AGU PUBLICATIONS

JGR

Journal of Geophysical Research: Biogeosciences

RESEARCH ARTICLE
10.1002/2013JG002553

Special Section:
Experiment-Model Integration in Terrestrial Ecosystem Study: Current Practices and Future Challenges

Key Points:
- Two temperate forest FACE experiments were simulated with 11 ecosystem models
- Transpiration biases were often caused by leaf area biases
- Accuracy was sometimes achieved with compensating biases in component variables

Supporting Information:
- Readme
- Text S1

Correspondence to:
A. P. Walker,
walkerap@ornl.gov

Citation:
Walker, A. P., et al. (2014), Comprehensive ecosystem model-data synthesis using multiple data sets at two temperate forest free-air CO_2 enrichment experiments: Model performance at ambient CO_2 concentration, *J. Geophys. Res. Biogeosci.*, *119*, 937–964, doi:10.1002/2013JG002553.

Received 21 NOV 2013
Accepted 25 APR 2014
Accepted article online 2 MAY 2014
Published online 27 MAY 2014

Comprehensive ecosystem model-data synthesis using multiple data sets at two temperate forest free-air CO_2 enrichment experiments: Model performance at ambient CO_2 concentration

Anthony P. Walker[1,2], Paul J. Hanson[1], Martin G. De Kauwe[3], Belinda E. Medlyn[3], Sönke Zaehle[4], Shinichi Asao[5], Michael Dietze[6], Thomas Hickler[7,8], Chris Huntingford[9], Colleen M. Iversen[1], Atul Jain[10], Mark Lomas[2], Yiqi Luo[11], Heather McCarthy[11], William J. Parton[5], I. Colin Prentice[3,12], Peter E. Thornton[1], Shusen Wang[13], Ying-Ping Wang[14], David Warlind[15], Ensheng Weng[16], Jeffrey M. Warren[1], F. Ian Woodward[2], Ram Oren[17,18], and Richard J. Norby[1]

[1]Environmental Sciences Division and Climate Change Science Institute, Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA, [2]Department of Animal and Plant Sciences, University of Sheffield, Sheffield, UK, [3]Department of Biological Sciences, Macquarie University, Sydney, New South Wales, Australia, [4]Biogeochemical Integration Department, Max Planck Institute for Biogeochemistry, Jena, Germany, [5]Department of Ecosystem Science and Sustainability, Colorado State University, Fort Collins, Colorado, USA, [6]Department of Earth and Environment, Boston University, Boston, Massachusetts, USA, [7]Biodiversity and Climate Research Centre (BiK-F) and Senckenberg Gesellschaft für Naturforschung, Frankfurt/Main, Germany, [8]Department of Physical Geography, Göethe University, Frankfurt/Main, Germany, [9]Centre for Ecology and Hydrology, Wallingford, UK, [10]Department of Atmospheric Sciences, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA, [11]Department of Microbiology and Plant Biology, University of Oklahoma, Norman, Oklahoma, USA, [12]Department of Life Sciences and Grantham Institute for Climate Change, Imperial College London, London, UK, [13]Canada Centre for Mapping and Earth Observation, Natural Resources Canada, Ottawa, Ontario, Canada, [14]CSIRO Marine and Atmospheric Research and Centre for Australian Weather and Climate Research, Aspendale, Victoria, Australia, [15]Department of Physical Geography and Ecosystem Science, Lund University, Lund, Sweden, [16]Department of Ecology and Evolutionary Biology, Princeton University, Princeton, New Jersey, USA, [17]Division of Environmental Science and Policy, Nicholas School of the Environment, Duke University, Durham, North Carolina, USA, [18]Department of Forest Ecology and Management, Swedish University of Agricultural Sciences, Umeå, Sweden

Abstract Free-air CO_2 enrichment (FACE) experiments provide a remarkable wealth of data which can be used to evaluate and improve terrestrial ecosystem models (TEMs). In the FACE model-data synthesis project, 11 TEMs were applied to two decadelong FACE experiments in temperate forests of the southeastern U.S.—the evergreen Duke Forest and the deciduous Oak Ridge Forest. In this baseline paper, we demonstrate our approach to model-data synthesis by evaluating the models' ability to reproduce observed net primary productivity (NPP), transpiration, and leaf area index (LAI) in ambient CO_2 treatments. Model outputs were compared against observations using a range of goodness-of-fit statistics. Many models simulated annual NPP and transpiration within observed uncertainty. We demonstrate, however, that high goodness-of-fit values do not necessarily indicate a successful model, because simulation accuracy may be achieved through compensating biases in component variables. For example, transpiration accuracy was sometimes achieved with compensating biases in leaf area index and transpiration per unit leaf area. Our approach to model-data synthesis therefore goes beyond goodness-of-fit to investigate the success of alternative representations of component processes. Here we demonstrate this approach by comparing competing model hypotheses determining peak LAI. Of three alternative hypotheses—(1) optimization to maximize carbon export, (2) increasing specific leaf area with canopy depth, and (3) the pipe model—the pipe model produced peak LAI closest to the observations. This example illustrates how data sets from intensive field experiments such as FACE can be used to reduce model uncertainty despite compensating biases by evaluating individual model assumptions.

1. Introduction

The terrestrial carbon cycle is a major source of interannual and intraannual variability in the global carbon cycle [*Canadell et al.*, 2007; *Le Quere et al.*, 2009]. Many of the uncertainties in Earth System model projections are related to uncertainties in the representation of the terrestrial carbon cycle and its response to environmental change, in particular, atmospheric CO_2 and climate [*Cramer et al.*, 2001; *Friedlingstein et al.*,

2006; *Sitch et al*., 2008; *Forster et al*., 2013; *Piao et al*., 2013]. To reduce this uncertainty, there is a need to first evaluate and identify sources of uncertainty in terrestrial ecosystem and biosphere models (TEMs) and then to improve TEMs using a wide range of ecosystem data. There have been a number of studies that have evaluated TEMs against different kinds of ecosystem-scale data including eddy covariance [*Schwalm et al*., 2010; *Dietze et al*., 2011; *Keenan et al*., 2012; *Schaefer et al*., 2012] and precipitation manipulations [*Hanson et al*., 2004; *Powell et al*., 2013]. This paper analyses simulations from a model-experiment intercomparison project in order to evaluate model predictions against data from the ambient CO_2 treatments in two temperate forest free-air CO_2 enrichment (FACE) experiments. The intercomparison involved multiple modeling groups and strong involvement from the experimentalists. This paper focuses on the ability of models to simulate ecosystems under ambient CO_2, while other papers provide similar discussion for water, carbon, and nitrogen cycle responses to elevated CO_2: *De Kauwe et al*. [2013, 2014] and *Zaehle et al*. [2014].

FACE experiments are ideal for model testing because they provide simultaneous data sets of multiple ecosystem properties at scales suitable for direct comparison with models [*Körner et al*., 2005; *Hendrey et al*., 1999; *Oren et al*., 2001; *Norby et al*., 2006; *Calfapietra et al*., 2001; *Zak et al*., 2011]. Situated in the southeastern U.S., the two forest FACE experiments used in this intercomparison, Duke and Oak Ridge, yielded rich and detailed data sets on the state and dynamics of temperate forest ecosystems across a range of temporal scales. Some data, such as weather data and sap flow, were resolved hourly, while the length of the FACE experiments used in this synthesis (~11 years) was sufficiently long to detect relatively slow feedbacks, such as nutrient limitation [*Johnson*, 2006; *Norby et al*., 2010]. This paper focuses on net primary productivity (NPP) and transpiration which drive carbon (C) and water fluxes, two of the main objectives of TEMs, and leaf area index (LAI) which plays a key role in simulating both NPP and transpiration by scaling leaf-level C and water fluxes to the forest canopy. Furthermore, leaf-level water C fluxes and water fluxes are linked via stomatal conductance [*De Kauwe et al*., 2013], so that NPP, transpiration, and LAI are all linked such that a bias in any one will lead to a bias in the other two. The purpose of this paper is threefold: (1) to detail the FACE model-experiment synthesis describing the two FACE experiments, the 11 TEMs, and how the models were applied to the FACE sites; (2) to evaluate and compare the models and their ability to simulate key ecosystem variables that were directly measured in the FACE studies—NPP, transpiration, and LAI—under ambient CO_2 conditions; and (3) to investigate the relationship between simulated transpiration and LAI to elucidate biases in transpiration driven by biases in LAI predictions.

While this paper focuses on ambient CO_2 conditions, it also discusses the consequences for the prediction of responses to elevated CO_2 that are detailed elsewhere. Goal two above uses the FACE experimental data to assess model performance or skill in prediction of ecosystem dynamics in response to environmental variability for each of the two ecosystems (Figure 1a). Model performance is quantified using a wide range of goodness-of-fit (GOF) statistics [*Nash and Sutcliffe*, 1970; *Smith and Rose*, 1995; *Moriasi et al*., 2007], to which we apply bootstrapping to estimate confidence in the GOF metrics.

Assessing GOF does not require understanding of the underlying modeling assumptions, and much can be gained from diagnosing and understanding key model assumptions that result in good fit to the data and that cause variability among model results [e.g., *De Kauwe et al*., 2013]. Goal three above aims to synthesize model and experimental data (Figure 1b) viewing models as coherent sets of quantitative hypotheses. Key hypotheses that lead to differences in predictions across models can be identified and categorized for evaluating different modeling hypotheses and assumptions (i.e., not just the individual models) using multiple observations. The model-data synthesis approach, illustrated in Figure 1b, generates recommendations for future model development and prioritizes hypotheses that require further experimental testing [*Medlyn et al*., 2005].

Simulating the Duke and Oak Ridge FACE experiments together provides a useful comparison of the C and water fluxes in different systems (evergreen versus deciduous) having similar climates. Simulations of annual NPP, daily transpiration, and daily LAI by 11 TEMs applied to the ambient CO_2 treatments are evaluated as key indicators of the state and dynamics of ecosystem carbon and water cycles at the two sites. We also assess GOF of annual NPP in the context of component variables. The capacity of the model-data synthesis approach is then demonstrated by evaluating the simulation of transpiration at the two FACE experiments. We hypothesize that a component of transpiration biases are caused by LAI biases and test this using a very simple conceptual model that expresses total plant transpiration as a rate of water use per unit

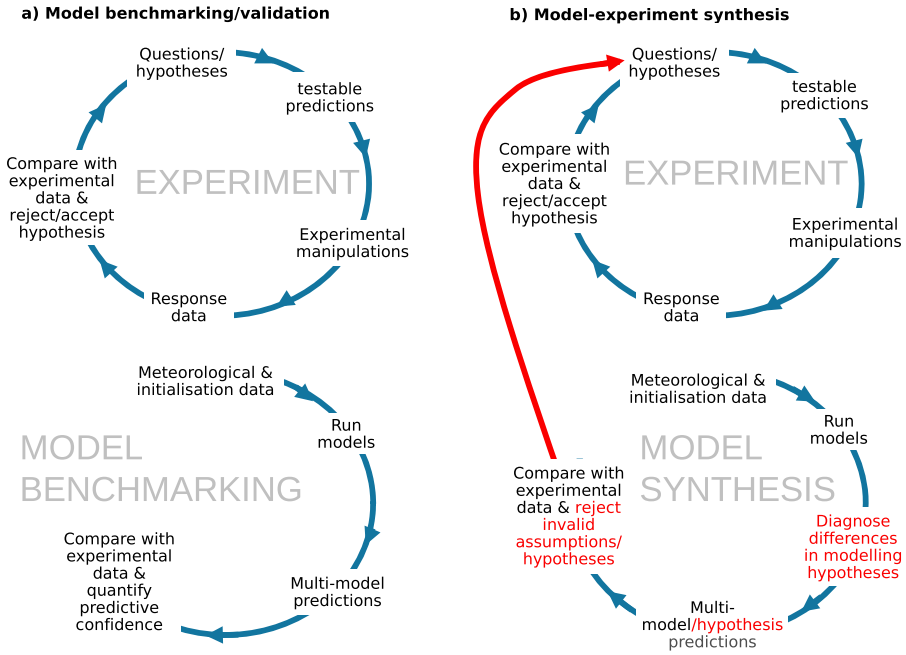

Figure 1. A schematic diagram of model-experiment data interactions: (a) assessment of goodness of fit (GOF) of model predictions to experimental data (benchmarking) and (b) model-experiment synthesis. The differences of model-experiment synthesis and benchmarking are highlighted in red. For model-experiment synthesis the modeling loop feeds back into the experimental loop, and another arrow could even be drawn whereby predictions are initially generated by the collection of hypotheses represented by a suite of models which feed directly into the experimental cycle.

leaf area index (LAI), multiplied by the LAI. The method used by each model to calculate LAI is investigated and the influence of the underlying assumptions discussed. Finally, we consider the implications of our findings for the simulation of elevated CO_2 effects on ecosystem carbon and water fluxes.

2. Methods

2.1. Site Descriptions

FACE experiments subject intact ecosystems to an atmosphere enriched in CO_2 [*Hendrey et al.*, 1999]. We simulated the Duke [*Oren et al.*, 2001; *McCarthy et al.*, 2010; *Drake et al.*, 2011] and the Oak Ridge [*Norby et al.*, 2006, p. 200, 2010; *Iversen et al.*, 2011] FACE experiments located in the southeastern USA. Both sites were situated in young (11 years old at the beginning of the experiments) closed-canopy, unmanaged plantation ecosystems. The two FACE sites were similar climatically, but they differed in soil type, species composition, and phenology—evergreen needleleaf-dominated canopy at Duke with a deciduous broadleaf understory and a deciduous broadleaf stand at Oak Ridge with little understory. Initial tree and soil conditions and other site and experimental details are reported in Table 1. Experimental design, measurement protocols, and ecosystem responses have been described in multiple papers, see above and below and others listed at http://face.env.duke.edu/ and http://face.ornl.gov/pubs.html.

2.1.1. The Duke FACE Experiment

The Duke FACE experiment [*Oren et al.*, 2001] was located within a 90 ha loblolly pine (*Pinus taeda* L.—Piedmont provenance) plantation situated in the Duke Forest, Chapel Hill, North Carolina (35.97°N, 79.08°W) (Table 1). The forest is on a moderately low fertility acidic loam supporting a site index (at age 25) for loblolly pine of 16 m (dominant trees were 13.2 m and 11 years old at the beginning of the experiment) and rooting depths were restricted to the upper 75 cm of the soil profile. The climate is typical of the warm-humid Piedmont region of the southeastern U.S. (mean annual temperature 15.5°C and mean annual precipitation 1150 mm), with precipitation evenly distributed throughout the year. The trees were planted in 1983 at a spacing of 2 m × 2.4 m. Canopy closure occurred around 1998, and peak stand LAI (~6 including hardwoods) was reached in 2001 [*McCarthy et al.*, 2007].

Table 1. Comparison of Duke and Oak Ridge FACE Experiment Characteristics

	Duke FACE	Oak Ridge FACE
Location	Orange County, North Carolina 35°58′N, 79°06′W Elevation 163 m	Roane County, Tennessee 35°54′N, 84°20′W Elevation 230 m
Soil Classification (U.S.)	Ultic Hapludalf	Aquic Hapludult
Soil texture	acidic loam (49% sand, 42% silt, 9% clay)	silty clay-loam (21% sand, 55% silt, 24% clay)
Soil C content (Mg ha^{-1})	101	74
Soil N content (Mg ha^{-1})	3	8
Mean annual Temp (°C)	15.5	13.9
Mean annual precipitation (mm)	1145	1371
N deposition (kg ha^{-1} y^{-1})	13.7	12–15
Site history (used in model initialization)	Pre-1800, deciduous broadleaf forest 1800, clear cut to grassland 1920, forest establishment allowed 1982, clear cut and burned, and plantation established	Pre-1750, deciduous broadleaf forest 1750, clear cut to C4 crop 1943, grassland established 1988, plantation established
Dominant species	*Pinus taeda* (L)	*Liquidambar styraciflua* (L)
Main other species present (understory unless specified)	*Liquidambar* styraciflua (some canopy trees), *Ulmus alata*, *Acer rubrum*, *Cornus florida*	*Elaeagnus umbellata*, *Microstegium vimineum*, *Lonicera japonica*, *Acer negundo*, *Liriodendron tulipifera*, *Lindera benzoin*
Nominal elevated CO_2 concentration (ppm)	Ambient +200	565
Treatment duration	1994–2010	1998–2009
Age at initiation	11	11
Number of plots	4 elevated, 4 ambient (1994–1996 one plot per treatment)	2 elevated, 3 ambient
Plot size (m^2)	527	314
Number trees per plot	86 *P. taeda* and 140 canopy and subcanopy broad-leaved individuals > 2 cm at 1.3 m	~90 *L. styraciflua*
Initial dominant height (m)	13.2	12.4
Initial peak leaf area index (LAI)	3.8	5.5
Initial basal area (m^2 ha^{-1})	32.5	29
Initial stem + branch mass (kg C m^{-2})	4.0	3.6
Initial leaf mass (kg C m^{-2})	0.3	0.17
Initial coarse root mass (kg C m^{-2})	0.9	1.4
Initial peak fine root mass (kg C m^{-2})	0.1	0.37
Initial leaf C:N	-[a]	24.3
Initial wood C:N	-	365
Root C:N[b]	-	67

[a]No data.
[b]Ambient treatment mean.

