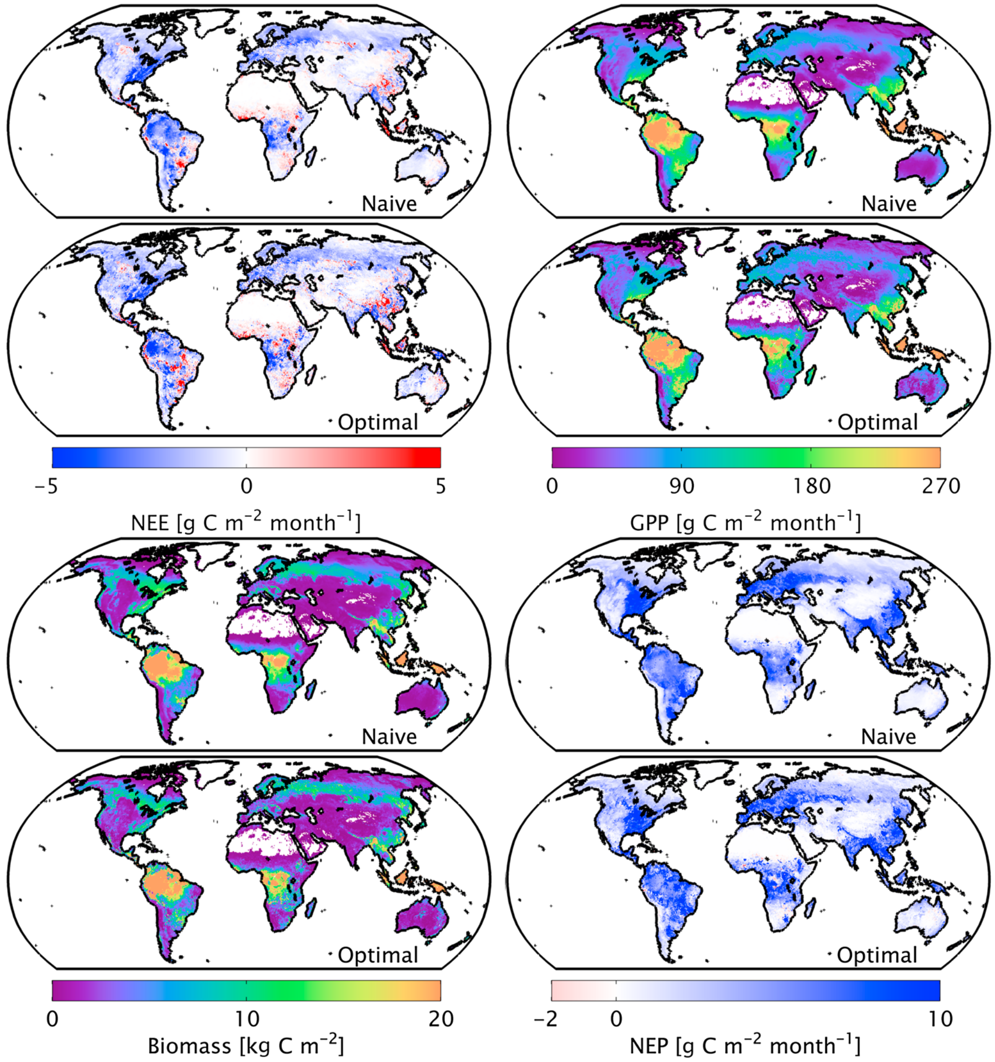

Figure 2. Spatial patterns of naïve and optimal cases. Maps show naïve and optimal case 1982–2008 means for NEE, GPP, biomass, and NEP.

also semantically equivalent across models, differences in NEP (5.32 and 5.76 Pg C per annum for naïve and optimal, respectively) are statistically insignificant relative to the uncertainty of the optimal case (Figure 1).

Furthermore, the lack of difference in global integrals is, as seen for NEE, supported by the small magnitudes of grid cell difference between cases (Figure 1) and the high degree of similarity in spatial morphology across cases (Figure 2) for NEP, GPP, and biomass. No region skews the global values with only a weak tendency for slightly larger differences in tropical forests, especially for GPP. For NEP, GPP, and biomass, the percent of grid cells where the difference between naïve and optimal values is less than the uncertainty of the optimal integration is 87%, 87%, and 86%, respectively (Figure 1).

Does that lack of a significant difference in integrated values indicate that the naïve case is "correct"? The naïve case presupposes equal weighting, i.e., one model-one vote. For composite model reliabilities $\left(\widetilde{R}_i\right)$, this implies weights of unity normalized by the number of ensemble members, i.e., uncertainty bounds derived from the 1000 replicates must contain a global mean $\widetilde{R}_i$ of 0.1 for each model. This is the case for 8 of the 10 models; ISAM and ORCHIDEE-LSCE are near misses where the upper uncertainty bounds are just below this cutoff (0.096 and 0.095, respectively). A similar pattern is seen with model rank, i.e., a one-number assessment of relative skill (Figure S13). Here model ranks show considerable overlap without any clear indication of "best" or "worst." Furthermore, even when focusing on a single-bootstrap replicate,