CO_2 enrichment commenced in August 1996 (targeted at ambient +200 ppmv) in the replicated experiment (four experimental plots in each treatment). There was heterogeneity in N availability at the site and elevated CO_2 experimental plots were paired (blocked) with ambient CO_2 control plots based on initial N availability [*DeLucia et al.*, 1999]. CO_2 enrichment occurred during daylight hours of the growing season targeting ambient +200 ppmv. The mean elevated CO_2 concentration during 1996–2004 was 571 ppmv, with 92% of 1 min CO_2 means within 20% of the target (average target = 573 ppmv).

2.1.2. The Oak Ridge FACE Experiment

The Oak Ridge FACE experiment [*Norby et al.*, 2006, p. 200] was located in a sweet gum (*Liquidambar styraciflua* L.) plantation on the Oak Ridge National Environmental Research Park, Tennessee (35.90°N, 84.33°W). The forest is on a low fertility silty clay-loam supporting a site index (at age 50) for sweet gum of 23–24 m (trees were 12.4 m tall and 11 years old at the beginning of the experiment). The climate at the site is typical of the humid southern Appalachian region (mean annual temperature 13.9°C and mean annual precipitation 1370 mm); weather records during the experiment were reported by [*Riggs et al.*, 2009]. At the start of the experiment, the trees had a fully developed canopy and were in a linear growth phase.

CO_2 enrichment began in May 1998 in two of five experimental plots (the remaining three plots were the control ambient CO_2 treatment). The FACE apparatus was constructed following the design employed at the Duke FACE experiment [*Hendrey et al*., 1999] and CO_2 enrichment targeted 565 ppmv. The mean elevated CO_2 concentration over the course of the experiment was 545 ppmv, and in 1998 and 1999, 90% of 1 min CO_2 means were within 20% of the target [*Norby et al*., 2001].

2.2. Experimental Data

Over the entire course of the experiments at both sites, the most continuous and integrative data over the spatial and temporal scales of the experiment were the measurements of NPP. NPP comprising leaf, wood and coarse root, and fine-root production was calculated from primary measurements of litter mass, specific leaf area, tree height and diameter, and minirhizotron or root ingrowth observations [*Norby et al*., 2001, 2003; *McCarthy et al*., 2007, 2010; *Iversen et al*., 2008; *Pritchard et al*., 2008].

Daily LAI at both sites was inferred from measurements of litterfall, specific leaf area (SLA), and canopy light interception [*Norby et al*., 2003; *McCarthy et al*., 2007]. At Duke, the native hardwood understory contributed a few emergent trees to the canopy that contributed substantially (~50% peak LAI) to the stand leaf area [*McCarthy et al*., 2007].

Transpiration data derived from sap flow observations were available at Duke from 1998 to 2007 and for the years 1999, 2004, 2007, and 2008 at Oak Ridge. At Duke (all years), the thermal dissipation probe (TDP) technique [*Granier*, 1987] was used to measure sap flow in up to eight loblolly pine trees and four sweet gum trees per plot [*Schäfer et al*., 2002; *Ward et al*., 2013]. Duke TDP probes were vertically spaced 10 cm apart and at two depths into the sapwood (0–2 or 2–4 cm) to estimate radial sap flux. At Oak Ridge, the compensated heat-pulse technique was used in 1999 and 2004 to measure hourly sap flow at 1.3 m height and 19 mm depth for four sweet gum trees in each of two ambient and elevated CO_2 plots [*Wullschleger and Norby*, 2001]. At Oak Ridge (2007 and 2008) TDP probes were installed at 1.3 m in up to five trees in each plot [*Warren et al*., 2011a, 2011b]. Probes were spaced vertically 5 cm apart, and installed at 1.5, 2.5, and 7.0 cm depths to estimate radial sap flux. Sap flow was scaled to total tree transpiration based on radial patterns of sap flow and sapwood depth considering potential error from a variety of sources [e.g., *Ewers and Oren*, 2000].

Other data collected at the sites included plant tissue N concentrations, soil water content, soil CO_2 efflux, soil carbon and nitrogen content and cycling, leaf physiology (photosynthesis and stomatal conductance), and tissue respiration, but these data sets were less comprehensive in their spatial or temporal coverage and were therefore less useful in this model-data synthesis.

2.3. The Models and Simulation Protocol

Eleven TEMs were used in the intercomparison: Community Atmosphere Biosphere Land Exchange (CABLE) [*Wang et al*., 2010, 2011], Community Land Model 4 (CLM4) [*Thornton et al*., 2007], the daily timestep version of the Century model (DAYCENT) [*Parton et al*., 2010], Ecological Assimilation of Land and Climate Observations (EALCO) [*Wang*, 2008], Ecosystem Demography 2.1 (ED2) [*Medvigy et al*., 2009], Generic Decomposition and Yield (GDAY) [*Comins and McMurtrie*, 1993], Integrated Science Assessment Model (ISAM) [*Jain and Yang*, 2005], Lund-Potsdam-Jena General Ecosystem Simulator (LPJ-GUESS) [*Smith et al*., 2001], O-CN [*Zaehle and Friend*, 2010], Sheffield Dynamic Global Vegetation Model (SDGVM) [*Woodward and Lomas*, 2004], and Terrestrial Ecosystem model (TECO) [*Weng and Luo*, 2008]. Models features are described in Table 2, and their primary functions and details are captured in a sequence of schematic images in Figure 2. Aspects of both structural similarity and diversity across the 11 models provide a good sample of ecosystem models developed over the last two decades. The models share a number of common features (Table 2), but notable differences exist that are briefly discussed in the individual model sections below.

With the exception of EALCO and ED2 (detailed in the individual model descriptions below), the simulations were initialized by spinning up the models to derive equilibrated stocks of C and N in vegetation and soils for the year 1750. Spin-ups were conducted by repeating the meteorological data recorded over the course of the experiments for 2000 years or until soil C had equilibrated. During the spin-up runs, each site was simulated assuming a deciduous broadleaf cover.

Following the spin-up, a transient "industrial" period was simulated from 1750 to the year prior to the start of the experiment. At both sites during this period several anthropogenic disturbance events and land use

Journal of Geophysical Research: Biogeosciences 10.1002/2013JG002553

Table 2. Operational, Structural and Physiological Characteristics for the 11 Models That Simulated FACE Experiments

Model	CABLE	CLM4	DAYCENT	EALCO	ED2	GDAY	ISAM	LPJ-GUESS	O-CN	SDGVM	TECO
Version	2.0	4.0	---	---	2.1	---	---	---	---	070607 modified	---
Modeler(s)	Wang, Y.-P.	Thornton	Parton	Wang, S.	Dietze	De Kauwe Medlyn	Jain	Wårlind Hickler	Zaehle	Walker	Weng Luo
Model Structure and Concept											
Initial conditions	Spun up	Spun up	Spun up	Specified and Partial spin up	Specified and Partial spin up	Spun up	Spun up	Spun up	Spun up	Spun up	Spun up
Dynamic vegetation	No	No	No	No	Age structure	No	No	Age structure	None	Age structure	None
N mass balance	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes
Leaf N	Predicted	Observed	Predicted	Predicted	Observed	Predicted	Observed	Predicted	Predicted	Observed	Predicted
Roots	Distributed	Distributed	Distributed	Distributed	Distributed	No	No	Distributed	Distributed	Distributed	Distributed
Carbon (and N) Pools											
Biomass pools[a]	L, W, R	L, SW, HW, CR (L and D) FR, S	L, B, W, CR, FR	L, SW, HW, CR, FR	L, WS, WH, R, S	L, B, SH, HW, R	L, W, R, LG, RG	L, WS, HW, FR	L, W, FR, F, S	L, SW, HW, CR, S	L, W, FR
C partitioning	Fixed fractions, phased phenology	Fixed fractions	Pipe model, resource capture, phased phenology	Fixed fractions	Pipe model and resource capture	Fixed fractions	Fixed fractions	Pipe model and resource capture	Pipe model and resource capture	Optimal leaf and fixed fractions	Fixed fractions, phased phenology
Litter soil C pools	3	4	2	3	1	4	4	2	6	4	1
SOM soil C pools	3	4	3	3	3	3	4	3	4	4	By layer
Physiological Characteristics[b]											
Photosynthesis	Farquhar N, SW	Collatz N, SW	empirical GPP = 2*NPP	Farquhar, leaf Ψ	Farquhar N	Farquhar N, SW	Farquhar	Collatz N	Friend and Kiang N, SW	Farquhar N, SW	Farquhar N
Canopy layers	1	1	1	10	10+	1	1	10+	10	LAI	10+
Sun and shade leaves	Yes	Yes	No	Yes	No	No	Yes	No	Yes	Yes	Yes
Canopy N concentration gradient	No	No	No	No	No	Assumed	No	Yes	Empirical	Beer's law	No
Canopy SLA gradients	No	Yes	No	No	No	Assumed	No	No	No	No	No
Leaf phenology	GDD	GDD	Prescribed	Tuned GDD	Prescribed	GDD	Prescribed	GDD	Tuned GDD	GDD	Predicted
Growth respiration	Yes	Yes	Yes	Yes	Yes	Total response 50% GPP	No	Yes	Yes	Leaves only	Yes
Maintenance respiration	*f*(*T*, *N*)	*f*(*T*, *N*)	*f*(*T*, *N*)	*f*(*T*, mass)	*f*(*T*, *N* indirect)	-	*f*(*T*, *N*)	*f*(*T*, *N*)	*f*(*T*, *N*, labile C)	*f*(*T*, *N*, mass)	*f*(*T*, mass)

[a]L = leaf; LG = ground-level leaves; B = branch; W = wood; WS = sapwood; WH = heartwood; R = roots; RG = ground-level roots; CR = coarse roots (L – live, D – dead); FR = fine roots; S = storage; F = fruits.
[b]*T* = temperature, *N* = mass nitrogen, CN = CN ratio, SW = soil water content, C = carbon, Ψ = water potential.

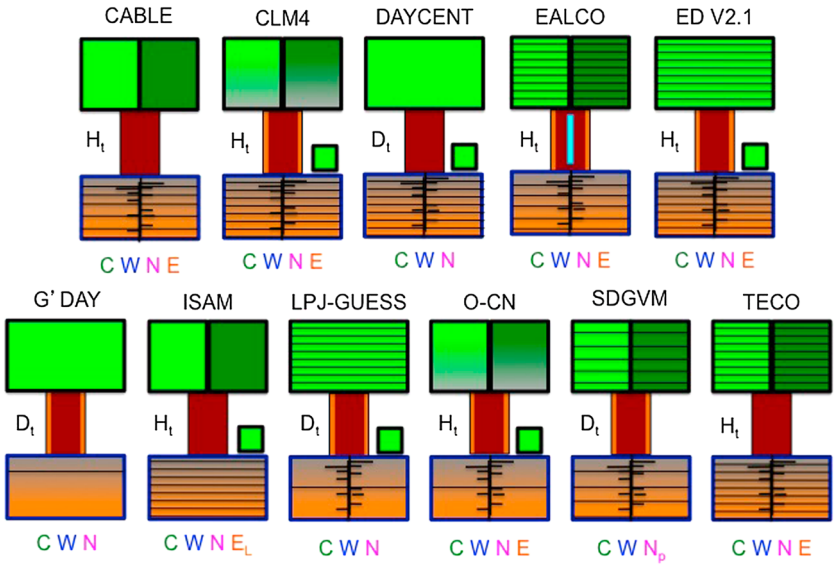

Figure 2. Structural representations of the 11 models used in the FACE model intercomparison. Canopy layering and Sun/shade assumptions are designated for each model. A single- versus dual-colored stem section indicates independent sapwood and heartwood simulations. Those models with a subcanopy "green box" include a ground-level vegetation component. Belowground details indicate soil layering and the presence or absence of roots. Belowground horizontal lines represent the approximate level of layering for that model. D_t and H_t indicate daily and hourly time steps, respectively. Letters subtending each model diagram indicate the presence of carbon (C), water (W), and nitrogen (N where Np is partial) cycles and the execution of a full energy balance (E where E_L is leaf only). EALCO includes stem water capacitance. Additional model details are itemized in Table 2.

changes occurred, first clearing the "natural" vegetation to agricultural land use and then some fallow period before site clearance and establishment of the plantations in which the experiments occurred. These site histories (detailed in Table 1) were used to initialize the modeled forest stands at the correct age and in a transient phase of ecosystem development similar to the experimental forests. Historical CO_2 driving data were taken from the record used by *Vetter et al*. [2008] created by combining the law dome ice-core record [*Etheridge et al*., 1998] with Mauna Loa and Antarctic flask measurements [*Keeling et al*., 2005]. Over the multiyear course of the experiments, daytime mean CO_2 concentrations from the experimental observations were used. N deposition data from *Dentener et al*. [2006] for the location of the experiments were in used in the transient phase of the spin-up. During the simulations of the experiments, N deposition was fixed at a rate of 13.7 kg N ha^{-1} yr^{-1} at Duke [*Sparks et al*., 2008] and 12.0 kg N ha^{-1} yr^{-1} at Oak Ridge [*Johnson et al*., 2004]. Fire was not simulated as there were no fires at either site.

All models were initialized with site data to most accurately represent conditions for the Duke and Oak Ridge FACE experiments. The initialization served to eliminate some of the parametric difference between the models and thus facilitated the model comparison. Which site data were used by each model for calibration is described in Table 3. With one exception the Duke Forest was simulated as a uniform pine plantation. ED2 was used to simulate a combined pine and hardwood forest because it has the capacity to do so, and it was consistent with the normal execution of that model.

The following brief descriptions are only intended to provide the reader with the general characteristics of each of the 11 models. The reader should look to the original source material for detailed descriptions of each model.

2.3.1. CABLE

The Community Atmosphere Biosphere Land Exchange (CABLE) model is the Australian community land surface model designed for coupling to a number of atmospheric models and Earth System models for air pollution forecast, numerical weather, and climate predictions [*Wang et al*., 2010, 2011]. CABLE simulates energy, water, carbon (C), nitrogen (N), and phosphorus (P) cycles in terrestrial ecosystems on a subdaily time step. CABLE was the first global model to include both N and P cycles [*Zhang et al*., 2011], although P limitation was not enabled in the current intercomparison as neither site was considered to be P limited. Disturbance is not normally simulated by CABLE.

Table 3. List of Site-Specific Parameters and Data Used to Calibrate Individual Models

Model	Site	Prescribed But Variable With Time	Leaf Traits	Physiological Traits	Stoichiometry	C Partitioning	Soil Properties
CABLE	Duke	-	-	-	-	-	-
CLM4		-	Top-of-canopy SLA, leaf lifespan	-	leaf C:N, wood C:N, leaf litter C:N	Fine root growth to leaf growth ratio, stem growth to leaf growth ratio	texture
DAYCENT		-	SLA	A_{max} (including CO_2 response), transpiration response to CO_2	-	-	texture, water holding capacity, bulk density
EALCO		-	SLA, max LAI	V_{cmax}/J_{max} to N relationship	-	Target C ratios among foliage, sapwood, and fine root biomass	texture
ED2		-	SLA	V_{cmax}	leaf C:N, wood C:N	leaf and stem allometry	texture
GDAY			Leaf lifespan (assumed the same for roots)	V_{cmax}/J_{max} to N relationship	-	NPP partitioning to leaves, wood, and fine roots	water holding capacity
ISAM		LAI (daily)	-	-	-	-	-
LPJ-GUESS		-	SLA	-	-	-	texture, water holding capacity
O-CN		-	SLA	-	-	-	texture, water holding capacity
SDGVM		Canopy N (annual)	SLA	V_{cmax}/J_{max} to N relationship	-	-	texture, water holding capacity, bulk density
TECO		-	SLA	-	-	-	water holding capacity
CABLE	Oak Ridge	-	-	-	-	-	-
CLM4		-	Top-of-canopy SLA, SLA increase rate, leaf length	-	-	Fine root growth to leaf growth ratio, coarse root growth to stem growth ratio	texture
DAYCENT		-	SLA	A_{max} (including CO_2 response), transpiration response to CO_2	-	-	texture, water holding capacity, bulk density
EALCO		-	SLA, max LAI	V_{cmax}/J_{max} to N relationship	-	Target C ratios among foliage, sapwood, and fine roots	texture
ED2		-	SLA	V_{cmax}	leaf C:N	leaf and stem allometries	texture
GDAY		-	Leaf growth and litterfall rates	V_{cmax}/J_{max} to N relationship	-	NPP partitioning to leaves, wood, and fine roots (including CO_2 response)	water holding capacity
ISAM		LAI (daily)	-	-	-	-	-
LPJ-GUESS		-	SLA	-	-	-	texture, water holding capacity
O-CN		-	SLA	-	-	-	texture, water holding capacity
SDGVM		Canopy N (annual)	SLA	V_{cmax}/J_{max} to N relationship	-	-	texture, water holding capacity, bulk density
TECO		-	SLA	-	-	-	water holding capacity

For this study, CABLE represented both the Duke and Oak Ridge Forests using the default evergreen needleleaf and deciduous broadleaf plant functional type (PFT) parameters [*Kowalczyk et al*., 2007].