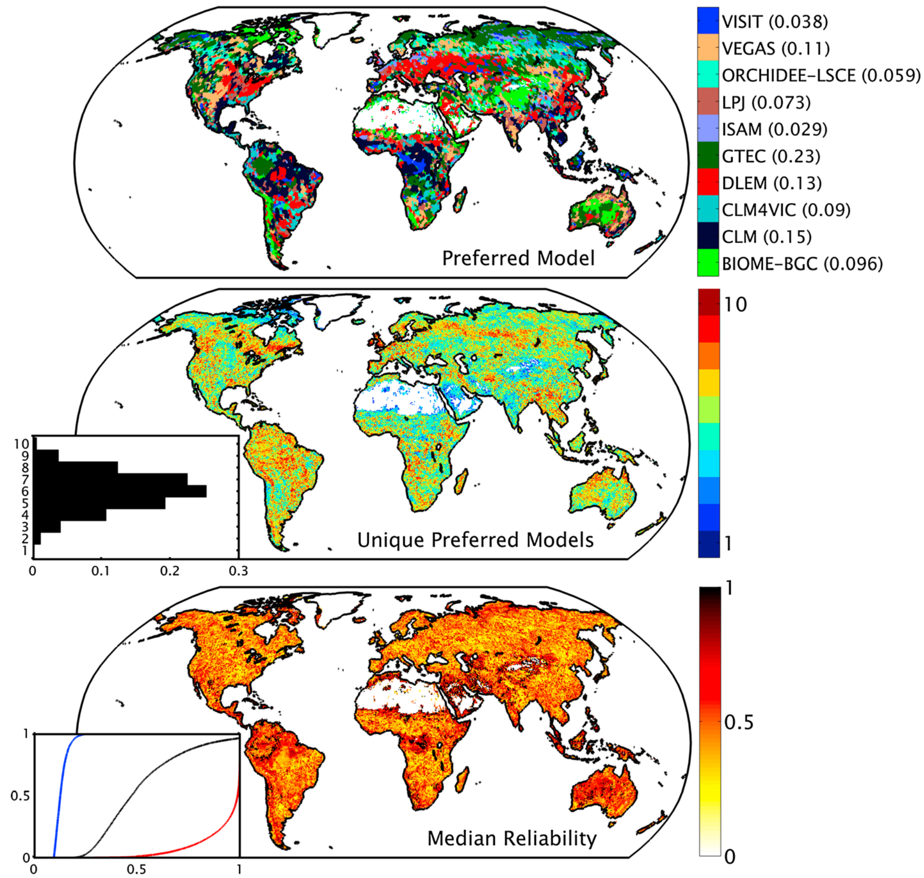

Figure 3. Preferred model. (top) Preferred model based on equal relative importance of all seven reference factors, the default optimal case. Values in parenthesis show fraction of vegetated land surface where a given model is preferred. A 3×3 majority filter is used for visualization purposes. (middle) Number of unique preferred models across all bootstrap replicates, inset shows histogram. (bottom) Median reliability of preferred model across all 1000 bootstrap replicates; inset shows cumulative distribution (*y* axis) over maximum (red), median (black), and minimum (blue) reliability (*x* axis).

a higher rank does not demonstrate that one model is "good" per se. As reliabilities do not exceed 0.25 (unity indicates perfect agreement between TBM and references), a higher rank only shows that the predictive skill of a higher ranked model is marginally higher than the next ranked model. Taken together, the equivalence in global model reliabilities and rank strongly implies that the benchmarking and complexity inherent in optimal integration add no value relative to the naïve case.

Collapsing $\widetilde{R}_i$ for each grid cell to ranks yields the preferred model (Figure 3). "Preferred" here indicates the highest composite $\widetilde{R}_i$. Applying this approach, the most skilled TBM is GTEC which is the preferred model for ~ 23% of the vegetated land surface. However, the preferred model is, as seen for global ranks, highly variable (Figure 3). Depending on reference factor importance, ~ 75% of all vegetated grid cells have between 4 and 7 different preferred models (Figure 3, inset) with only 33 of 55,457 vegetated grid cells having the same preferred model throughout. Lastly, while there is the suggestion (Figure 3) that some TBMs exhibit higher skill levels, the associated variability emphasizes the equivalence of models (Figure 3, inset). That is, a given TBM only posts higher reliability scores under a particular set of references and relative importance of those reference factors. These conditions are not identifiable a priori such that skill-based discrimination is not feasible as the signal (actual model skill) is dwarfed by the noise (plausible approaches to asses actual model skill).

4. Implications

The equivalence of the naïve and optimal cases is a troubling but robust finding of this study. The difference between both integrations is small in magnitude and less than the uncertainty associated with the optimal integration. This holds for global aggregates and is the overwhelmingly dominant pattern on a grid cell basis. Equivalence also applies to both semantically identical (GPP, biomass, and NEP) and semantically diverse (NEE) simulation outputs. Taken together, this indicates that TBM skill is largely indistinguishable as well as malleable in that over a plausible set of skill assessments (i.e., the variants in REA from bootstrapping) a model's reliability ranges widely.

To better understand the interplay between TBM skill, ensemble integration, and benchmarking, several innovations are needed: As with the atmospheric component of Earth system models, the land component evaluated here must be regularly subject to perturbed-physics ensembles (where parameterizations are varied within some tolerance). This is motivated by parameter tuning [*Bindoff et al.,* 2013; *Flato et al.,* 2013] and the social anchoring tendency of models to regress to the mean value of an existing ensemble or reference [*Knutti*, 2010; *Sanderson and Knutti,* 2012]. A systemic exploration of parameter-based divergence in model outputs is needed to quantify and isolate sources of uncertainty and "detune" models (i.e., uncover compensatory errors) [*Collins et al.,* 2011]. A second innovation concerns steady state spin-up. Models are routinely run to equilibrium states, where change in carbon stocks is zero within some tolerance [e.g., *Huntzinger et al.,* 2013] prior to actual simulation. However, the resultant initial carbon pool sizes vary dramatically both for fully coupled Earth system models [*Exbrayat et al.,* 2014] and TBMs. For the MsTMIP ensemble evaluated here, starting soil carbon pools range from 409 to 2118 relative to a reference value of 890 to 1660 Gt C [*Todd-Brown et al.,* 2013]. Given the interplay between carbon pool size and carbon flux insuring a model's equilibrated state is similar to observations will materially affect TBM skill.

Systemically varying TBM structure [*Curry and Webster,* 2011; *McWilliams*, 2007] is also a needed innovation. This is especially warranted given the recent emphasis on more comprehensive treatments of Earth climate system dynamics. This additional complexity does not guarantee more accurate projections [*Knutti and Sedláček,* 2013] but represents another structural component to assess. Here a change in model building is needed such that discrete subroutines can be altered systematically. Target subroutines must include known problematic processes (e.g., phenology [*Richardson et al.,* 2012], net land use flux [*Pongratz et al.,* 2014], or carbon allocation [*De Kauwe et al.,* 2014]) as well as, in the case of MsTMIP, key processes with uneven (or absent) structural representation [*Huntzinger et al.,* 2014] such as carbon-nitrogen interactions [*Zaehle et al.,* 2014], phosphorous limitation, fire emissions, forest management, and forest age structure. Note that this is a refinement of the prescribed protocol used in MsTMIP which fixes nonstructural TBM characteristics but does not guarantee that the ensemble range in structural characteristics equates to a systematic sampling of all possible modeling algorithms.

A further protocol refinement concerns the use of offline runs. While this effectively controls for model-specific implementations of atmospheric coupling, it can be considered biased as interactions between the surface energy budget and atmospheric conditions are missing. This suggests a nested experimental design whereby the components of a fully coupled Earth system model (land, cryosphere, atmosphere, and ocean) are, in conjunction with the semifactorial base runs, systemically varied. A full factorial design with systematically toggleable subroutines across all Earth system model domains, in turn, requires a deeper understanding of the trade-offs between ensemble size, model complexity, and computational resources [*Ferro et al.,* 2012]. A corollary to this approach is to move model development toward using stochastic treatments of unresolved processes [*Palmer et al.,* 2014], and the realization that treating ensemble spread as uncertainty is an approximation [*Curry and Webster,* 2011; *Parker*, 2010].

Another key innovation concerns "ground truth" for gridded model outputs. Here the analyst must contend with multiple plausible references [e.g., *Mitchard et al.,* 2014; *Schwalm et al.,* 2013] and/or references with large uncertainty bounds [*Todd-Brown et al.,* 2013]. For point-based data upscaled to gridded reference products, like the GPP product used here, representativeness is a further concern [*Schwalm et al.,* 2011]. The resultant ambiguity surrounding ground truth can render model reliability a pliable construct. As such, we suggest a parallel track of MIPs and DIPs, i.e., data intercomparison projects where "data" encompass observationally based reference products. Only when reference data sets themselves have been reconciled and their uncertainty quantified at scales that typify TBM simulations can we unambiguously assess TBM skill. This highlights an advantage of skill-based integration that generalizes to accommodate MIP- and/or DIP-based uncertainties (using χ^2-based metrics) [*Schwalm et al.,* 2010] where available. MIPs and DIPs must also be viewed as necessary vehicles to explicitly link TBM skill gradients to intrinsic model structural characteristics. Effectively mapping uncertainty-aware skill gradients to structural attributes [*Schwalm et al.,* 2010; *Xia et al.,* 2013] has great potential to inform future development of TBMs by identifying subroutines associated with higher skill.

Finally, it is important to emphasize that the TBM equivalence shown here is in the context of carbon metabolism for a given model ensemble with a given set of references. Previous work [*Schwalm et al.,* 2013] showed similar results in model skill assessment using evapotranspiration from fully coupled CMIP5 runs, and we expect this overall result to generalize across multiple land surface processes, especially when ground truth is ambiguous. The equivalence between naïve and optimal cases is, however, not a reason to abandon skill-based integration or TBM skill assessment in general. Advancing our understanding across the full taxonomy of uncertainties is necessary to resolve actual model skill as well as issues of MME integration and interpretation. This taxonomy includes uncertainty relative to parameterization, steady state spin-up (i.e., initial conditions), structure, reference data, and forcing data (relatively well established in the land surface modeling community) [e.g., *Barman et al.,* 2014a, 2014b; *Fekete et al.,* 2004; *Haddeland et al.,* 2011; *Jain et al.,* 2013].