2.3.2. CLM4 (Version 4.0)

The Community Land Model (CLM4) is the land surface model of the Community Earth System Model (CESM) [*Thornton et al*., 2007; *Oleson et al*., 2010], simulating energy, water, C, and N cycles, conserving both mass and energy. CLM4 was one of the first global C cycle models to include a mass-balanced N cycle [*Thornton et al*., 2007]. Normally, CLM4 simulates fire disturbance as a function of fuel availability and soil water content [*Oleson et al*., 2010], though this was turned off for these simulations.

CLM4 was spun up with default PFT parameters to set initial conditions for the site (e.g., soil C), but various parameters were updated to initialize the experimental period based on site observations. N deposition was taken from the standard CLM4 data set [*Galloway et al*., 2004]. Also, the daily fraction of mineral N lost to denitrification was changed from 0.5 to 0.1 day^{-1} resulting in a longer residence time of mineral nitrogen within the ecosystem.

2.3.3. DAYCENT

DAYCENT [*Parton et al*., 2010] is a version of the CENTURY model [*Parton et al*., 1994] with an added vegetation component operating on a daily time step. DAYCENT simulates C, N, and water cycles, typically for predicting soil organic C and trace gas fluxes under agricultural conditions. DAYCENT is a growth-centric model that does not explicitly simulate leaf physiology (e.g., photosynthesis and stomatal conductance). NPP is determined by a prescribed potential rate that is downregulated by nutrient, water, and temperature stress. NPP also increases, while transpiration decreases, as linear functions of atmospheric CO_2 concentration, the slopes of which were empirically determined using data from each site. GPP was assumed to be twice NPP and autotrophic respiration (R_a) was not simulated. Disturbance is not normally simulated by DAYCENT.

2.3.4. EALCO

The Ecological Assimilation of Land and Climate Observations (EALCO) model [*Wang et al*., 2007; *Wang*, 2008] was developed to assimilate a wide range of Earth observation data to study the impacts of environmental change on water resources and ecosystems for applications ranging from local to continental scales. EALCO simulates energy, water, C, and N cycles. A unique feature of EALCO is the dynamic coupling scheme [*Wang*, 2008] which uses nested numerical algorithms to solve the governing system of equations. Disturbance is not normally simulated by EALCO.

EALCO used a constrained spin-up whereby initial total soil C and N pools were prescribed with observations, but the pool sizes within the total pool were spun up to equilibrium based on 4 years of site meteorological data. Initial plant C and N pools were prescribed with observations.

2.3.5. ED2 (Version 2.1)

The Ecosystem Demography (ED) model [*Moorcroft et al*., 2001] is a complex ecosystem model which uses a forest structure approximation to scale individual-level ecophysiology, stand-level competition, and landscape-level stand age distributions to regional-scale dynamics. Stand structure is represented by PFT-specific, age-segregated cohorts and evolves from the competition between cohorts varying in size and stem density. ED2 uses a subdaily time step land surface scheme to simulate energy, water, C, and N cycles within each cohort [*Medvigy et al*., 2009]. ED2 represents fine-scale gap generation, fire, and land cover change [*Albani et al*., 2006] and has also been used to evaluate insect disturbance, for example, by the hemlock wooly adelgid [*Albani et al*., 2010].

At both sites ED2 was run for each sample plot, initialized with the diameter breast high (DBH) and species of each individual tree. Total soil C and N were initialized with site-level means, while a short spin-up was used to initialize variables for which no measurements were available, i.e., transient responses in nonstructural C, litter, soil temperature, and soil water content. Site data were used to adjust the southern pine and early-successional hardwood PFTs (preexisting ED PFTs) to more closely reflect loblolly pines and sweet gum. An experimental code for dynamic fine-root allocation in response to resource limitation was also employed.

2.3.6. GDAY

The Generic Decomposition and Yield (GDAY) model is a simple, daily time step ecosystem model that represents C, N, and water dynamics at the stand scale [*Comins and McMurtrie*, 1993; *Medlyn et al*., 2000]. The model was originally developed as a research tool to investigate the general behavior of CO_2 and N interactions [*Comins and McMurtrie*, 1993] and therefore does not include many of the details of the other models. The advantage of GDAY is its simplicity and tractability; the behavior of GDAY has been analyzed

theoretically and is well understood [*Kirschbaum et al*., 1994; *Mcmurtrie and Comins*, 1996]. GDAY was modified to represent deciduous phenology (described below) for this exercise. Disturbance is not normally simulated by GDAY.

2.3.7. ISAM

The Integrated Science Assessment Model (ISAM) [*Jain and Yang*, 2005] is an Earth System model that has been used to assess responses of the terrestrial biosphere to historical changes in cropland cover, and environmental change. ISAM simulates C, N [*Yang et al*., 2009], and water cycles. ISAM simulates secondary forest using a number of successional classes.

ISAM was applied to the two FACE sites using default PFT parameterizations and using prescribed daily LAI data at both sites.

2.3.8. LPJ-GUESS

The Lund-Potsdam-Jena General Ecosystem Simulator (LPJ-GUESS) [*Smith et al*., 2001] is an individual-based dynamic vegetation-ecosystem model, applicable at local to global scales. LPJ-GUESS simulates vegetation biogeography, ecosystem development, water, and C and N cycling at a daily time step. Vegetation dynamics are simulated adopting a forest gap model approach, distinguishing different size and age classes, which compete for light and resources within a patch. The N cycle has only recently been implemented and is based on the CENTURY model [*Smith et al*., 2013]. In standard LPJ-GUESS simulations, generic patch-destroying disturbances, representing for example windstorms, pest outbreaks, or harvest, are simulated stochastically with a mean disturbance interval of 100 years, while wildfire is normally modeled prognostically based on current fuel load and soil moisture.

The planting year was calibrated, such that the simulated tree height at the beginning of the experiment corresponded with the observed tree height. The number of saplings was prescribed to match the real planting density of the dominant trees. The model was run with default PFT parameters [*Ahlstrom et al*., 2012] using the shade intolerant evergreen needle-leaved PFT at Duke and the shade intolerant deciduous broad-leaved PFT at Oak Ridge National Laboratory (ORNL).

2.3.9. O-CN

O-CN [*Zaehle and Friend*, 2010] is a further development of the Organizing Carbon and Hydrology in Dynamic Ecosystems land surface model [*Krinner et al*., 2005] that was developed to simulate terrestrial-climate feedbacks within the Laboratoire de Météorologie Dynamique (LMDz) Earth system model [*Marti et al*., 2005]. O-CN simulates energy, water, C, and N cycles at an hourly time scale. Disturbance is not normally simulated by O-CN.

The model was applied in its default parameterization, with the exception that leaf turnover time was adjusted to be consistent with the observations at Duke, and the days of bud-burst and leaf senescence at the Oak Ridge sites were adjusted to match average observations.

2.3.10. SDGVM

The Sheffield Dynamic Global Vegetation Model (SDGVM) [*Woodward and Lomas*, 2004] was developed to simulate the global C cycle and global biogeography in response to climate. C and water cycles conserve mass, while canopy N is normally simulated through an empirical relationship to soil C [*Woodward et al*., 1995]. Ecologically, SDGVM simulates a dynamic vegetation age structure, and mortality occurs via self-thinning and maximum age. Fire disturbance is simulated by an empirical function of temperature and precipitation restricted by a location-specific fire return interval [*Kantzas et al*., 2013]. At these FACE sites, SDGVM was found to strongly underpredict canopy nitrogen and consequently V_{cmax}, so canopy N and the V_{cmax} to leaf N relationship was prescribed based on observed data (see supporting information). Also, photosynthetically active radiation (PAR) was strongly overpredicted by SDGVM and so SDGVM was driven with the mean of the annual PAR cycle.

2.3.11. TECO

The Terrestrial Ecosystem model (TECO) is an hourly time step ecosystem model [*Weng and Luo*, 2008], simulating water, C, and N cycles. TECO was designed to simulate C flows in response to environmental change at specific sites [*Weng and Luo*, 2008, 2011]. Disturbance is not normally simulated by the TECO model.

Where data were not available, default parameterizations for evergreen needleleaf and deciduous broadleaf PFTs were used with phenology parameterized to reproduce observations. The model was initialized with the initial C storage in the slow turnover pools (i.e., woody biomass and slow SOM) that it met the observed initial plant and soil C after 1 year.

2.4. Model Analysis

2.4.1. Goodness-of-Fit Statistics

We employ a number of metrics to assess and interpret model goodness of fit (GOF). To assess model GOF we use three metrics: model efficiency (EF), root-mean-square error (RMSE), and the coefficient of determination (r^2). Most GOF metrics include the sum of squares of the prediction error (SSPE) calculated as

$$\mathrm{SSPE} = \sum_{i=1}^{n} (o_i - p_i)^2 \tag{1}$$

where o_i is the ith observation, p_i is the ith model prediction, and n is the total number of paired observation-model comparisons. Model efficiency EF [*Nash and Sutcliffe*, 1970] is defined as

$$\mathrm{EF} = 1 - \frac{\mathrm{SSPE}}{\sum_{i=1}^{n} (o_i - \bar{o})^2} \tag{2}$$

where $\bar{o}$ is the mean of the observations across all n time steps. EF tells us how well the predictions fit the observations using the mean of the observations as a benchmark. An EF value of one represents a perfect fit, and an EF above zero indicates that the simulation model predictions are a better predictor of the observed values than the mean of the observed values. The root-mean-square error (RMSE) is calculated as

$$\mathrm{RMSE} = \frac{1}{n}(\mathrm{SSPE})^{0.5} \tag{3}$$

where n is the total number of comparisons between observed and predicted data. RMSE quantifies the mean absolute error between the predictions and observations. The coefficient of determination (r^2) measures the proportion of variance in the observations explained by the predictions, i.e., how well the variability in the observations is captured by the predictions.

The above metrics assess GOF, but they do not provide information on the sources of this error. To account for different sources of predictive error, we use Theil's inequality coefficients [*Smith and Rose*, 1995; *Paruelo et al*., 1998] to decompose the variance between the observations and the modeled data, i.e., the SSPE, into three components resulting from model bias (U_{bias}), difference from one in the slope of observed to predicted relationship (U_{slope}), and from random differences or nonlinearity in the relationship (U_{error}). U_{bias} describes the proportion of model error resulting from a bias (mean error equation (6) below) and is calculated as

$$U_{\mathrm{bias}} = \frac{n(\bar{o} - \bar{p})^2}{\mathrm{SSPE}} \tag{4}$$

where $\bar{p}$ is the mean of the predictions. U_{slope} describes the proportion of model error resulting from a difference in the slope from one, most likely due to different sensitivity of the model to environmental drivers compared to observations and is calculated as

$$U_{\mathrm{slope}} = \frac{(\beta - 1)^2 \sum_{i=1}^{n} (p_i - \bar{p})^2}{\mathrm{SSPE}} \tag{5}$$

where β is the slope of an ordinary least squares (OLS) linear regression of the observed on the predicted values. The observed (O) against predicted (P) regression slope was calculated using OLS linear regression with the "lm" function in R [*R Core Development Team*, 2011], and this tells us whether variability in predictions is greater or less sensitive to drivers of variability than the observations. U_{error} is the proportion of error assigned to random errors or nonlinear systematic errors (such as phase changes in seasonal cycles) and is calculated as the regression residual sum of squares divided by the SSPE. To assess the magnitude of model bias, we calculate the mean error (ME) between observations and predictions:

$$\mathrm{ME} = \frac{1}{n} \sum_{i=1}^{n} o_i - p_i \tag{6}$$

OLS linear regressions assume that only the dependent variable is measured with any uncertainty; consequently, observations should be regressed on predictions and regression of predicted values on

observed values results in biased coefficients [*Piñeiro et al*., 2008]. For this reason, we derive linear regressions of observed values to predicted values and use observed minus predicted values in equations (5)–(7). Unfortunately this means that interpretation of the metrics can be nonintuitive as negative model bias means that predictions are greater than observations and vice versa. And the slope coefficient of an OLS linear regression (β) describes biases in variability with a slope below one, indicating that a model is more sensitive to drivers of variability than the observations, and vice versa. Most GOF statistics do not provide any assessment of statistical confidence [*Moriasi et al*., 2007], and it is therefore difficult to assess whether GOF of one model is statistically different from another, or even if it is statistically different from the observations. Confidence intervals for EF, RMSE, and ME were calculated by bootstrapping. A distribution of values was generated by randomly resampling the data with replacement 1000 times and calculating a value for each resampled data set. The "boot" function [*Canty and Ripley*, 2012] in R [*R Core Development Team*, 2011] was used to resample the data, and confidence intervals were based on the percentiles of the bootstrapped distributions. When referred to below, statistical significance is at $P < 0.05$.

2.4.2. Structural Analysis for Interpretation of Transpiration and NPP

LAI is used to scale leaf-level calculations of water and C fluxes to the canopy in all models. LAI is simulated primarily by two processes, one that predicts the peak LAI and the other predicting the dynamics of LAI, or phenology. To analyze the model structures that determine rates of transpiration, following *Schäfer et al*. [2002] transpiration was decomposed into two components: (1) the transpiration per unit leaf area index (T/LAI) and (2) LAI. We further decomposed LAI into the peak LAI (LAI_{peak}) and the phenological state as a proportion of LAI_{peak} (LAI_{phen}):

$$T = T/LAI \,.\, LAI = T/LAI \,.\, LAI_{phen} \,.\, LAI_{peak} \quad (7)$$

The decomposition allows model transpiration to be corrected for biases in model LAI by replacing either, or both, modeled LAI_{phen} and LAI_{peak} by observed values. This simple decomposition assumes that transpiration is proportional to LAI, which is an oversimplification that breaks down when the difference between modeled and observed LAI are large. Nevertheless, this approach is useful for attributing differences among models in transpiration to differences in the LAI components. Examining differences in the assumptions that lead to modeled LAI_{peak} and LAI_{phen} then allows us to identify some of the reasons for different predictions among the models.

Following *Zaehle et al*. [2014], we also use the simple decomposition of NPP into component variables, N uptake (N_{up}) and N use efficiency (NUE). NUE is defined as $NUE = NPP/N_{up}$, such that $NPP = N_{up} \times NUE$. The decomposition separates the N constraint on NPP into the stoichiometric constraint (NUE, the N required for growth), and the N uptake constraint (the N available for growth).

2.4.2.1. Modeling Peak LAI

Peak LAI depends on leaf growth, leaf turnover (litterfall) rate, and specific leaf area (SLA). Often turnover plays a lesser role in determining peak LAI than leaf allocation and growth due to differential timing of these two processes. For some models LAI is the primary variable, and leaf allocation is adjusted to maintain or achieve a target LAI. The models in this synthesis use either fixed partitioning coefficients, allometric scaling, or an optimization approach to simulating LAI or leaf growth.

Using fixed partitioning coefficients, "fractions of the assimilated carbon [are] allocated to each organ" [*Franklin et al*., 2012]. Leaf growth is simulated by multiplying total C available for growth (i.e., NPP) by a foliage partitioning coefficient. CABLE, CLM4, and GDAY used fixed coefficients. SLA was a prescribed model parameter with the exception of CLM4 where SLA increases linearly with canopy depth leading to an exponential increase in LAI with leaf C. Both GDAY and CLM were parameterized with observed leaf partitioning coefficients. CABLE applied PFT default, fixed coefficients that varied according to phenological phase.

Allometric scaling is when "relationships between organs […] vary with individual size but not with the environment" [*Franklin et al*., 2012]. ED2, LPJ-GUESS, and O-CN all use allometric scaling, maintaining a prescribed constant LAI:sapwood area ratio at maximal foliar development, according to the pipe-model hypothesis [*Shinozaki et al*., 1964]. The pipe model implies a functional relationship whereby sapwood area must be sufficient to supply the canopy water demand resulting from a given LAI. The mass of wood needed to maintain a given sapwood area increases with tree height; therefore, the leaf mass:wood mass ratio decreases with tree height. Therefore, peak LAI is constrained by increasing wood growth demands as trees

increase in size. Using the pipe model, peak LAI is also sensitive to the rate of sapwood turnover to heartwood. ED2 simulates LAI as a function of DBH within each cohort, scaling to the stand scale by stem density and cohort land area. LPJ-GUESS parameterized LAI:sapwood area ratios with observations while O-CN used PFT defaults. These models typically also assume a functional balance between leaf and root mass, such that the requirement for uptake of water and nutrients, which scales with root mass, increases with leaf mass, and resource limitation.

SDGVM optimizes peak LAI based on the principle that over 1 year the sum of leaf C and annually integrated respiration of the lowest LAI layer should not exceed annually integrated C assimilation of that layer. DAYCENT, EALCO, and ISAM constrained peak LAI using observed values. Thus, the hypotheses used to simulate LAI include fixed partitioning coefficients of NPP based on general PFT relationships (CABLE), fixed partitioning coefficients with a linear decrease in SLA through the canopy (CLM4), the pipe model (ED2, LPJ-GUESS, and O-CN), and an optimization to maximize canopy C export (SDGVM).