As is, the enduring popularity of the naïve case is based both on ease (e.g., no references are needed) and the higher skill generally shown by the naïve case relative to most or all ensemble members singly. While it is possible that land surface carbon metabolism has predictability limits similar to atmospheric dynamics [*Slingo and Palmer,* 2011]—variously termed σ_{climate}, "irreducible imprecision," or "irreducible ignorance" [*McWilliams,* 2007; *Walker et al.,* 2003]—only a full inventory of uncertainty types will allow an intelligent skill-based integration and reveal if TBMs are subject to "reducible ignorance" (where additional insight and predictive skill are achievable) [*Luo et al.,* 2014] or irreducible ignorance (where predictive skill is limited).

Acknowledgments
C.R.S. was supported by National Aeronautics and Space Administration (NASA) grants NNX12AP74G, NNX10AG01A, and NNX11AO08A. J.B.F. carried out this research at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with NASA. Funding for the Multiscale synthesis and Terrestrial Model Intercomparison Project (MsTMIP; http://nacp.ornl.gov/MsTMIP.shtml) activity was provided through NASA ROSES grant NNX10AG01A. Data management support for preparing, documenting, and distributing model driver and output data was performed by the Modeling and Synthesis Thematic Data Center at Oak Ridge National Laboratory (ORNL; http://nacp.ornl.gov), with funding through NASA ROSES grant NNH10AN681. Finalized MsTMIP data products are archived at the ORNL DAAC (http://daac.ornl.gov). This is MsTMIP contribution 5. Acknowledgments for specific MsTMIP participating models: Biome-BGC: Biome-BGC code was provided by the Numerical Terradynamic Simulation Group at the University of Montana. The computational facilities provided by NASA Earth Exchange at NASA Ames Research Center. CLM: This research is supported in part by the U.S. Department of Energy (DOE), Office of Science, Biological and Environmental Research. Oak Ridge National Laboratory is managed by UTBATTELLE for DOE under contract DE-AC05-00OR22725. CLM4VIC: CLM4VIC simulations were supported in part by the U.S. Department of Energy (DOE), Office of

References

Abramowitz, G. (2010), Model independence in multi-model ensemble prediction, *Aust. Meteorol. Oceanogr. J.*, *59*, 3–6.
Abramowitz, G., and H. Gupta (2008), Towards a model space and model independence metric, *Geophys. Res. Lett.*, *35*, L05705, doi:10.1029/2007GL032834.
Anav, A., P. Friedlingstein, M. Kidston, L. Bopp, P. Ciais, P. Cox, C. Jones, M. Jung, R. Myneni, and Z. Zhu (2013), Evaluating the land and ocean components of the global carbon cycle in the CMIP5 Earth System Models, *J. Clim.*, *26*(18), 6801–6843.
Annan, J. D., and J. C. Hargreaves (2010), Reliability of the CMIP3 ensemble, *Geophys. Res. Lett.*, *37*, L02703, doi:10.1029/2009GL041994.
Annan, J. D., J. C. Hargreaves, and K. Tachiiri (2011), On the observational assessment of climate model performance, *Geophys. Res. Lett.*, *38*, L24702, doi:10.1029/2011GL049812.
Barman, R., A. K. Jain, and M. Liang (2014a), Climate-driven uncertainties in modeling terrestrial energy and water fluxes: A site-level to global-scale analysis, *Global Change Biol.*, *20*(6), 1885–1900.
Barman, R., A. K. Jain, and M. Liang (2014b), Climate-driven uncertainties in modeling terrestrial gross primary production: A site level to global-scale analysis, *Global Change Biol.*, *20*(5), 1394–1411.
Beer, C., et al. (2010), Terrestrial gross carbon dioxide uptake: Global distribution and covariation with climate, *Science*, *329*(5993), 834–838.
Bindoff, N., et al. (2013), Detection and attribution of climate change: From global to regional, in *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, edited by T. Stocker et al., pp. 867–952, Cambridge Univ. Press, Cambridge, U. K., and New York.
Booth, B. B., C. D. Jones, M. Collins, I. J. Totterdell, P. M. Cox, S. Sitch, C. Huntingford, R. A. Betts, G. R. Harris, and J. Lloyd (2012), High sensitivity of future global warming to land carbon cycle processes, *Environ. Res. Lett.*, *7*(2), 024002.
Cadule, P., P. Friedlingstein, L. Bopp, S. Sitch, C. D. Jones, P. Ciais, S. L. Piao, and P. Peylin (2010), Benchmarking coupled climate-carbon models against long-term atmospheric CO_2 measurements, *Global Biogeochem. Cycles*, *24*, GB2016, doi:10.1029/2009GB003556.
Chapin, F. S., III, et al. (2006), Reconciling carbon-cycle concepts, terminology, and methods, *Ecosystems*, *9*(7), 1041–1050.
Christensen, J. H., and F. Boberg (2012), Temperature dependent climate projection deficiencies in CMIP5 models, *Geophys. Res. Lett.*, *39*, L24705, doi:10.1029/2012GL053650.
Ciais, P., A. V. Borges, G. Abril, M. Meybeck, G. Folberth, D. Hauglustaine, and I. A. Janssens (2008), The impact of lateral carbon fluxes on the European carbon balance, *Biogeosciences*, *5*(5), 1259–1271.
Collins, M., B. B. Booth, B. Bhaskaran, G. R. Harris, J. M. Murphy, D. M. Sexton, and M. J. Webb (2011), Climate model errors, feedbacks and forcings: A comparison of perturbed physics and multi-model ensembles, *Clim. Dyn.*, *36*(9–10), 1737–1766.
Cramer, W., D. Kicklighter, A. Bondeau, B. Moore III, G. Churkina, B. Nemry, A. Ruimy, and A. Schloss (1999), Comparing global models of terrestrial net primary productivity (NPP): Overview and key results, *Global Change Biol.*, *5*(S1), 1–15.
Curry, J. A., and P. J. Webster (2011), Climate science and the uncertainty monster, *Bull. Am. Meteorol. Soc.*, *92*(12), 1667–1682.
De Kauwe, M. G., et al. (2014), Where does the carbon go? A modeldata intercomparison of vegetation carbon allocation and turnover processes at two temperate forest free-air CO_2 enrichment sites, *New Phytol.*, *203*, 883–899.
Eckel, F. A., and C. F. Mass (2005), Aspects of effective mesoscale, short-range ensemble forecasting, *Weather Forecast.*, *20*, 328–350.
Epstein, E. S. (1969), Stochastic dynamic prediction, *Tellus*, *21*(6), 739–759.
Eum, H., P. Gachon, R. Laprise, and T. Ouarda (2012), Evaluation of regional climate model simulations versus gridded observed and regional reanalysis products using a combined weighting scheme, *Clim. Dyn.*, *38*, 1433–1457.
Exbrayat, J. F., N. R. Viney, H. G. Frede, and L. Breuer (2013), Using multi-model averaging to improve the reliability of catchment scale nitrogen predictions, *Geosci. Model Dev.*, *6*(1), 117–125.