2.4.2.2. Modeling LAI Phenology

Phenology is determined by the interaction of budburst, leaf growth rate, and leaf turnover rate. Budburst (or initiation of needle growth) was simulated using either an empirical calibration of phenology with climatic data or passively, i.e., no specific model structure controlling timing and the dynamics of leaf biomass are an emergent property of leaf growth and litterfall. For evergreen PFTs, CLM4, and GDAY simulated phenology passively, while LPJ-GUESS assumed no seasonal variation in LAI. For evergreen PFTs, ED2, and O-CN only simulated active timing of litterfall, rapid needle growth occurs in the spring when C assimilation increases and when the LAI:sapwood area ratio is below the target.

In contrast to simulation of evergreen PFTs, deciduous leaf C dynamics in CLM4 and GDAY were entirely controlled by phenology; all C for leaf growth comes from C stored in the previous year and the timing leaf growth and litterfall were entirely determined by active phenology routines. CABLE, CLM4, GDAY, LPJ-GUESS, and SDGVM represented budburst and senescence in seasonally deciduous PFTs (and for evergreens in CABLE and SDGVM) using algorithms calibrated with satellite-derived phenology and climate observations (variously cumulative growing degree days—GDD, day length, and soil temperature) [*White et al*., 1997; *Botta et al*., 2000; *Smith et al*., 2001; *Zhang et al*., 2004; *Picard et al*., 2005]. DAYCENT, EALCO, O-CN, and TECO calibrated a formulation of budburst with site observations based on either running mean air temperature, GDD, or GDD and soil temperature, respectively. For this study ED prescribed deciduous phenology but is more generally based on GDD. ISAM used daily LAI as a model input (repeating 2007 in 2008). TECO initiates leaf growth with stored C, then leaf C allocation is a fixed fraction of C allocated to growth, which is a function of available C and temperature, until maximum LAI is reached. TECO simulates maximum LAI as a function of tree height; therefore, peak LAI is either the achieved maximum or is an emergent property of the dynamics of leaf growth and turnover.

3. Results

3.1. Simulation of Annual NPP and Transpiration at Ambient CO_2

Model predictions of annual NPP and transpiration under ambient CO_2 are compared against measurements in Figures 3a–3d, and many models were within the bounds of observational uncertainty. Clear outliers were CLM4 and SDGVM, which strongly overpredicted NPP at both sites, ED2 which overpredicted NPP at Duke and underpredicted NPP at Oak Ridge, and ISAM which underpredicted NPP at Duke. A majority of models captured either the magnitude or the interannual variability in observed annual NPP, and EALCO and LPJ-GUESS captured both (Figures 3a and 3b). There was no evidence to suggest that NPP was better captured by models which better represent light interception and canopy scaling by numerically solving C assimilation in each of a number of canopy layers (EALCO, ED2, LPJ-GUESS, SDGVM, and TECO—Figure 2).

Several models (EALCO, ED2, GDAY, O-CN, SDGVM, and TECO) captured the decline in NPP at Oak Ridge from 2003/4 to 2007 (Figure 3b) but most of these models (with the exception of O-CN and GDAY) also showed an increase in NPP from 2007 to 2008, which was not consistent with the data. Models that show a general NPP decline but an increase in 2008 (ED2, SDGVM, and TECO) may have not simulated the strength of N limitation by the end of the experiment at Oak Ridge [*Norby et al*., 2010] or may also have missed carry-over effects from the strong drought in 2007 [*Warren et al*., 2011b].

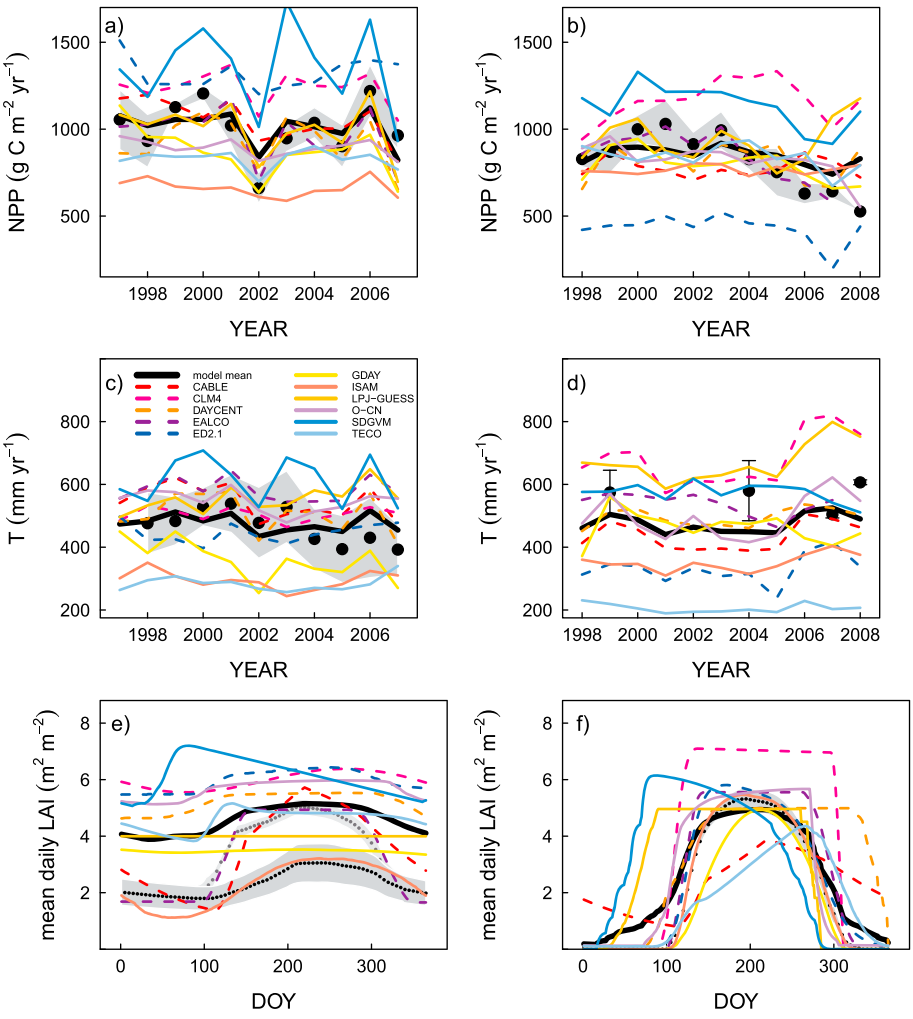

Figure 3. (a, b) Annual net primary production (NPP; g C m^{-2} y^{-1}), (c, d) annual transpiration (mm y^{-1}), and (e, f) daily mean LAI (m^2 m^{-2}) at Duke in Figures 3a, 3c, and 3e and Oak Ridge in Figures 3b, 3d, and 3f for the ambient treatments. The mean of all model results is represented by the thick black line. Model results are represented by the colored lines, observed results by the black circles with the 95% confidence interval (CI) shaded in grey. Where data are present for limited years, the 95% CIs are shown by error bars (see Figure 3d), where no error bars are visible, they are within the space occupied by the data point. Observations in Figure 3e show both the pine LAI (lighter grey shading and grey circles) and the whole stand LAI (i.e., including broad-leaved LAI). For clarity, LAI points are shown only every 5 days.

For annual transpiration at Duke (Figure 3c), most models were within the observed range of uncertainty though there was a general overprediction in the later years of the experiment. SDGVM consistently overpredicted annual transpiration, while GDAY, ISAM, and TECO consistently underpredicted annual transpiration (Figure 3c). At Oak Ridge, the mean model prediction of annual transpiration was generally low biased when compared against the 4 years of data (Figure 3d). However, the mean results from a wide range of model results.

Goodness-of-fit (GOF) statistics of model predictions with data are shown in Figures 4 (Duke) and 5 (Oak Ridge). For annual NPP, the model efficiency (EF) 95% confidence intervals (95% CIs) for most models contained zero (Figures 4a and 5a), indicating that those models were no statistically better or worse than the mean (across years) of the observations as predictors of annual NPP. That is, the models captured the mean annual NPP or some of the interannual variability but not both—to be better than the mean of the observations as a predictor of those observations might be considered a minimum standard for model performance. However, the EF confidence intervals were wide, indicating that for a given model there were large year-to-year differences in prediction accuracy. RMSE and ME CIs were also wide; however, statistical

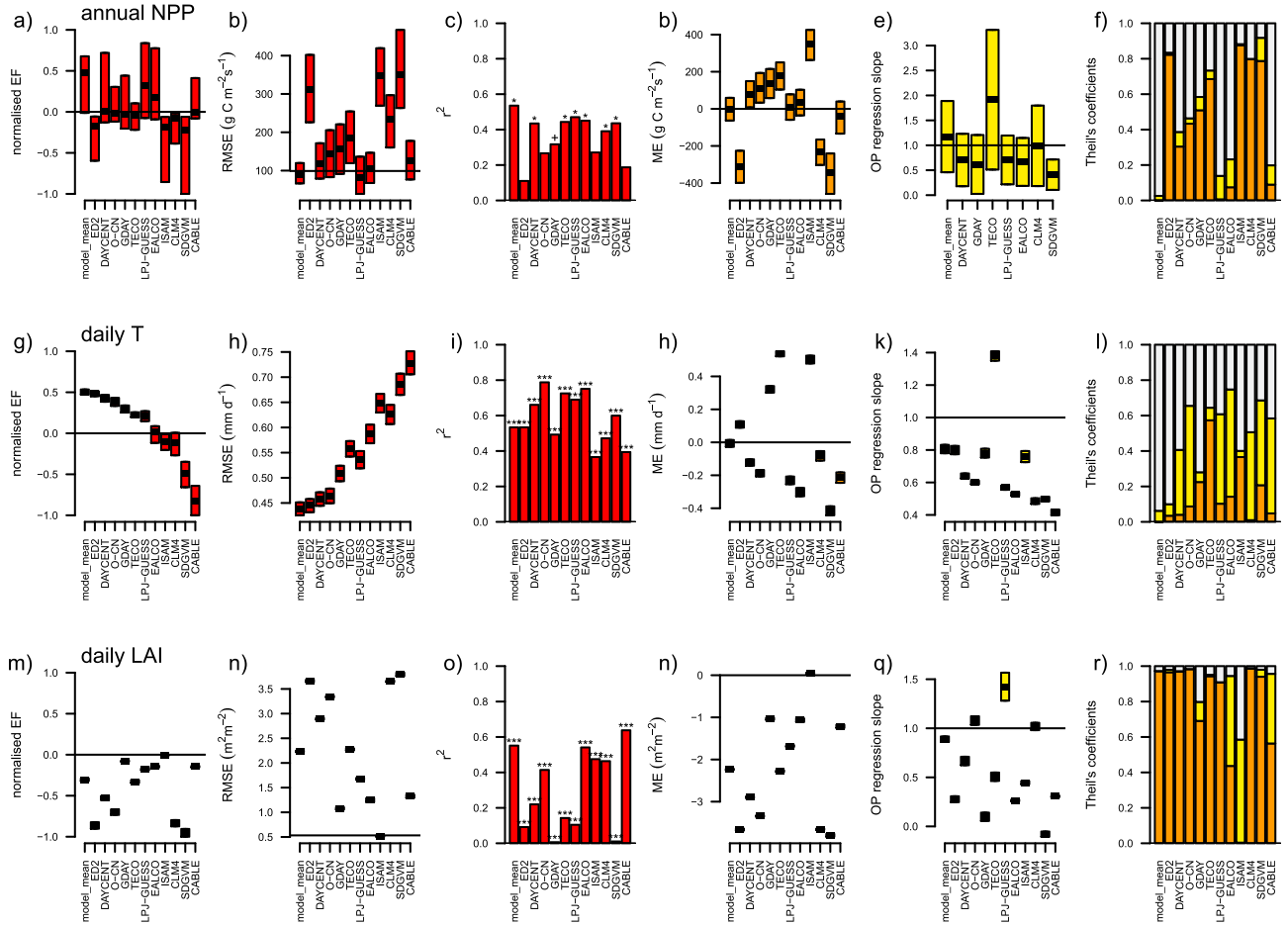

Figure 4. (a–f) Goodness-of-fit (GOF) measures for all models and multimodel mean at reproducing ambient CO_2 annual NPP, (g–l) daily transpiration, and (m–r) daily LAI (pine-only LAI) at Duke. GOF statistics are model efficiency (EF), root-mean-square error (RMSE), coefficient of determination (r^2), mean error (ME), slope bias (slope), and Theil's coefficients. Negative EF values were normalized by dividing by the most negative EF value. The horizontal lines on the RMSE plots indicate the mean over time of the observed 95% confidence interval. Note that observed values were regressed on predicted values so a positive magnitude bias indicates a negative model bias, and a sensitivity bias >1 indicates that the model was under sensitive to drivers of interannual variability in the observations. Theil's coefficients assign the proportion of model error accounted for by mean error (ME; orange), slope bias (slope; yellow), and nonsystematic error (grey).

differences between the models and observations were more apparent in these measures of error. Therefore, overall model errors in annual NPP were detectable, but 11 years of data were insufficient to accurately assess model ability to capture interannual variability in NPP using EF. Higher-frequency transpiration data were available so discussion of transpiration GOF is deferred to daily transpiration in the section below.

A notable exception from large EF uncertainty in annual NPP was EALCO at Oak Ridge which had a highly significant, high EF (Figure 5a). The spin-up method for EALCO resulted in high initial soil organic matter in rapidly cycling pools which declined over time and decreased N availability, resulting in declining NPP. The declining trend of NPP in EALCO was overlain by accurate reproduction of interannual variability; however, EALCO did not submit results for 2008 which separated models with a decreasing NPP trend into those driven by climate and those by declining N availability.

Many of the models RMSE 95% CIs were overlapping with or just outside of the 95% CI of the annual NPP observations (Figures 4b and 5b). The strong biases of CLM4, ED2, and SDGVM were reflected by significantly negative EF (Figures 4a and 5a) and large RMSE (Figures 4b and 5b). Theil's coefficients (Figures 4f and 5f) show these errors to result from large mean error (ME) biases (±300 gC m^{-2} y^{-1}—Figures 4d and 5d). At Duke, ISAM, GDAY, O-CN, and TECO also had MEs (Figures 4d and 4f) which led to poor prediction of annual NPP, reflected in high RMSEs (Figure 4b). Despite negative EF scores, CLM4 and SDGVM captured a

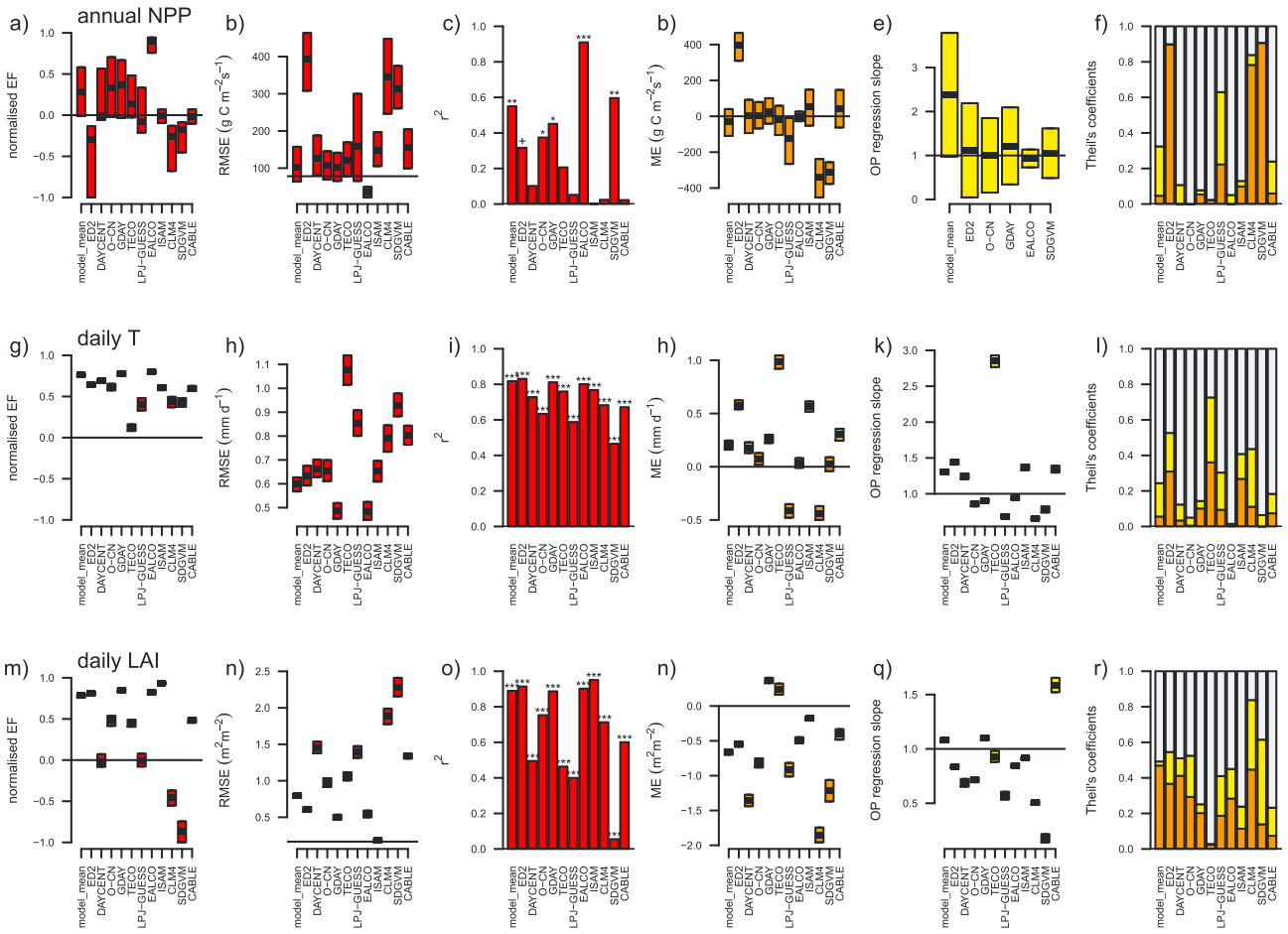

Figure 5. Goodness of fit (GOF) measures for annual NPP, daily transpiration, and daily LAI at Oak Ridge. See Figure 4 caption for details.

significant amount of interannual variability at Duke (and Oak Ridge for SDGVM), reflected in the significant ($P < 0.05$) r^2 values (Figures 5c and 6c).