Science, Biological and Environmental Research (BER) through the Earth System Modeling program and performed using the Environmental Molecular Sciences Laboratory(EMSL), a national scientific user facility sponsored by the U.S. DOE-BER and located at Pacific Northwest National Laboratory (PNNL). Participation of M. Huang in the MsTMIP synthesis is supported by the U.S. DOE-BER through the Subsurface Biogeochemical Research Program (SBR) as part of the SBR Scientific Focus Area (SFA) at the Pacific Northwest National Laboratory (PNNL). PNNL is operated for the U.S. DOE by BATTELLE Memorial Institute under contract DE-AC05-76RLO1830. DLEM: The Dynamic Land Ecosystem Model (DLEM) developed in the International Center for Climate and Global Change Research at Auburn University has been supported by NASA Interdisciplinary Science Program (IDS), NASA Land Cover/Land Use Change Program (LCLUC), NASA Terrestrial Ecology Program, NASA Atmospheric Composition Modeling and Analysis Program (ACMAP); NSF Dynamics of Coupled Natural-Human System Program (CNH), Decadal and Regional Climate Prediction using Earth System Models (EaSM); DOE National Institute for Climate Change Research; USDA AFRI Program; and EPA STAR Program. Integrated Science Assessment Model (ISAM) simulations were supported by the U.S. National Science Foundation (NSF-AGS-12-43071 and NSF-EFRI-083598), the USDA National Institute of Food and Agriculture (NIFA) (2011-68002–30220), the U.S. Department of Energy (DOE) Office of Science (DOE-DE-SC0006706), and the NASA Land Cover and Land Use Change Program (NNX14AD94G). ISAM simulations were carried out at the National Energy Research Scientific Computing Center (NERSC), which is supported by the Office of Science of the U.S. Department of Energy under contract DE-AC02-05CH11231, and at the Blue Waters sustained-petascale computing, University of Illinois at Urbana-Champaign, which is supported by the National Science Foundation (awards OCI-0725070 and ACI-1238993) and the state of Illinois. LPJ-wsl: This work was conducted at LSCE, France, using a modified version of the LPJ version 3.1 model, originally made available by the Potsdam Institute for Climate Impact Research. ORCHIDEE-LSCE: ORCHIDEE is a global land surface model developed at the IPSL institute in France. The simulations were performed with the support of the GhG Europe FP7 grant with computing facilities provided by LSCE (Laboratoire des Sciences du Climat et de l'Environnement) or TGCC (Très Grand Centre de Calcul). VISIT: VISIT was developed at the National Institute for

Exbrayat, J.-F., A. J. Pitman, and G. Abramowitz (2014), Response of microbial decomposition to spin-up explains CMIP5 soil carbon range until 2100, *Geosci. Model Dev.*, *7*, 2683–2692, doi:10.5194/gmd-7-2683-2014.

Fekete, B. M., C. J. Vörösmarty, J. O. Roads, and C. J. Willmott (2004), Uncertainties in precipitation and their impacts on runoff estimates, *J. Clim.*, *17*, 294–304.

Ferro, C. A., T. E. Jupp, F. H. Lambert, C. Huntingford, and P. M. Cox (2012), Model complexity versus ensemble size: Allocating resources for climate prediction, *Philos. Trans. R. Soc. A*, *370*(1962), 1087–1099.

Fisher, J. B., D. N. Huntzinger, C. R. Schwalm, and S. Sitch (2014), Modeling the terrestrial biosphere, *Annu. Rev. Environ. Resour.*, *39*, 91–123, doi:10.1146/annurev-environ-012913-093456.

Flato, G., et al. (2013), Evaluation of climate models, in *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, edited by T. Stocker et al., pp. 741–866, Cambridge Univ. Press, Cambridge, U. K., and New York.

Friedlingstein, P., et al. (2006), Climate-carbon cycle feedback analysis: Results from the C4MIP model intercomparison, *J. Clim.*, *19*(14), 3337–3353.

Gibbs, H. K., S. Brown, J. O. Niles, and J. A. Foley (2007), Monitoring and estimating tropical forest carbon stocks: Making REDD a reality, *Environ. Res. Lett.*, *2*(4), 045023.

Giorgi, F., and L. O. Mearns (2002), Calculation of average, uncertainty range, and reliability of regional climate changes from AOGCM simulations via the "reliability ensemble averaging"(REA) method, *J. Clim.*, *15*(10), 1141–1158.

Hacker, J. P., S. Y. Ha, C. Snyder, J. Berner, F. A. Eckel, E. Kuchera, M. Pocernich, S. Rugg, J. Schramm, and X. Wang (2011), The US Air Force Weather Agency's mesoscale ensemble: Scientific description and performance results, *Tellus, Ser. A*, *63*(3), 625–641.

Haddeland, I., et al. (2011), Multimodel estimate of the global terrestrial water balance: Setup and first results, *J. Hydrometeorol.*, *12*(5), 869–884.

Hawkins, E., and R. Sutton (2009), The potential to narrow uncertainty in regional climate predictions, *Bull. Am. Meteorol. Soc.*, *90*, 1095–1107.

Hayes, D., and D. Turner (2012), The need for "apples-to-apples" comparisons of carbon dioxide source and sink estimates, *Eos Trans. AGU*, *93*(41), 404–405, doi:10.1029/2012EO410007.

Houghton, R. A. (2005), Aboveground forest biomass and the global carbon balance, *Global Change Biol.*, *11*(6), 945–958.

Huntingford, C., J. A. Lowe, B. B. B. Booth, C. D. Jones, G. R. Harris, L. K. Gohar, and P. Meir (2009), Contributions of carbon cycle uncertainty to future climate projection spread, *Tellus, Ser. B*, *61*(2), 355–360.

Huntzinger, D. N., et al. (2012), North American Carbon Program (NACP) regional interim synthesis: Terrestrial biospheric model intercomparison, *Ecol. Modell.*, *232*, 144–157.

Huntzinger, D. N., et al. (2013), The North American Carbon Program Multi-scale synthesis and Terrestrial Model Intercomparison Project—Part 1: Overview and experimental design, *Geosci. Model Dev.*, *6*, 2121–2133, doi:10.5194/gmd-6-2121-2013.

Huntzinger, D. N., C. Schwalm, A. M. Michalak, K. Schaefer, Y. Wei, R. B. Cook, and A. Jacobson (2014), NACP MsTMIP summary of model structure and characteristics, ORNL DAAC, Oak Ridge, Tenn., doi:10.3334/ORNLDAAC/1228. [Available at http://daac.ornl.gov.]

Intergovernmental Panel on Climate Change (2007), *Climate Change 2007: The Physical Science Basis, Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, edited by S. Solomon et al., Cambridge Univ. Press, Cambridge, U. K., and New York.

Intergovernmental Panel on Climate Change (2010), *Meeting Report of the Intergovernmental Panel on Climate Change Expert Meeting on Assessing and Combining Multi Model Climate Projections*, edited by T. F. Stocker et al., 117 pp., IPCC Working Group I Technical Support Unit, Univ. of Bern, Bern, Switzerland.

Intergovernmental Panel on Climate Change (2013), *Climate Change 2013: The Physical Science Basis, Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, edited by T. F. Stocker, Q. Dahe, and G.-K. Plattner, Cambridge Univ. Press, Cambridge, U. K., and New York.

Ito, A. (2010), Changing ecophysiological processes and carbon budget in East Asian ecosystems under near-future changes in climate: Implications for long-term monitoring from a process-based model, *J. Plant Res.*, *123*, 577–588, doi:10.1007/s10265-009-0305-x.

Jain, A. K., and X. Yang (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*, GB2015, doi:10.1029/2004GB002349.

Jain, A. K., P. Meiyappan, Y. Song, and J. I. House (2013), CO_2 emissions from land-use change affected more by nitrogen cycle, than by the choice of land-cover data, *Global Change Biol.*, *19*(9), 2893–2906.

Jung, M., et al. (2011), Global patterns of land-atmosphere fluxes of carbon dioxide, latent heat, and sensible heat derived from eddy covariance, satellite, and meteorological observations, *J. Geophys. Res.*, *116*, G00J07, doi:10.1029/2010JG001566.

Keith, H., B. G. Mackey, and D. B. Lindenmayer (2009), Re-evaluation of forest biomass carbon stocks and lessons from the world's most carbon-dense forests, *Proc. Natl. Acad. Sci. U.S.A.*, *106*(28), 11,635–11,640.

King, A. W., W. M. Post, and S. D. Wullschleger (1997), The potential response of terrestrial carbon storage to changes in climate and atmospheric CO_2, *Clim. Change*, *35*, 199–227.

Knutti, R. (2010), The end of model democracy?, *Clim. Change*, *102*(3–4), 395–404.

Knutti, R., and J. Sedláček (2013), Robustness and uncertainties in the new CMIP5 climate model projections, *Nat. Clim. Change*, *3*(4), 369–373.