NPP results from many integrated processes and can be viewed as something of a composite variable that integrates multiple processes. It is possible that compensating biases exist whereby biases of opposite sign in component processes can offset each other when combined into the composite variable which results in spuriously high GOF. Different conclusions of model skill can be drawn by analyzing some of the component processes of NPP compared to NPP alone. Figure 6 shows the decomposition of NPP into two component variables—N uptake (N_{up}) and N use efficiency (NUE)—[*Zaehle et al*., 2014] in the ambient treatment at both sites. Two of the three models that best captured NPP at Oak Ridge (EALCO and O-CN) overpredicted N_{up} and underpredicted NUE, resulting in good predictions of NPP. In other words, the models that correctly simulated the magnitude of observed NPP were doing so with compensating biases in component variables. CABLE at Duke and GDAY at Oak Ridge most accurately captured all three variables NPP, N_{up}, and NUE. C and N dynamics are the focus of another study within the FACE-MDS [*Zaehle et al*., 2014].

All of these models represent N uptake and NUE in different ways. NUE is determined by tissue C:N stoichiometry and the partitioning of new growth between various tissues with different C:N stoichiometries. Various assumptions are made on the flexibility of tissue stoichiometry, varying from fixed, to flexible within prescribed bounds, to flexible [*Zaehle et al*., 2014]. Partitioning assumptions are also diverse, as touched upon above and discussed in detail in *De Kauwe et al*. [2014]. A caveat for comparisons to observations of NUE and N uptake are that the measurements are not independent and depend upon scaling assumptions and sampling uncertainty. Nevertheless, the variability in the predictions of NUE shows that the modeling assumptions that determine NUE are crucial in predicting NPP.

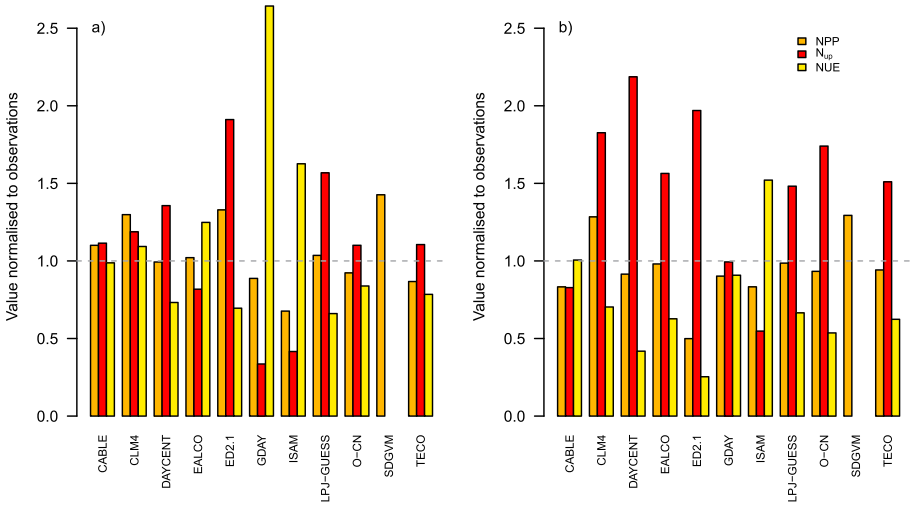

Figure 6. Simulated mean annual net primary production (NPP; orange), nitrogen uptake (N_{up}; red), and nitrogen use efficiency (NPP/N_{up}; yellow) for 1998 to 2005 (a) at Duke and for 1999 to 2008 (b) at Oak Ridge. All data are normalized by the observed mean values for the same averaging period.

3.2. Simulation of Daily Transpiration and LAI at Ambient CO_2

3.2.1. Duke Forest

At Duke, there were significant modeled errors in both the magnitude and timing of daily transpiration (Figures 4, 5, and 7 and Table 4). Many models (CABLE, CLM4, EALCO, LPJ-GUESS, O-CN, and SDGVM)

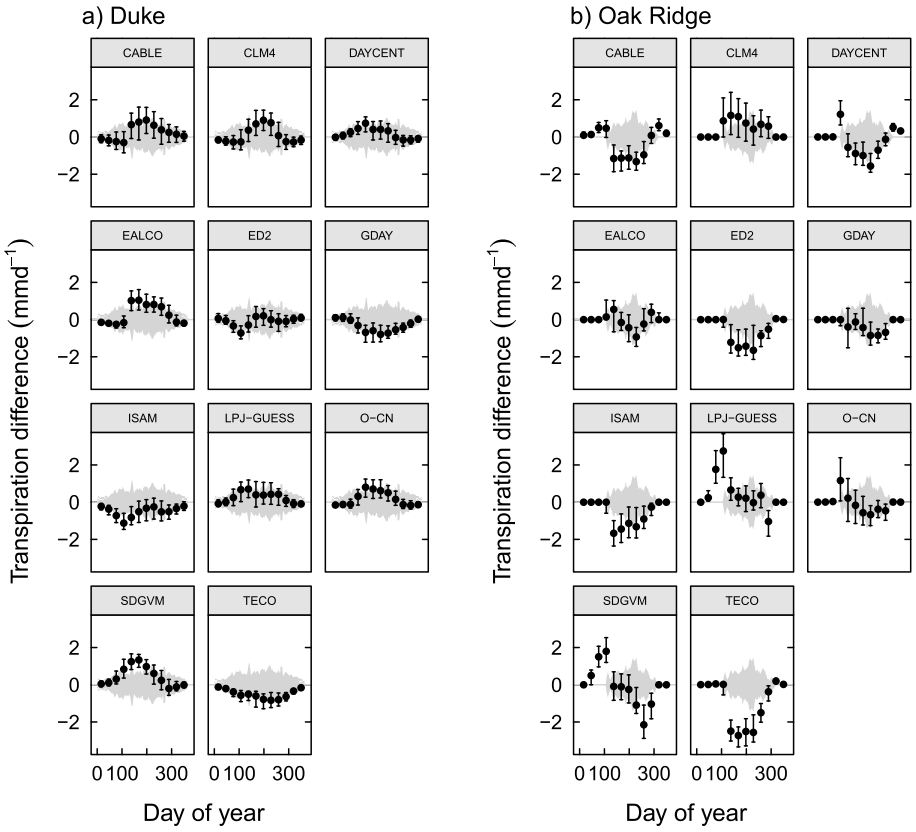

Figure 7. The difference between modeled transpiration and observed transpiration at (a) Duke (1998–2007) and (b) Oak Ridge (1999, 2004, 2007, and 2008). Monthly mean difference (points) and interquartile range of the difference (error bars). Grey-shaded area represents the 95% CI of the observations.

Table 4. Seasonal Biases in Transpiration and LAI at Duke and Oak Ridge[a]

		Duke				Oak Ridge			
Model	Variable	Spring	Summer	Autumn	Winter	Spring	Summer	Autumn	Winter
CABLE	Transpiration	•/+	+	+	•	+/−	-	−/+	+
CLM4		+	+	•	•	+	+	+	•
DAYCENT		•/+	•	•	•	+/−	-	−/+	•
EALCO		•/+	+	+	•	+	•	•/+	•
ED2		•/-	•	•	•	•/-	-	-	•
GDAY		•	-	•	•	•	-	-/•	•
ISAM		-	•	-	•	•/-	-	-/•	•
LPJ-GUESS		+	+	•/+	•	+	•	•/-	•
O-CN		•/+	+	•	•	•/+	•	•	•
SDGVM		+	+	•/+	•	+	•	-	•/+
TECO		-	-	-	•	•/-	-	-	•
CABLE	LAI	•	+	+	•/+	+/−	-	−/+	+
CLM4		+	+	+	+	•/+	+	+	•
DAYCENT		+	+	+	+	•/+	•	+	+/•
EALCO		•/+	+	•/+	•	•	•/+	+	•
ED2		+	+	+	+	•	•/+	•	•
GDAY		+	•	+	+	-	-/•	•	•
ISAM		•/-	•/-	•	•	•	•	•	•
LPJ-GUESS		+	+	+	+	+	•	•/-	•
O-CN		+	+	+	+	•/+	•	+/−	•
SDGVM		+	+	+	+	+	•	-	•/+
TECO		+	+	+	+	•/-	-	−/+	•

[a]Biases were judged based on the GOF statistics presented in Figures 4 and 5 and the comparisons to data in Figures 3e and 3f and Figures 7a and 7b. Cross symbol refers to a high bias, • no bias, and - a low bias. A combination of symbols indicates that the model exhibited a combination of those biases in that season.

overestimated transpiration over the year or during the peak season, while TECO consistently underestimated transpiration through the year, and ISAM underestimated transpiration in the spring (Table 4).

The comparisons with LAI data at Duke Forest were confounded by the broadleaf component of the forest (predominantly in the understory) that significantly contributed to the whole stand LAI (Figure 3e and see *McCarthy et al.* [2007]). Models represented the mixture of evergreens and broadleaves in multiple ways as a result of different representations of forest structural heterogeneity, and only ED2 represents vertical structure of different PFTs within the forest stand. The variability across models in representation of the PFT heterogeneity at Duke invariably confounds the comparison to observations; therefore, the discussion of transpiration in relation to LAI at Duke is deliberately limited as comparing the representations of forest structure is beyond the scope of this analysis. Interestingly ED2, which resolved vertical stand structure and individual tree distribution, overpredicted peak LAI despite detailed parameterization of stand structure, DBH, and LAI:DBH. The LAI overprediction was caused by higher than observed NPP (Figures 3a and 4d) which led to overprediction of DBH growth and thus through the DBH:LAI relationship, LAI.

In some models (CLM4, O-CN, and SDGVM) the high summer transpiration biases at Duke (Figure 7a) were matched by high summer biases in stand LAI (Figure 3e). The remaining models with high summer transpiration bias (CABLE, EALCO and LPJ-GUESS) also had high LAI biases when compared to the pine LAI only (Figure 3e), suggesting that perhaps the pines, which comprised most of the canopy leaf area (i.e., not including the understory), were the primary contributors to stand transpiration. None of the models (ED2, GDAY, ISAM, and TECO; Table 4) with low biased daily transpiration showed a low LAI bias. By contrast, all models that showed high biases in transpiration in spring, summer, and autumn (CABLE, CLM4, DAYCENT, LPJ-GUESS, O-CN, and SDGVM) also exhibited high biases in LAI.

3.2.2. Oak Ridge Forest

Many models at Oak Ridge underpredicted daily transpiration over the whole season or during the peak season (see CABLE, DAYCENT, ED2, GDAY, ISAM, and TECO in Figure 7b). TECO had a low, midseason LAI bias but in contrast with transpiration biases, LAI biases were a result of non-systematic error caused by a general low bias, a shifted peak LAI followed by a late season high bias. For DAYCENT the low mid-season

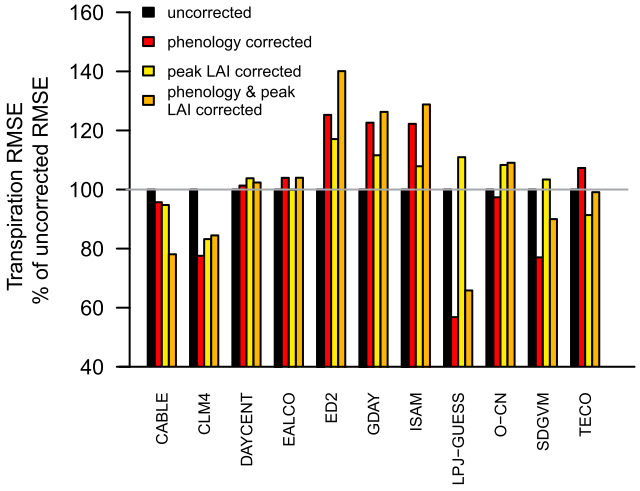

Figure 8. Root-mean-square error in daily transpiration predictions at Oak Ridge, uncorrected (black bars), corrected for phenology bias (red bars), corrected for maximum LAI bias (yellow bars), and corrected for both phenology and maximum LAI (orange bars). RMSE values are normalized to the uncorrected value.

transpiration bias was not matched by a low mid-season LAI bias (Figure 3f). With both positive and negative transpiration biases (Figure 7b), DAYCENT transpiration errors were a result of nonsystematic error (Figure 5l), while LAI biases were strongly affected by a large ME caused by very late onset of senescence (Figure 3f). In contrast to transpiration biases, LAI simulations by ISAM and ED2 had a slope bias below one and a negative ME.

CLM4 was the only model to simulate a high transpiration bias over the whole season at Oak Ridge (Table 4 and Figure 7b) and CLM4 showed the strongest LAI bias of all models (Figure 3f). CLM4 maintained a significant LAI bias in the midseason (Figures 3f, 5p and 5r) while transpiration biases were reduced (Figure 7b), CLM4 also had higher transpiration errors during the early season and showed early initiation of budburst. CABLE simulations of transpiration and LAI were high biased throughout the winter season (Table 4).CABLE, CLM4, DAYCENT, GDAY, LPJ-GUESS, O-CN, and SDGVM showed significant early- and late-season errors in transpiration at Oak Ridge (Figure 7b). Theil's coefficients show that nonsystematic error was a major source of error between predictions and observations of daily transpiration in the deciduous Oak Ridge Forest (Figure 5l). Nonsystematic error includes both random error and error caused by phase or period shifts in the seasonal cycles which changes the sign of the error over the course of the year. LPJ-GUESS and SDGVM showed a clear phase shift in predicted daily transpiration demonstrated by the opposing sign of the errors in the spring and autumn, LAI predictions of these two models were also out of phase with the observations (Figure 3f). O-CN showed a smaller phase shift with errors most pronounced in the early season as were the LAI errors (Figure 3f). EALCO showed some small transpiration errors in the early and late season that were also apparent in LAI predictions.

DAYCENT showed positive transpiration errors in the early and late seasons, i.e., an increase in transpiration period, and the LAI seasonal cycle was also of increased period. However, transpiration errors were larger in the spring, while LAI error was more pronounced in the autumn. Transpiration errors in GDAY were negative in both the spring and autumn (Figure 7b), while LAI was only smaller than observed in the spring (Figure 3f). ED2 showed some early season high bias in LAI that was not matched by transpiration errors.

3.2.3. Correction of Transpiration Error for LAI Error

We used a simple conceptual model (equation (7)) as an aid to detect and quantify transpiration errors caused by biases in phenology and biases in peak LAI. The simple model assumes that the relationship of transpiration to LAI is linear, which is a simplification but one that is useful and has been used previously [e.g., *Schäfer et al*., 2002]. Because the duel (conifer plus hardwood) stand structure at Duke was not represented by the majority of models, we only apply the correction of transpiration for LAI biases (as described by equation (7)) at Oak Ridge (Figure 8). Reductions in RMSE represent improvements in the simulation of daily transpiration once corrected for the components of the LAI bias, while increases in the RMSE mean that correction for LAI biases make transpiration errors worse indicating compensating errors between simulating the biophysics of transpiration (transpiration per unit leaf area) and LAI.

At Oak Ridge, CABLE, DAYCENT, ED2, ISAM, and TECO showed negative midseason errors in transpiration. CABLE and TECO also simulated a low peak LAI, which when corrected for slightly improved simulations of transpiration (Figure 8). Although correction for peak LAI in CABLE reduced midseason transpiration errors, it also increased the winter errors; therefore, there was an interaction when correcting for both peak LAI and phenology that improved transpiration simulation by 22%. In contrast, TECO transpiration errors were

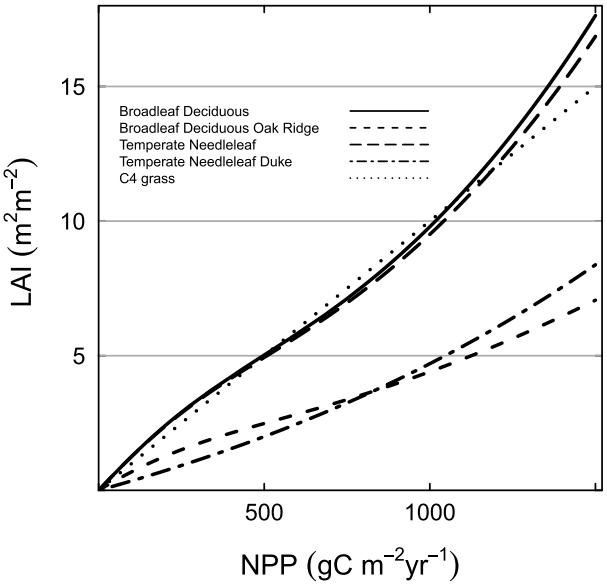

Figure 9. LAI as a function of NPP as hypothesized by CLM4, parameterized for the default and Oak Ridge broadleaf deciduous PFTs, the default and Duke temperate evergreen needleleaf PFTs, and C4 grasses.

extremely negative (Figure 7b) and while correction of peak LAI improved simulation of transpiration over the year, correction for phenology made simulations worse. So when corrections were combined there was little change in TECO transpiration error indicating an extremely low bias in simulating the biophysics of transpiration in TECO.

Correction for the small peak and phenology LAI biases in ED2, GDAY, and ISAM increased the transpiration RMSE by ~28–40% (albeit against a lower baseline RMSE, Figure 5h). Although the errors in the mean seasonal LAI cycle were small (Figure 3f), there was interannual variability in both phenology and peak that caused LAI errors (for ISAM which prescribed LAI this was because 2007 LAI was used repeated in 2008). The increase in transpiration RMSE when corrected for LAI suggests that LAI biases in ED2, GDAY, and ISAM were compensating biases in the simulation of the biophysics of transpiration.

CLM4 transpiration errors were improved once corrected for LAI phenology bias and peak LAI bias individually. However, correction for both LAI biases simultaneously did not further improve the RMSE (Figure 9), suggesting that the phenology bias (when LAI errors were largest) caused transpiration errors but that the overall high LAI bias was compensating a low bias in the simulation of transpiration per unit leaf area.

For some models with an early-season transpiration bias (CABLE, LPJ-GUESS, and SDGVM), correction of transpiration for the LAI phenology bias improved the simulation of transpiration by up to ~40% (Figure 8). LPJ-GUESS and SDGVM both showed a late-season low transpiration bias and LAI bias. In LPJ-GUESS, senescence was too rapid, while senescence in SDGVM was initiated too early (Figure 3f). The correction of transpiration for LAI biases (modeled T/LAI multiplied by observed LAI) only works if modeled transpiration is nonzero, which was not the case in the late season for these models. Therefore, accurate simulation of LAI phenology in both LPJ-GUESS and SDGVM would likely improve transpiration simulation even more that suggested by the reduction in RMSE once corrected for LAI phenology bias.

Transpiration errors of DAYCENT and O-CN were not improved when corrected for phenology bias, despite positive errors in both transpiration and LAI during the spring. Both models also had negative transpiration errors during the late season (day 250–300) but positive LAI errors over the same period; therefore, early season transpiration errors were reduced by LAI phenology correction in the early season but were increased during the late season. There was little change in EALCO transpiration RMSE when corrected for LAI because both phenology and peak LAI were prescribed.

4. Discussion

Many of the models reproduced annual NPP, and daily transpiration, in ambient CO_2 conditions with a reasonable degree of accuracy. However, we have shown that for some models, accuracy in prediction of both annual NPP and daily transpiration was achieved by biases of opposing sign in component variables—NUE and N uptake for NPP and transpiration per unit LAI and LAI for transpiration—what we call compensating biases in component processes. If we had drawn conclusions of model performance based only on the GOF statistics in Figures 4 and 5, we would have missed the compensating biases and could have had false confidence in many of the models. Without analyzing N uptake and NUE, EALCO would have been considered the most accurate predictor of NPP; however, Figure 6 shows that overall GDAY was a more accurate predictor of NPP and its component variables under ambient CO_2 conditions. For a full analysis of

the compensating biases in the C and N component processes of NPP in response to elevated CO_2, see *Zaehle et al*. [2014].

The underlying principle of multimodel benchmarking is to find the "best" or most predictively accurate models given experimental data as a benchmark. Many variables are composites of other variables, for example, NPP is the product of nitrogen use efficiency and nitrogen uptake [*Zaehle et al*., 2014], or biases in V_{cmax} parameterizations can be compensating biases in the representation of photosynthesis and canopy scaling [*Bonan et al*., 2011]. Therefore, it is possible, and we have shown, that accurate GOF for one variable can result from compensating biases among component variables, variables which one may not be able to test for GOF due to unavailable data. We advise caution in interpretation of GOF metrics where potential compensating biases are impossible to assess or could be overlooked, as this could lead to false confidence in model performance.

Based on the EF statistic, as predictors of annual NPP, most of the models were statistically indistinguishable from the mean of the observations. However, uncertainty in NPP EF for most models was large indicating that 11 years of annual data were insufficient to accurately assess model ability (GOF) at simulating interannual variability. In part, the uncertainty in evaluating model performance at simulating annual NPP was due to variable model sensitivities to drought convolved with the impact of and subsequent recovery from stochastic events; such as an ice storm at Duke [*McCarthy et al*., 2006b] and a severe wind event at Oak Ridge [*Warren et al*., 2011b]. Stochastic events and subsequent recovery were not simulated by the models. There was some evidence to suggest that models captured observed reductions in NPP in response to drought (see 2002 at Duke and 2007 at Oak Ridge; Figures 3a and 3b), although there was a broad range of responses in these years across the models. Sensitivity of the models to drought results from variable representations of soil water dynamics, the form of the physiological response function, and whether C assimilation, stomatal conductance or both are affected. Modeled C and water fluxes are sensitive to the soil water stress assumptions [e.g., *De Kauwe et al*., 2013; *Powell et al*., 2013] and model-data synthesis could help to progress our understanding of how to accurately represent soil water stress. Without assessing the uncertainty in EF, had the models been ranked according to EF, we would have missed the conclusion that the uncertainty was too large to really distinguish one model from another. To avoid false confidence in models' abilities to predict an ecosystem function it is critical to fully understand the competing models' structures and the hypotheses which they represent. Often, understanding deviations of complex model results from those of a simple, tractable model can help to identify important component processes as demonstrated by the simple relationship proposed between transpiration and LAI (for another example, see *De Kauwe et al*. [2013]). Once competing hypotheses have been identified, appropriate analyses can be developed to compare competing hypotheses to the data (for an example, see *Zaehle et al*. [2014] and *Rastetter et al*. [1992]).

Further, reducing the generality of interpretations of GOF accuracy to broader regional or global scales is that models were initialized using site-specific data, including disturbance/land use history, soil characteristics, and plant traits. We do not have these data for every model grid square at larger spatial scales, abstracting most of the models from their normal mode of operation in which representative PFT traits and natural successional processes are assumed.

In this study, LAI, in particular phenology, was best simulated by models that used some form of calibration to site data, which in turn facilitated accuracy in the simulation of transpiration (Table 4 and Figures 3 and 7). However, it is not desirable nor possible to use site data to calibrate regional or global model simulations and the applicability of global trait relationships [e.g., *Kattge et al*., 2009] or simulating adaptive traits [e.g., *Pavlick et al*., 2013; *Scheiter et al*., 2013] could also be tested. Global databases of land use history are becoming available, and land use change in global C cycle simulations has been shown to reduce the twentieth century land C sink [*Gerber et al*., 2013]. While beyond the scope of this current set of simulations, a sensitivity analysis of the models used in this intercomparison to various components of site history would make a useful study.

4.1. The Relationship of LAI and Transpiration

We have shown that biases in LAI—both in the peak and phenology—result in biases in the simulation of transpiration. For example, SDGVM simulates early onset of leaf growth and senescence (Figure 3), and this leads to similar seasonal transpiration biases (Figure 7). As another example, CABLE does not explicitly represent stored C in plants; rather, the LAI over the winter months for deciduous forests represents stored C

for spring leaf growth. This formulation predictably leads to transpiration biases over the winter months in deciduous PFTs.

LAI is a key ecosystem property and model variable that scales leaf gas exchange to the canopy and determines light capture (absorbed photosynthetically active radiation—APAR) and the biophysical interaction with the atmosphere [*Richardson et al*., 2010, 2013]. Correction of LAI biases will lead to improved simulation of transpiration (Figure 8) which in models (e.g., CABLE) coupled to atmospheric circulation models will improve the representation of the biophysical feedback of the land surface on the climate system. However, we have also demonstrated that some models had compensating biases in the simulation of transpiration, these compensating biases must be resolved before improvement LAI simulation accuracy will result in improved transpiration accuracy. In the following sections we examine the model hypotheses that determine peak LAI and LAI phenology and assess these competing hypotheses based on the experimental observations. Many models calibrated either peak LAI or LAI phenology to observations. As calibration is not a predictive method, we do not discuss these models in detail in the relevant sections below.

4.1.1. Peak LAI

For modeling peak LAI, the biological theories represented by the models are fixed partitioning coefficients and SLA (CABLE, DAYCENT, and GDAY), the pipe model (O-CN, LPJ-GUESS, and to an extent ED2), optimization for canopy C export (SDGVM), and fixed partitioning combined with a linear increase in SLA through the canopy (CLM4). Both CLM4 and SDGVM overpredicted LAI, while O-CN and LPJ-GUESS made a reasonable approximation.

CABLE, DAYCENT, and GDAY all use the partitioning coefficient approach (see above). The ability of these models to reproduce peak LAI depended on accurate simulation of NPP and whether they used site data to parameterize SLA and leaf partitioning coefficients. At Oak Ridge, GDAY simulated peak LAI close to the observations (Figure 3f) using partitioning coefficients and SLA parameterized with the site data. As turnover at the site was minimal prior to peak LAI (and did not occur until peak LAI was reached in GDAY), accurate simulation of NPP resulted in the accurate simulation of peak LAI. CABLE was a poorer predictor of peak LAI at Oak Ridge because generic plant functional type (PFT) parameters were employed.

CLM4 also uses the partitioning coefficient approach but strongly overpredicted LAI due to the assumption of a linear increase in SLA through the canopy [*Thornton and Zimmermann*, 2007]. A linear increase in SLA means that each additional LAI layer is less costly in terms of leaf C than the previous layer. Thus, as NPP increases, leaf C increases at the same rate, but LAI increases at an exponential rate (Text S1 in the supporting information). CLM4 was parameterized with site data for SLA at the top of the canopy, the rate of increase of SLA through the canopy, and leaf C partitioning. However, the overprediction of NPP led to an even greater overprediction of peak LAI.

The default mode of CLM4 increases wood partitioning as a function of NPP, and all other parameters necessary to simulate allocation and LAI are fixed so that LAI can be written as a function of NPP (Text S1). The relationship between LAI and NPP for broad classes of CLM4 PFTs is shown in Figure 9, demonstrating that overprediction of LAI is likely to be a common feature of CLM4 especially for broadleaf deciduous PFTs.

While increasing SLA with canopy depth is a valid assumption [*Norby and Iversen*, 2006; *White and Scott*, 2006; *Lloyd et al*., 2010], the formulation of CLM4 leads to overprediction of LAI within the range of forest NPP (Figure 9). Constraining SLA to a maximum value [*White and Scott*, 2006] would limit the overprediction of LAI. The pipe model imposes area:volume scaling between leaf area and sapwood volume which would also constrain LAI at high productivity, though this would require representation of tree structure in CLM4.

SDGVM strongly overpredicts peak LAI. LAI in SDGVM maximizes canopy C export by targeting an annual C balance (assimilation-respiration-leaf C construction costs) of zero in the lowest LAI layer. The optimization is sensitive to model parameterization of canopy N scaling and dark respiration, which may have been misrepresented. Also, the optimization does not account for the structural C necessary to support the lowest LAI layer (as posited by the pipe model) nor supporting root tissue [*McMurtrie and Dewar*, 2013], which could be missing costs from the optimization calculation that could help to constrain LAI.

O-CN and LPJ-GUESS employ the pipe model to simulate allocation, of which peak LAI is an emergent property, and both models accurately reproduced peak LAI at Oak Ridge. Using the pipe model, LAI is sensitive to the LAI:sapwood area ratio, sapwood turnover, and SLA. Other than SLA, both O-CN and

LPJ-GUESS used default PFT parameters. The reasonable reproduction of Oak Ridge LAI by both models suggests that with accurate simulations of NPP, the pipe model assumption is relatively robust, constraining LAI within the bounds of the observations (Figure 3f).It is well established that there is a relationship between leaf area and sapwood area as proposed by the pipe model [*Shinozaki et al*., 1964; *Oohata and Shinozaki*, 1979] although the exact nature of the relationship is still the subject of some debate [*McDowell et al*., 2002; *Schneider et al*., 2011].

4.1.2. LAI Phenology

CLM4, ED2, and GDAY all simulate phenology passively (as described in the methods) for evergreen PFTs. In GDAY passive phenology resulted in no seasonal cycle in pine LAI at Duke (Figure 3e) and restricted amplitude of the cycle for CLM4 and ED compared with observations. The restricted amplitude of CLM4, ED2, and GDAY resulted from a constant needle turnover rate, in contrast to the observations, which showed distinct seasonality with a senescent phase during the late season (Figure 3e and see *McCarthy et al*. [2007]). LPJ-GUESS assumes no variability in evergreen LAI over the year, in clear contrast to the observations. The remaining models used similar methods to simulate phenology at both sites so we discuss them in combination.

In contrast to the observations at Oak Ridge, CABLE simulates a canopy with a LAI between 1 and 2 during the winter months as a reserve to initiate physiological activity in the following spring. Predictably, the overprediction of winter LAI led to overprediction of winter transpiration (Figures 7 and 8). Prioritized partitioning of NPP in TECO at Oak Ridge caused the timing of peak LAI in TECO to be delayed until immediately before senescence. Timing of the peak LAI was delayed because the initial, rapid leaf growth phase was not sufficiently long to achieve the peak LAI. The delay was exacerbated by the default SLA in TECO being ~20% lower than observations, slowing the accumulation of LAI.

At Oak Ridge the sweet gums were of a more northerly provenance and hence had delayed budburst compared to local trees, somewhat confounding the evaluation of the timing of leaf growth. However, LPJ-GUESS (at Oak Ridge) and SDGVM (at both sites) initiated budburst extremely early in the year (Figures 3e and 3f). The SDGVM phenology formulation uses a cumulative growing degree day (GDD) formulation that was evaluated in Siberia [*Picard et al*., 2005], which may not be more generally applicable to temperate regions. GDAY and CLM4 accurately reproduced the timing of budburst at Oak Ridge using the formulation of *White et al*. [1997], most likely because the formulation was calibrated using North American data. CABLE also reproduced budburst accurately using the satellite formulation of *Zhang et al*. [2004].

CABLE, CLM4, and GDAY all accurately reproduced the timing of leaf growth initiation and senescence; however, there were notable differences in rates of leaf growth and senescence which lead to marked differences in their ability to reproduce the seasonality of LAI at Oak Ridge (Figure 3). The only difference between GDAY and CLM4 was the parameterization of the duration of the leaf growth and senescence season but had important consequences for reproducing the timing of transpiration. Timing of senescence in SDGVM occurs when leaves reach a fixed age and therefore early senescence was tied to early budburst.

At Oak Ridge most models prescribed or calibrated LAI phenology (DAYCENT, EALCO, O-CN, and TECO) in some way with observations. And if not calibrated at the site scale to empirical data, phenology schemes were calibrated at the regional or global scale. While several of the regional calibrations of satellite observations of greenness to climate data [*White et al*., 1997; *Zhang et al*., 2004] captured phenology at the two sites reasonably, such calibrations are limited in their mechanistic representation of leaf growth initiation in spring. And a large suit of models have been shown to perform poorly at a wider range of sites in North America [*Richardson et al*., 2012]. A primary reason for deciduous phenology is the avoidance of frost damage to leaves [*Woodward*, 1987] and a growing degree day (GDD) formulation of budburst assumes that the risk of frost is simply a function of GDDs. However, this assumption may not be valid in the warm humid climate of the southeastern U.S., and perhaps multiple environmental indicators may be employed to predict frost risk, and hence budburst, more accurately.

4.2. Potential Consequences for Simulating Responses to Elevated CO_2

LAI biases have been shown to be important in modeled GPP biases [*Richardson et al*., 2012; *Schaefer et al*., 2012], and LAI biases are also likely to affect the magnitude and seasonality of the GPP response to elevated CO_2 (eCO_2). Low LAI ecosystems have more opportunity to increase the fraction of absorbed

photosynthetic radiation (*f*APAR) as *f*APAR saturates as LAI increases. FACE experiments have shown that the peak LAI response [*Norby and Zak*, 2011] and the response of *f*APAR [*Norby et al*., 2005] to eCO_2 are higher in low LAI systems. Low LAI systems are also likely to have a greater response to eCO_2 [*Ewert*, 2004] as a higher fraction of the canopy is light saturated. Light-saturated photosynthesis is more sensitive to eCO_2 because at light saturation eCO_2 relieves substrate limitation of the Calvin-Benson cycle which increases the carboxylation rate of ribulose-1,5-bisphosphate carboxylase-oxygenase (RuBisCO) [*Farquhar et al*., 1980]. However, at Duke NPP responses to CO_2 were strongly related to LAI responses [*McCarthy et al*., 2006a], and more generally, NPP response to eCO_2 in low LAI systems was often directly proportional to the response of *f*APAR [*Norby et al*., 2005], suggesting that the NPP response in low LAI systems could be mostly explained by the increase in LAI.