Knutti, R., R. Furrer, C. Tebaldi, J. Cermak, and G. A. Meehl (2010), Challenges in combining projections from multiple climate models, *J. Clim.*, *23*, 2739–2758.

Krinner, G., N. Viovy, N. de Noblet-Ducoudre, J. Ogee, J. Polcher, P. Friedlingstein, P. Ciais, S. Sitch, and I. C. Prentice (2005), A dynamic global vegetation model for studies of the coupled atmosphere-biosphere system, *Global Biogeochem. Cycles*, *19*, GB1015, doi:10.1029/2003GB002199.

Lei, H., M. Huang, L. R. Leung, D. Yang, X. Shi, J. Mao, D. J. Hayes, C. R. Schwalm, Y. Wei, and S. Liu (2014), Sensitivity of global terrestrial gross primary production to hydrologic states simulated by the Community Land Model using two runoff parameterizations, *J. Adv. Model. Earth Syst.*, *6*, 658–679, doi:10.1002/2013MS000252.

Le Quéré, C., et al. (2013), The global carbon budget 1959–2011, *Earth Syst. Sci. Data*, *5*, 165–185, doi:10.5194/essd-5-165-2013.

Luo, Y., T. F. Keenan, and M. Smith (2014), Predictability of the terrestrial carbon cycle, *Global Change Biol.*, *21*, 1737–1751.

Luo, Y. Q., et al. (2012), A framework for benchmarking land models, *Biogeosciences*, *9*(10), 3857–3874.

Lynch, P. (2008), The origins of computer weather prediction and climate modeling, *J. Comput. Phys.*, *227*(7), 3431–3444.

Mao, J., P. E. Thornton, X. Shi, M. Zhao, and W. M. Post (2012), Remote sensing evaluation of CLM4 GPP for the period 2000–09, *J. Clim.*, *25*, 5327–5342, doi:10.1175/JCLI-D-11-00401.1.

Masson, D., and R. Knutti (2011), Climate model genealogy, *Geophys. Res. Lett.*, *38*, L08703, doi:10.1029/2011GL046864.

McWilliams, J. C. (2007), Irreducible imprecision in atmospheric and oceanic simulations, *Proc. Natl. Acad. Sci. U.S.A.*, *104*(21), 8709–8713.

Meehl, G. A., C. Covey, K. E. Taylor, T. Delworth, R. J. Stouffer, M. Latif, B. McAvaney, and J. F. Mitchell (2007), The WCRP CMIP3 multimodel dataset: A new era in climate change research, *Bull. Am. Meteorol. Soc.*, *88*(9), 1383–1394.

Environmental Studies, Japan. This work was mostly conducted during a visiting stay at Oak Ridge National Laboratory.

The Editor thanks an anonymous reviewer for assisting in the evaluation of this paper.

Melillo, J. M. (1995), Vegetation/ecosystem modeling and analysis project: Comparing biogeography and biogeochemistry models in a continental-scale study of terrestrial ecosystem responses to climate change and CO_2 doubling, *Global Biogeochem. Cycles*, *9*, 407–437, doi:10.1029/95GB02746.

Mitchard, E. T., et al. (2014), Markedly divergent estimates of Amazon forest carbon density from ground plots and satellites, *Global Ecol. Biogeogr.*, *23*(8), 935–946, doi:10.1111/geb.12168.

Palmer, T., P. Düben, and H. McNamara (2014), Stochastic modelling and energy-efficient computing for weather and climate prediction, *Philos. Trans. R. Soc. A*, *372*(2018), 20140118.

Parker, W. S. (2010), Predicting weather and climate: Uncertainty, ensembles and probability, *Stud. Hist. Philos. Sci. B*, *41*(3), 263–272.

Piao, S., et al. (2013), Evaluation of terrestrial carbon cycle models for their response to climate variability and to CO_2 trends, *Global Change Biol.*, *19*(7), 2117–2132.

Pongratz, J., C. H. Reick, R. A. Houghton, and J. I. House (2014), Terminology as a key uncertainty in net land use and land cover change carbon flux estimates, *Earth Syst. Dyn.*, *5*(1), 177–195.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski (2005), Using Bayesian model averaging to calibrate forecast ensembles, *Mon. Weather Rev.*, *133*(5), 1155–1174.

Regnier, P., et al. (2013), Anthropogenic perturbation of the carbon fluxes from land to ocean, *Nat. Geosci.*, *6*(8), 597–607.

Ricciuto, D., A. W. King, D. Dragoni, and W. M. Post (2011), Parameter and prediction uncertainty in an optimized terrestrial carbon cycle model: Effects of constraining variables and data record length, *J. Geophys. Res.*, *116*, G01033, doi:10.1029/2010JG001400.

Richardson, A. D., et al. (2012), Terrestrial biosphere models need better representation of vegetation phenology: Results from the North American Carbon Program Site Synthesis, *Global Change Biol.*, *18*(2), 566–584.

Ruesch, A., and H. K. Gibbs (2008), New IPCC Tier-1 Global Biomass Carbon Map for the Year 2000, Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, Oak Ridge, Tenn. [Available at http://cdiac.ornl.gov.]