Despite reasonable predictions of the magnitude of NPP by many models (Figures 3a and 3b), they achieved such projections with considerable inaccuracies in their underlying N cycle simulations (Figures 6c and 6d). For example, almost every model strongly overpredicted N uptake at Oak Ridge, including the model with the best reproduction of NPP (EALCO). Similar results were observed for transpiration. Correcting transpiration in GDAY, ED2, and ISAM for LAI resulted in worse simulation of transpiration which suggests that the LAI bias was compensating a transpiration bias of opposite sign.

The problems that the models have simulating ecosystem state and dynamics at ambient CO_2 are likely to affect their performance at simulating ecosystem response to elevated CO_2. N cycling interacts with C cycling and plays a major role in the response to eCO_2 of the Duke and Oak Ridge forests [*McCarthy et al*., 2010; *Norby et al*., 2010; *Drake et al*., 2011; *Garten et al*., 2011] and the models used in this study to simulate these ecosystems [*Zaehle et al*., 2014]. For example, the lower than observed NUE simulated by many of the models (Figures 4c and 4d) results in the simulated forest using and sequestering more N which may exacerbate the strength of the simulated N limitation under eCO_2 [*Zaehle et al*., 2014].

Similarly, model biases at simulating transpiration are likely to bias ecosystem responses to eCO_2 via the effect of CO_2 decreasing stomatal conductance. Decreased stomatal conductance is likely to increase soil water content which is likely to impact transpiration and primary productivity [*Schäfer et al*., 2002; *McCarthy et al*., 2010; *Morgan et al*., 2011; *De Kauwe et al*., 2013]. In summary, model development is still necessary to reduce model uncertainty in simulating ecosystem C, N, and water dynamics in ambient CO_2 conditions to provide an accurate baseline from which to make confident predictions of terrestrial ecosystem dynamics on a rising CO_2 Earth.

4.3. Concluding Remarks

Models are tools that can be used to interpret terrestrial ecosystem dynamics in response to observed or manipulated environmental change. They are also requisite tools for making projections associated with future environmental change. As a measure of the quality of these tools, we assess their ability to predict some key aspects of current terrestrial ecosystems or the biosphere—e.g., carbon fluxes and water fluxes. We have shown that model accuracy in key ecosystem properties is sometimes achieved with compensating biases which are likely to bias model predictions of ecosystem dynamics in response to environmental change. Compensating biases demonstrate that we cannot use GOF statistics alone as metrics for the predictive ability of a model. The use of GOF statistics without consideration of compensating biases could result in over confidence in a model's predictive ability which could ultimately result in misguided environmental policy.

To provide a more useful approach to interpretation of model results, we advocate the comparison of models, to each other, and with experimental data, based on their underlying hypotheses and assumptions (Figure 1). The model-data synthesis method draws multimodel intercomparisons into the scientific method of hypothesis, prediction, and experiment. Finally, we encourage a more iterative process of model intercomparison and experimental data synthesis to help identify, and therefore facilitate reductions in, uncertainty in model predictions to further our understanding of the biosphere.

Acknowledgments

This effort was conducted by the "Benchmarking ecosystem response models with experimental data from long-term CO_2 enrichment experiments" Working Group supported by National Center for Ecological Analysis and Synthesis (NCEAS; grant EF-0553768). The Oak Ridge and Duke FACE sites and additional synthesis were supported by the U.S. Department of Energy (DOE) Office of Science's Biological and Environmental Research (BER). Running the simulations was supported by funding available to the individual modeling groups. Additional support for A.P.W. was provided by a UK National Centre for Earth Observation (NCEO) sponsored PhD. M.D.K. was also supported by ARC Discovery grant DP1094791. Much of the data used in this model-data synthesis project can be found on the FACE Data Management System on the ORNL Carbon Dioxide Information Analysis Center (CDIAC) website (http://public.ornl.gov/face/), and the model data will be made available on that site in due course. Please contact the corresponding author for more information.
This manuscript has been authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

References

Ahlstrom, A., G. Schurgers, A. Arneth, and B. Smith (2012), Robustness and uncertainty in terrestrial ecosystem carbon response to CMIP5 climate change projections, *Environ. Res. Lett.*, *7*(4), doi:10.1088/1748-9326/7/4/044008.

Albani, M., D. Medvigy, G. C. Hurtt, and P. R. Moorcroft (2006), The contributions of land-use change, CO_2 fertilization, and climate variability to the eastern US carbon sink, *Global Change Biol.*, *12*(12), 2370–2390.

Albani, M., P. R. Moorcroft, A. M. Ellison, D. A. Orwig, and D. R. Foster (2010), Predicting the impact of hemlock woolly adelgid on carbon dynamics of eastern United States forests, *Can. J. For. Res.-Rev. Can. Rech. For.*, *40*(1), 119–133, doi:10.1139/X09-167.

Bonan, G. B., P. J. Lawrence, K. W. Oleson, S. Levis, M. Jung, M. Reichstein, D. M. Lawrence, and S. C. Swenson (2011), Improving canopy processes in the Community Land Model version 4 (CLM4) using global flux fields empirically inferred from FLUXNET data, *J. Geophys. Res.*, *116*, doi:10.1029/2010JG001593. [Available at: //WOS:000290933300002.]

Botta, A., N. Viovy, P. Ciais, P. Friedlingstein, and P. Monfray (2000), A global prognostic scheme of leaf onset using satellite data, *Global. Change Biol.*, *6*(7), 709–725, doi:10.1046/j.1365-2486.2000.00362.x.

Calfapietra, C., B. Gielen, M. Sabatti, P. De Angelis, G. Scarascia-Mugnozza, and R. Ceulemans (2001), Growth performance of Populus exposed to "Free air carbon dioxide enrichment" during the first growing season in the POPFACE experiment, *Ann. For. Sci.*, *58*(8), 819–828.

Canadell, J. G., C. Le Quere, M. R. Raupach, C. B. Field, E. T. Buitenhuis, P. Ciais, T. J. Conway, N. P. Gillett, R. A. Houghton, and G. Marland (2007), Contributions to accelerating atmospheric CO_2 growth from economic activity, carbon intensity, and efficiency of natural sinks, *Proc. Natl. Acad. Sci. U.S.A.*, *104*(47), 18,866–18,870.

Canty, A., and B. D. Ripley (2012), Boot: Bootstrap R (S-Plus) functions. R package version 1.3-11.

Comins, H. N., and R. E. McMurtrie (1993), Long-term response of nutrient-limited forests to CO_2 enrichment: Equilibrium behavior of plant-soil models, *Ecol. Appl.*, *3*(4), 666–681, doi:10.2307/1942099.

Cramer, W., et al. (2001), Global response of terrestrial ecosystem structure and function to CO_2 and climate change: Results from six dynamic global vegetation models, *Global Change Biol.*, *7*(4), 357–373.

De Kauwe, M. G., et al. (2013), Forest water use and water use efficiency at elevated CO_2: A model-data intercomparison at two contrasting temperate forest FACE sites, *Global Change Biol.*, *19*(6), 1759–1779, doi:10.1111/gcb.12164.

De Kauwe, M. G., et al. (2014), Where does the carbon go? A model-data intercomparison of carbon allocation at two temperate forest free-air CO_2 enrichment sites, *New Phytol.*

DeLucia, E. H., et al. (1999), Net primary production of a forest ecosystem with experimental CO_2 enrichment, *Science*, *284*(5417), 1177–1179, doi:10.1126/science.284.5417.1177.

Dentener, F., et al. (2006), Nitrogen and sulfur deposition on regional and global scales: A multimodel evaluation, *Global Biogeochem. Cycles*, *20*, GB4003, doi:10.1029/2005GB002672.

Dietze, M. C., et al. (2011), Characterizing the performance of ecosystem models across time scales: A spectral analysis of the North American Carbon Program site-level synthesis, *J. Geophys. Res.*, *116*, G04029, doi:10.1029/2011JG001661.

Drake, J. E., et al. (2011), Increases in the flux of carbon belowground stimulate nitrogen uptake and sustain the long-term enhancement of forest productivity under elevated CO(2), *Ecol. Lett.*, *14*(4), 349–357, doi:10.1111/j.1461-0248.2011.01593.x.

Etheridge, D. M., L. P. Steele, R. L. Langenfelds, R. J. Francey, J. M. Barnola, and V. I. Morgan (1998), Historical CO_2 record from the Law Dome DE08, DE08-2, and DSS ice cores, in *Trends: A Compendium of Data on Global Change*, Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, U.S. Department of Energy, Oak Ridge, Tenn. [Available at http://cdiac.ornl.gov/trends/co2/lawdome.htm.]

Ewers, B. E., and R. Oren (2000), Analyses of assumptions and errors in the calculation of stomatal conductance from sap flux measurements, *Tree Physiol.*, *20*(9), 579–589, doi:10.1093/treephys/20.9.579.

Ewert, F. (2004), Modelling plant responses to elevated CO_2: How important is leaf area index?, *Ann. Bot.*, *93*(6), 619–627, doi:10.1093/aob/mch101.

Farquhar, G. D., S. V. Caemmerer, and J. A. Berry (1980), A biochemical-model of photosynthetic CO_2 assimilation in leaves of C-3 species, *Planta*, *149*(1), 78–90.

Forster, P. M., T. Andrews, P. Good, J. M. Gregory, L. S. Jackson, and M. Zelinka (2013), Evaluating adjusted forcing and model spread for historical and future scenarios in the CMIP5 generation of climate models, *J. Geophys. Res. Atmos.*, *118*, 1139–1150, doi:10.1002/jgrd.50174.

Franklin, O., J. Johansson, R. C. Dewar, U. Dieckmann, R. E. McMurtrie, Å. Brännström, and R. Dybzinski (2012), Modeling carbon allocation in trees: A search for principles, *Tree Physiol.*, *32*(6), 648–666, doi:10.1093/treephys/tpr138.

Friedlingstein, P., et al. (2006), Climate-carbon cycle feedback analysis: Results from the (CMIP)-M-4 model intercomparison, *J. Clim.*, *19*(14), 3337–3353.

Galloway, J. N., et al. (2004), Nitrogen cycles: Past, present, and future, *Biogeochemistry*, *70*(2), 153–226, doi:10.1007/s10533-004-0370-0.

Garten, C. T., C. M. Iversen, and R. J. Norby (2011), Litterfall (15)N abundance indicates declining soil nitrogen availability in a free-air CO(2) enrichment experiment, *Ecology*, *92*(1), 133–139.

Gerber, S., L. O. Hedin, S. G. Keel, S. W. Pacala, and E. Shevliakova (2013), Land-use change and nitrogen feedbacks constrain the trajectory of the land carbon sink, *Geophys. Res. Lett.*, *40*, 5281–5222, doi:10.1002/grl.50957.

Granier, A. (1987), Evaluation of transpiration in a Douglas-fir stand by means of sap flow measurements, *Tree Physiol.*, *3*(4), 309–320, doi:10.1093/treephys/3.4.309.

Hanson, P. J., et al. (2004), Oak forest carbon and water simulations: Model intercomparisons and evaluations against independent data, *Ecol. Monogr.*, *74*(3), 443–489, doi:10.1890/03-4049.

Hendrey, G. R., D. S. Ellsworth, K. F. Lewin, and J. Nagy (1999), A free-air enrichment system for exposing tall forest vegetation to elevated atmospheric CO_2, *Global Change Biol.*, *5*(3), 293–309, doi:10.1046/j.1365-2486.1999.00228.x.

Iversen, C. M., J. Ledford, and R. J. Norby (2008), CO_2 enrichment increases carbon and nitrogen input from fine roots in a deciduous forest, *New Phytol.*, *179*(3), 837–847, doi:10.2307/25150505.

Iversen, C. M., T. D. Hooker, A. T. Classen, and R. J. Norby (2011), Net mineralization of N at deeper soil depths as a potential mechanism for sustained forest production under elevated [CO_2], *Glob. Change Biol.*, *17*(2), 1130–1139, doi:10.1111/j.1365-2486.2010.02240.x.

Jain, A. K., and X. J. 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. [Available at: //WOS:000229145600003.]

Johnson, D. W. (2006), Progressive N limitation in forests: Review and implications for long-term responses to elevated CO_2, *Ecology*, *87*(1), 64–75, doi:10.1890/04-1781.

Johnson, D. W., W. Cheng, J. D. Joslin, R. J. Norby, N. T. Edwards, and D. E. T. Jr (2004), Effects of elevated CO_2 on nutrient cycling in a sweetgum plantation, *Biogeochemistry*, *69*(3), 379–403, doi:10.2307/1469835.

Kantzas, E., M. Lomas, and S. Quegan (2013), Fire at high latitudes: Data-model comparisons and their consequences, *Global Biogeochem. Cycles*, *27*, 677–691, doi:10.1002/gbc.20059.

Kattge, J., W. Knorr, T. Raddatz, and C. Wirth (2009), Quantifying photosynthetic capacity and its relationship to leaf nitrogen content for global-scale terrestrial biosphere models, *Global Change Biol.*, *15*(4), 976–991, doi:10.1111/j.1365-2486.2008.01744.x.

Keeling, C. D., S. C. Piper, R. B. Bacastow, M. Wahlen, T. Whorf, M. Heimann, and H. A. Meijer (2005), Atmospheric CO_2 and $13CO_2$ exchange with the terrestrial biosphere and oceans from 1978 to 2000: Observations and carbon cycle implications, in *A History of Atmospheric CO_2 and Its Effects on Plants, Animals, and Ecosystems*, edited by J. R. Ehleringer, T. E. Cerling, and M. D. Dearing, pp. 83–113, Springer Verlag, New York.

Keenan, T. F., et al. (2012), Terrestrial biosphere model performance for inter-annual variability of land-atmosphere CO_2 exchange, *Global Change Biol.*, *18*(6), 1971–1987, doi:10.1111/j.1365-2486.2012.02678.x.

Kirschbaum, M. U. F., et al. (1994), Modelling forest response to increasing CO_2 concentration under nutrient-limited conditions, *Plant Cell Environ.*, *17*(10), 1081–1099, doi:10.1111/j.1365-3040.1994.tb02007.x.

Körner, C., R. Asshoff, O. Bignucolo, S. Hättenschwiler, S. G. Keel, S. Peláez-Riedl, S. Pepin, R. T. W. Siegwolf, and G. Zotz (2005), Carbon flux and growth in mature deciduous forest trees exposed to elevated CO_2, *Science*, *309*(5739), 1360–1362, doi:10.1126/science.1113977.

Kowalczyk, E. A., Y.-P. Wang, R. M. Law, H. L. Davies, and J. L. McGregor (2007), The CSIRO Atmosphere Biosphere Land Exchange (CABLE) model for use in climate models and as an offline model, CSIRO Marine and Atmospheric Research Technical paper, Aspendale, Victoria, Aust.

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. [Available at: //000227525800001.]

Le Quere, C., et al. (2009), Trends in the sources and sinks of carbon dioxide, *Nat. Geosci.*, *2*(12), 831–836, doi:10.1038/ngeo689.

Lloyd, J., et al. (2010), Optimisation of photosynthetic carbon gain and within-canopy gradients of associated foliar traits for Amazon forest trees, *Biogeosciences*, *7*(6), 1833–1859, doi:10.5194/bg-7-1833-2010.

Marti, O., P. Braconnot, and J. Bellier (2005), The new IPSL climate system model: IPSL-CM4., IPSL, Paris, France.

McCarthy, H. R., R. Oren, A. C. Finzi, and K. H. Johnsen (2006a), Canopy leaf area constrains $[CO_2]$-induced enhancement of productivity and partitioning among aboveground carbon pools, *Proc. Natl. Acad. Sci. U.S.A.*, *103*(51), 19,356–19,361, doi:10.1073/pnas.0609448103.

McCarthy, H. R., R. Oren, H.-S. Kim, K. H. Johnsen, C. Maier, S. G. Pritchard, and M. A. Davis (2006b), Interaction of ice storms and management practices on current carbon sequestration in forests with potential mitigation under future CO_2 atmosphere, *J. Geophys. Res.*, *111*, D15103 doi:10.1029/2005JD006428. [Available at: //WOS:000239786800002.]

McCarthy, H. R., R. Oren, A. C. Finzi, D. S. Ellsworth, H.-S. Kim, K. H. Johnsen, and B. Millar (2007), Temporal dynamics and spatial variability in the enhancement of canopy leaf area under elevated atmospheric CO_2, *Global Change Biol.*, *13*(12), 2479–2497, doi:10.1111/j.1365-2486.2007.01455.x.

McCarthy, H. R., R. Oren, K. H. Johnsen, A. Gallet-Budynek, S. G. Pritchard, C. W. Cook, S. L. LaDeau, R. B. Jackson, and A. C. Finzi (2010), Re-assessment of plant carbon dynamics at the Duke free-air CO(2) enrichment site: Interactions of atmospheric CO(2) with nitrogen and water availability over stand development, *New Phytol.*, *185*(2), 514–528, doi:10.1111/j.1469-8137.2009.03078.x.

McDowell, N., et al. (2002), The relationship between tree height and leaf area: Sapwood area ratio, *Oecologia*, *132*(1), 12–20, doi:10.1007/s00442-002-0904-x.

McMurtrie, R. E., and H. N. Comins (1996), The temporal response of forest ecosystems to doubled atmospheric CO_2 concentration, *Global Change Biol.*, *2*(1), 49–57, doi:10.1111/j.1365-2486.1996.tb00048.x.

McMurtrie, R. E., and R. C. Dewar (2013), New insights into carbon allocation by trees from the hypothesis that annual wood production is maximized, *New Phytol.*, *199*(4), 981–990, doi:10.1111/nph.12344.

Medlyn, B. E., R. E. McMurtrie, R. C. Dewar, and M. P. Jeffreys (2000), Soil processes dominate the long-term response of forest net primary productivity to increased temperature and atmospheric CO_2 concentration, *Can. J. For. Res.-Revue Canadienne De Recherche Forestiere*, *30*(6), 873–888, doi:10.1139/cjfr-30-6-873.

Medlyn, B. E., A. P. Robinson, R. Clement, and R. E. McMurtrie (2005), On the validation of models of forest CO_2 exchange using eddy covariance data: Some perils and pitfalls, *Tree Physiol.*, *25*(7), 839–857, doi:10.1093/treephys/25.7.839.

Medvigy, D., S. C. Wofsy, J. W. Munger, D. Y. Hollinger, and P. R. Moorcroft (2009), Mechanistic scaling of ecosystem function and dynamics in space and time: Ecosystem demography model version 2, *J. Geophys. Res.*, *114*, G01002, doi:10.1029/2008JG000812.

Moorcroft, P. R., G. C. Hurtt, and S. W. Pacala (2001), A method for scaling vegetation dynamics: The ecosystem demography model (ED), *Ecol. Monogr.*, *71*(4), 557–585.

Morgan, J. A., D. R. LeCain, E. Pendall, D. M. Blumenthal, B. A. Kimball, Y. Carrillo, D. G. Williams, J. Heisler-White, F. A. Dijkstra, and M. West (2011), C4 grasses prosper as carbon dioxide eliminates desiccation in warmed semi-arid grassland, *Nature*, *476*(7359), 202–205, doi:10.1038/nature10274.

Moriasi, D. N., J. G. Arnold, M. W. Van Liew, R. L. Bingner, R. D. Harmel, and T. L. Veith (2007), Model evaluation guidelines for systematic quantification of accuracy in watershed simulations, *Trans. ASABE*, *50*(3), 885–900.

Nash, J. E., and J. V. Sutcliffe (1970), River flow forecasting through conceptual models. Part I—A discussion of principles, *J. Hydrol.*, *10*(3), 282–290, doi:10.1016/0022-1694(70)90255-6.

Norby, R. J., and C. M. Iversen (2006), Nitrogen uptake, distribution, turnover, and efficiency of use in a CO_2-enriched sweetgum forest, *Ecology*, *87*(1), 5–14.

Norby, R. J., and D. R. Zak (2011), Ecological lessons from free-air CO_2 enrichment (FACE) experiments, in *Annual Review of Ecology, Evolution, and Systematics*, vol. 42, edited by D. J. Futuyma, H. B. Shaffer, and D. Simberloff, pp. 181–203, Annual Reviews, Palo Alto.

Norby, R. J., D. E. Todd, J. Fults, and D. W. Johnson (2001), Allometric determination of tree growth in a CO_2-enriched sweetgum stand, *New Phytol.*, *150*(2), 477–487, doi:10.1046/j.1469-8137.2001.00099.x.

Norby, R. J., J. D. Sholtis, C. A. Gunderson, and S. S. Jawdy (2003), Leaf dynamics of a deciduous forest canopy: No response to elevated CO_2, *Oecologia*, *136*(4), 574–584, doi:10.1007/s00442-003-1296-2.

Norby, R. J., et al. (2005), Forest response to elevated CO_2 is conserved across a broad range of productivity, *Proc. Natl. Acad. Sci. U.S.A.*, *102*(50), 18,052–18,056.

Norby, R. J., S. D. Wullschleger, P. J. Hanson, C. A. Gunderson, T. J. Tschaplinski, and J. D. Jastrow (2006), CO_2 enrichment of a deciduous forest: The Oak Ridge FACE experiment, in *Ecological Studies*, edited by J. Nosberger et al., pp. 231–251, Springer, Berlin. [Available at: //BCI:BCI200700031191.]

Norby, R. J., J. M. Warren, C. M. Iversen, B. E. Medlyn, and R. E. McMurtrie (2010), CO(2) enhancement of forest productivity constrained by limited nitrogen availability, *Proc. Natl. Acad. Sci. U.S.A.*, *107*(45), 19,368–19,373, doi:10.1073/pnas.1006463107.

Oleson, K. W., et al. (2010), *Technical Description of Version 4.0 of the Community Land Model (CLM)*, National Centre for Atmospheric Research, Boulder, Colo.

Oohata, S., and K. Shinozaki (1979), A statistical model of plant form—Further analysis of the pipe model theory, *Jpn. J. Ecol.*, *29*(4), 323–335.

Oren, R., et al. (2001), Soil fertility limits carbon sequestration by forest ecosystems in a CO_2-enriched atmosphere, *Nature*, *411*(6836), 469–472.

Parton, W., D. Ojima, C. Cole, and D. Schimel (1994), A general model for soil organic matter dynamics—Sensitivity to litter chemistry, texture and management, in *Quantitative Modeling of Soil Forming Processes*, edited by R. B. Bryant and R. W. Arnold, pp. 147–167, Soil Science Society of America, Madison.

Parton, W. J., P. J. Hanson, C. Swanston, M. Torn, S. E. Trumbore, W. Riley, and R. Kelly (2010), ForCent model development and testing using the enriched background isotope study experiment, *J. Geophys. Res.*, *115*, G04001, doi:10.1029/2009JG001193.

Paruelo, J. M., E. G. Jobbágy, O. E. Sala, W. K. Lauenroth, and I. C. Burke (1998), Functional and structural convergence of temperate grassland and shrubland ecosystems, *Ecol. Appl.*, *8*(1), 194–206, doi:10.1890/1051-0761(1998)008[0194:FASCOT]2.0.CO;2.

Pavlick, R., D. T. Drewry, K. Bohn, B. Reu, and A. Kleidon (2013), The Jena Diversity-Dynamic Global Vegetation Model (JeDi-DGVM): A diverse approach to representing terrestrial biogeography and biogeochemistry based on plant functional trade-offs, *Biogeosciences*, *10*(6), 4137–4177, doi:10.5194/bg-10-4137-2013.

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*, 2117–2132, doi:10.1111/gcb.12187.

Picard, G., S. Quegan, N. Delbart, M. R. Lomas, T. Le Toan, and F. I. Woodward (2005), Bud-burst modelling in Siberia and its impact on quantifying the carbon budget, *Global Change Biol.*, *11*(12), 2164–2176, doi:10.1111/j.1365-2486.2005.01055.x.

Piñeiro, G., S. Perelman, J. P. Guerschman, and J. M. Paruelo (2008), How to evaluate models: Observed vs. predicted or predicted vs. observed?, *Ecol. Modell.*, *216*(3–4), 316–322, doi:10.1016/j.ecolmodel.2008.05.006.

Powell, T. L., et al. (2013), Confronting model predictions of carbon fluxes with measurements of Amazon forests subjected to experimental drought, *New Phytol.*, doi:10.1111/nph.12390.

Pritchard, S. G., A. E. Strand, M. L. McCormack, M. A. Davis, and R. Oren (2008), Mycorrhizal and rhizomorph dynamics in a loblolly pine forest during 5 years of free-air-CO_2-enrichment, *Global Change Biol.*, *14*(6), 1252–1264.

R Core Development Team (2011), R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna. [Available at http://www.R-project.org.]

Rastetter, E., R. Mckane, G. Shaver, and J. Melillo (1992), Changes in C-storage by terrestrial ecosystems - how C-N interactions restrict responses to CO_2 and temperature, *Water Air. Soil Pollut.*, *64*(1–2), 327–344, doi:10.1007/BF00477109.

Richardson, A. D., et al. (2010), Influence of spring and autumn phenological transitions on forest ecosystem productivity, *Philos. Trans. R. Soc. B-Biol. Sci.*, *365*(1555), 3227–3246, doi:10.1098/rstb.2010.0102.

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, doi:10.1111/j.1365-2486.2011.02562.x.

Richardson, A. D., M. S. Carbone, T. F. Keenan, C. I. Czimczik, D. Y. Hollinger, P. Murakami, P. G. Schaberg, and X. Xu (2013), Seasonal dynamics and age of stemwood nonstructural carbohydrates in temperate forest trees, *New Phytol.*, *197*(3), 850–861, doi:10.1111/nph.12042.

Riggs, J. S., M. L. Tharp, and R. J. Norby (2009), *ORNL FACE Weather Data*., Carbon Dioxide Information Analysis Center, U.S. Department of Energy, Oak Ridge National Laboratory, Oak Ridge, Tenn. [Available at: http://cdiac.ornl.gov.]

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.

Schäfer, K. V. R., R. Oren, C. T. Lai, and G. G. Katul (2002), Hydrologic balance in an intact temperate forest ecosystem under ambient and elevated atmospheric CO_2 concentration, *Global Change Biol.*, *8*(9), 895–911, doi:10.1046/j.1365-2486.2002.00513.x.

Scheiter, S., L. Langan, and S. I. Higgins (2013), Next-generation dynamic global vegetation models: Learning from community ecology, *New Phytol.*, *198*(3), 957–969, doi:10.1111/nph.12210.

Schneider, R., F. Berninger, C.-H. Ung, A. Mäkelä, D. E. Swift, and S. Y. Zhang (2011), Within crown variation in the relationship between foliage biomass and sapwood area in jack pine, *Tree Physiol.*, *31*(1), 22–29, doi:10.1093/treephys/tpq104.

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. [Available at: //WOS:000285258300001.]

Shinozaki, K., K. Yoda, K. Hozumi, and T. Kira (1964), A quantitative analysis of plant form—The pipe model theory, 1, *Jpn. J. Ecol.*, *14*(3), 97–105.

Sitch, S., et al. (2008), Evaluation of the terrestrial carbon cycle, future plant geography and climate-carbon cycle feedbacks using five Dynamic Global Vegetation Models (DGVMs), *Global Change Biol.*, *14*(9), 2015–2039.

Smith, E. P., and K. A. Rose (1995), Model goodness-of-fit analysis using regression and related techniques, *Ecol. Modell.*, *77*(1), 49–64, doi:10.1016/0304-3800(93)E0074-D.

Smith, B., I. C. Prentice, and M. T. Sykes (2001), Representation of vegetation dynamics in the modelling of terrestrial ecosystems: Comparing two contrasting approaches within European climate space, *Global Ecol. Biogeogr.*, *10*(6), 621–637.

Smith, B., D. Wårlind, A. Arneth, T. Hickler, and P. Leadley (2013), Implications of incorporating nitrogen cycling and limitations on plant production in an individual-based dynamic vegetation model, *Biogeosci. Discuss.*, *10*, 18,613–18,685, doi:10.5194/bgd-10-18613-2013.

Sparks, J. P., J. Walker, A. Turnipseed, and A. Guenther (2008), Dry nitrogen deposition estimates over a forest experiencing free air CO_2 enrichment, *Global Change Biol.*, *14*(4), 768–781, doi:10.1111/j.1365-2486.2007.01526.x.

Thornton, P. E., and N. E. Zimmermann (2007), An improved canopy integration scheme for a land surface model with prognostic canopy structure, *J. Clim.*, *20*(15), 3902–3923, doi:10.1175/jcli4222.1.

Thornton, P. E., J.-F. Lamarque, N. A. Rosenbloom, and N. M. Mahowald (2007), Influence of carbon-nitrogen cycle coupling on land model response to CO(2) fertilization and climate variability, *Global Biogeochem. Cycles*, *21*, GB4018, doi:10.1029/2006GB002868. [Available at: //WOS:000251690600001.]

Vetter, M., et al. (2008), Analyzing the causes and spatial pattern of the European 2003 carbon flux anomaly using seven models, *Biogeosciences*, *5*(2), 561–583, doi:10.5194/bg-5-561-2008.

Wang, S. (2008), Simulation of evapotranspiration and its response to plant water and CO_2 transfer dynamics, *J. Hydrometeorol.*, *9*(3), 426–443, doi:10.1175/2007JHM918.1.

Wang, S., A. P. Trishchenko, and X. Sun (2007), Simulation of canopy radiation transfer and surface albedo in the EALCO model, *Clim. Dyn.*, *29*(6), 615–632, doi:10.1007/s00382-007-0252-y.

Wang, Y. P., R. M. Law, and B. Pak (2010), A global model of carbon, nitrogen and phosphorus cycles for the terrestrial biosphere, *Biogeosciences*, *7*(7), 2261–2282, doi:10.5194/bg-7-2261-2010.

Wang, Y. P., E. Kowalczyk, R. Leuning, G. Abramowitz, M. R. Raupach, B. Pak, E. van Gorsel, and A. Luhar (2011), Diagnosing errors in a land surface model (CABLE) in the time and frequency domains, *J. Geophys. Res.*, *116*, G01034, doi:10.1029/2010JG001385.

Ward, E. J., D. M. Bell, J. S. Clark, and R. Oren (2013), Hydraulic time constants for transpiration of loblolly pine at a free-air carbon dioxide enrichment site, *Tree Physiol.*, *33*(2), 123–134, doi:10.1093/treephys/tps114.

Warren, J. M., E. Pötzelsberger, S. D. Wullschleger, P. E. Thornton, H. Hasenauer, and R. J. Norby (2011a), Ecohydrologic impact of reduced stomatal conductance in forests exposed to elevated CO_2, *Ecohydrology*, *4*(2), 196–210, doi:10.1002/eco.173.

Warren, J. M., R. J. Norby, and S. D. Wullschleger (2011b), Elevated CO(2) enhances leaf senescence during extreme drought in a temperate forest, *Tree Physiol.*, *31*(2), 117–130, doi:10.1093/treephys/tpr002.

Weng, E., and Y. Luo (2008), Soil hydrological properties regulate grassland ecosystem responses to multifactor global change: A modeling analysis, *J. Geophys. Res.*, *113*, G03003, doi:10.1029/2007JG000539.

Weng, E., and Y. Luo (2011), Relative information contributions of model vs. data to short- and long-term forecasts of forest carbon dynamics, *Ecol. Appl.*, *21*(5), 1490–1505.

White, J. D., and N. A. Scott (2006), Specific leaf area and nitrogen distribution in New Zealand forests: Species independently respond to intercepted light, *For. Ecol. Manage.*, *226*(1–3), 319–329, doi:10.1016/j.foreco.2006.02.0001.

White, M. A., P. E. Thornton, and S. W. Running (1997), A continental phenology model for monitoring vegetation responses to interannual climatic variability, *Global Biogeochem. Cycles*, *11*(2), 217–234, doi:10.1029/97GB00330.

Woodward, F. I. (1987), *Climate and Plant Distribution*, Cambridge Univ. Press, Cambridge, U. K.

Woodward, F. I., and M. R. Lomas (2004), Vegetation dynamics - simulating responses to climatic change, *Biol. Rev.*, *79*(3), 643–670.

Woodward, F. I., T. M. Smith, and W. R. Emanuel (1995), A global land primary productivity and phytogeography model, *Global Biogeochem. Cycles*, *9*(4), 471–490, doi:10.1029/95GB02432.

Wullschleger, S. D., and R. J. Norby (2001), Sap velocity and canopy transpiration in a sweetgum stand exposed to free-air CO_2 enrichment (FACE), *New Phytol.*, *150*(2), 489–498, doi:10.1046/j.1469-8137.2001.00094.x.

Yang, X., V. Wittig, A. K. Jain, and W. Post (2009), Integration of nitrogen cycle dynamics into the Integrated Science Assessment Model for the study of terrestrial ecosystem responses to global change, *Global Biogeochem. Cycles*, *23*, GB4029, doi:10.1029/2009GB003474.

Zaehle, S., and A. D. Friend (2010), Carbon and nitrogen cycle dynamics in the O-CN land surface model: 1. Model description, site-scale evaluation, and sensitivity to parameter estimates, *Global Biogeochem. Cycles*, *24*, GB1005, doi:10.1029/2009GB003521. [Available at: //WOS:000274463400001.]

Zaehle, S., et al. (2014), Evaluation of 11 terrestrial carbon–nitrogen cycle models against observations from two temperate free-air CO_2 enrichment studies, *New Phytol.*, *202*, 803–822, doi:10.1111/nph.12697.

Zak, D. R., K. S. Pregitzer, M. E. Kubiske, and A. J. Burton (2011), Forest productivity under elevated CO_2 and O_3: Positive feedbacks to soil N cycling sustain decade-long net primary productivity enhancement by CO_2, *Ecol. Lett.*, *14*(12), 1220–1226, doi:10.1111/j.1461-0248.2011.01692.x.

Zhang, X., M. A. Friedl, C. B. Schaaf, and A. H. Strahler (2004), Climate controls on vegetation phenological patterns in northern mid- and high latitudes inferred from MODIS data, *Global Change Biol.*, *10*(7), 1133–1145, doi:10.1111/j.1529-8817.2003.00784.x.

Zhang, Q., Y. P. Wang, A. J. Pitman, and Y. J. Dai (2011), Limitations of nitrogen and phosphorous on the terrestrial carbon uptake in the 20th century, *Geophys. Res. Lett.*, *38*, L22701, doi:10.1029/2011GL049244.