Sanderson, B., and R. Knutti (2012), Climate change projections: Characterizing uncertainty using climate models, in *Climate Change Modeling Methodology*, pp. 235–259, Springer, New York.

Schaefer, K., et al. (2012), A model-data comparison of gross primary productivity: Results from the North American Carbon Program site synthesis, *J. Geophys. Res.*, *117*, G03010, doi:10.1029/2012JG001960.

Schwalm, C. R., et al. (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, doi:10.1029/2009JG001229.

Schwalm, C. R., C. A. Williams, and K. Schaefer (2011), Carbon consequences of global hydrologic change, 1948–2009, *J. Geophys. Res.*, *116*, G03042, doi:10.1029/2011JG001674.

Schwalm, C. R., D. N. Huntinzger, A. M. Michalak, J. B. Fisher, J. S. Kimball, B. Mueller, K. Zhang, and Y. Zhang (2013), Sensitivity of inferred climate model skill to evaluation decisions: A case study using CMIP5 evapotranspiration, *Environ. Res. Lett.*, *8*(2), 024028.

Sigalotti, L. D. G., E. Sira, J. Klapp, and L. Trujillo (2014), Environmental fluid mechanics: Applications to weather forecast and climate change, in *Computational and Experimental Fluid Mechanics With Applications to Physics, Engineering and the Environment*, pp. 3–36, Springer Int., Heidelberg, Germany.

Sitch, S., et al. (2003), Evaluation of ecosystem dynamics, plant geography and terrestrial carbon cycling in the LPJ dynamic vegetation model, *Global Change Biol.*, *9*, 161–185.

Slingo, J., and T. Palmer (2011), Uncertainty in weather and climate prediction, *Philos. Trans. R. Soc. A*, *369*(1956), 4751–4767.

Stefanova, L., and T. N. Krishnamurti (2002), Interpretation of seasonal climate forecast using Brier skill score, the Florida State University superensemble, and the AMIP-I dataset, *J. Clim.*, *15*(5), 537–544.

Stephenson, D. B., M. Collins, J. C. Rougier, and R. E. Chandler (2012), Statistical problems in the probabilistic prediction of climate change, *Environmetrics*, *23*(5), 364–372.

Taylor, K. E., R. J. Stouffer, and G. A. Meehl (2012), An overview of CMIP5 and the experiment design, *Bull. Am. Meteorol. Soc.*, *93*(4), 485–498.

Thornton, P. E., et al. (2002), Modeling and measuring the effects of disturbance history and climate on carbon and water budgets in evergreen needleleaf forests, *Agric. For. Meteorol.*, *113*, 185–222.

Tian, H. Q., et al. (2012), Century-scale response of ecosystem carbon storage to multifactorial global change in the Southern United States, *Ecosystems*, *15*(4), 674–694, doi:10.1007/s10021-012-9539-x.

Todd-Brown, K. E. O., J. T. Randerson, W. M. Post, F. M. Hoffman, C. Tarnocai, E. A. G. Schuur, and S. D. Allison (2013), Causes of variation in soil carbon simulations from CMIP5 Earth system models and comparison with observations, *Biogeosciences*, *10*, 1717–1736, doi:10.5194/bg-10-1717-2013.

von Storch, H., and F. Zwiers (2013), Testing ensembles of climate change scenarios for "statistical significance", *Clim. Change*, *117*(1–2), 1–9.

Walker, W. E., P. Harremoes, J. Rotmans, J. P. van der Sluijs, M. B. A. van Asselt, P. Janssen, and M. P. Krayer von Krauss (2003), Defining uncertainty: A conceptual basis for uncertainty management in model-based decision support, *Integr. Assess.*, *4*, 5–17.

Warszawski, L., K. Frieler, V. Huber, F. Piontek, O. Serdeczny, and J. Schewe (2013), The Inter-Sectoral Impact Model Intercomparison Project (ISI-MIP): Project framework, *Proc. Natl. Acad. Sci. U.S.A.*, *111*(9), 3228–3232.

Wei, Y., et al. (2014), The North American Carbon Program Multi-scale Synthesis and Terrestrial Model Intercomparison Project—Part 2: Environmental driver data, *Geosci. Model Dev.*, *7*, 2875–2893, doi:10.5194/gmd-7-2875-2014.

Xia, J., Y. Luo, Y. P. Wang, and O. Hararuk (2013), Traceable components of terrestrial carbon storage capacity in biogeochemical models, *Global Change Biol.*, *19*(7), 2104–2116.

Zaehle, S., S. Sitch, B. Smith, and F. Hattermann (2005), Effects of parameter uncertainties on the modeling of terrestrial biosphere dynamics, *Global Biogeochem. Cycles*, *19*, GB3020, doi:10.1029/2004GB002395.

Zaehle, S., et al. (2014), Evaluation of 11 terrestrial carbonnitrogen cycle models against observations from two temperate free-air CO_2 enrichment studies, *New Phytol.*, *202*(3), 803–822.

Zeng, N., A. Mariotti, and P. Wetzel (2005), Terrestrial mechanisms of interannual CO_2 variability, *Global Biogeochem. Cycles*, *19*, GB1016, doi:10.1029/2004GB002273.

Zhao, Z. C., Y. Luo, and J. B. Huang (2013), A review on evaluation methods of climate modeling, *Adv. Clim. Change Res.*, *4*(3), 137–144.