Evaluation of continental carbon cycle simulations with North American flux tower observations

Citation	Brett M. Raczka, Kenneth J. Davis, Deborah Huntzinger, Ronald P. Neilson, Benjamin Poulter, Andrew D. Richardson, Jingfeng Xiao, Ian Baker, Philippe Ciais, Trevor F. Keenan, Beverly Law, Wilfred M. Post, Daniel Ricciuto, Kevin Schaefer, Hanqin Tian, Enrico Tomelleri, Hans Verbeeck, and Nicolas Viovy 2013. Evaluation of continental carbon cycle simulations with North American flux tower observations. Ecological Monographs 83:531–556. doi:10.1890/12-0893.1
DOI	10.1890/12-0893.1
Publisher	Ecological Society of America
Version	Version of Record
Citable Link	http://hdl.handle.net/1957/47060
Terms of Use	http://cdss.library.oregonstate.edu/sa-termsofuse

Ecological Monographs, 83(4), 2013, pp. 531–556

Evaluation of continental carbon cycle simulations with North American flux tower observations

Brett M. Raczka,[1,15] Kenneth J. Davis,[1] Deborah Huntzinger,[2] Ronald P. Neilson,[3] Benjamin Poulter,[4] Andrew D. Richardson,[5] Jingfeng Xiao,[6] Ian Baker,[7] Philippe Ciais,[4] Trevor F. Keenan,[5] Beverly Law,[8] Wilfred M. Post,[9] Daniel Ricciuto,[10] Kevin Schaefer,[11] Hanqin Tian,[12] Enrico Tomelleri,[13] Hans Verbeeck,[14] and Nicolas Viovy[4]

[1]*Department of Meteorology, Pennsylvania State University, 503 Walker Building, University Park, Pennsylvania 16802-5013 USA*
[2]*School of Earth Science and Environmental Sustainability, Northern Arizona University, P.O. Box 5694, Flagstaff, Arizona 86011-5694 USA*
[3]*Department of Botany and Plant Pathology, Oregon State University, 2082 Cordley Hall, Corvallis, Oregon 97331-2902 USA*
[4]*Laboratoire des Sciences du Climat et l'Environnement, LSCE CEA CNRS UVSQ, 91191 Gif Sur Yvette, France*
[5]*Department of Organismic and Evolutionary Biology, Harvard University, 22 Divinity Avenue, Cambridge, Massachusetts 02138 USA*
[6]*Earth Systems Research Center, Institute for the Study of Earth, Oceans, and Space, University of New Hampshire, 8 College Road, Durham, New Hampshire 03824-3525 USA*
[7]*Atmospheric Science Department, Colorado State University, 200 West Lake Street, Fort Collins, Colorado 80523 USA*
[8]*Department of Forest Ecosystems and Society, Oregon State University, 321 Richardson Hall, Corvallis, Oregon 97331 USA*
[9]*Earth Science Division, Oak Ridge National Laboratory, P.O. Box 2008, Oak Ridge, Tennessee 37831-6301 USA*
[10]*Environmental Sciences Division, Oak Ridge National Laboratory, P.O. Box 2008, Oak Ridge, Tennessee 37831-6301 USA*
[11]*National Snow and Ice Data Center, Cooperative Institute for Research in Environmental Sciences, 449 UCB, University of Colorado, Boulder, Colorado 80309-0449 USA*
[12]*International Center for Climate and Global Change Research, School of Forestry and Wildlife Sciences, SFWS Building, 602 Duncan Drive, Auburn University, Auburn, Alabama 36849-5418 USA*
[13]*Max Planck Institute for Biogeochemistry, Hans-Knöll-Straße 10, 07745 Jena, Germany*
[14]*Laboratory of Plant Ecology, Department of Applied Ecology and Environmental Biology, Ghent University, Coupure links 653, 9000 Ghent, Belgium*

Abstract. Terrestrial biosphere models can help identify physical processes that control carbon dynamics, including land–atmosphere CO_2 fluxes, and have great potential to predict the terrestrial ecosystem response to changing climate. The skill of models that provide continental-scale carbon flux estimates, however, remains largely untested. This paper evaluates the performance of continental-scale flux estimates from 17 models against observations from 36 North American flux towers. Fluxes extracted from regional model simulations were compared with co-located flux tower observations at monthly and annual time increments. Site-level model simulations were used to help interpret sources of the mismatch between the regional simulations and site-based observations. On average, the regional model runs overestimated the annual gross primary productivity (5%) and total respiration (15%), and they significantly underestimated the annual net carbon uptake (64%) during the time period 2000–2005. Comparison with site-level simulations implicated choices specific to regional model simulations as contributors to the gross flux biases, but not the net carbon uptake bias. The models performed the best at simulating carbon exchange at deciduous broadleaf sites, likely because a number of models used prescribed phenology to simulate seasonal fluxes. The models did not perform as well for crop, grass, and evergreen sites. The regional models matched the observations most closely in terms of seasonal correlation and seasonal magnitude of variation, but they have very little skill at interannual correlation and minimal skill at interannual magnitude of variability. The comparison of site vs. regional-level model runs demonstrated that (1) the interannual correlation is higher for site-level model runs, but the skill remains low; and (2) the underestimation of year-to-year variability for all fluxes is an inherent weakness of the models. The best-performing regional models that did not use flux tower calibration were CLM-CN, CASA-GFEDv2, and SIB3.1. Two flux tower calibrated, empirical models, EC-MOD and MOD17+, performed as well as the best process-based models. This suggests that (1) empirical, calibrated models can perform as well as complex, process-based models and (2) combining process-based model structure with relevant constraining data could significantly improve model performance.

Key words: *carbon fluxes; flux towers; model–data comparison; terrestrial biosphere models.*

Introduction

Future global climate predictions are uncertain. A significant portion of global climate prediction uncertainty stems from the inability to predict the amount of

Manuscript received 4 June 2012; revised 13 February 2013; accepted 25 February 2013; final version received 22 March 2013. Corresponding Editor (ad hoc): C. A. Williams.
[15] E-mail: bmr205@psu.edu

anthropogenic CO_2 from fossil fuel emissions that will be reabsorbed into the terrestrial Earth system (Randall et al. 2007). A significant improvement in the prediction of the terrestrial carbon cycle is necessary to develop a well-informed projection of the natural carbon cycle, and to design effective global carbon management strategies (Friedlingstein et al. 2006).

Model–data comparisons have the potential to identify terrestrial biosphere models (TBMs) that provide the most accurate portrayal of current terrestrial carbon-cycling processes. Unfortunately, the evaluation of TBM skill at a continental or regional level is limited due to a lack of observations across the same spatial domain. For example, regional atmospheric inversions can be used as a basis of comparison to TBMs (Peters et al. 2007, Rayner et al. 2008); however, the observation network that inversions rely upon remains too sparse to reliably resolve subcontinental fluxes. In addition, regional carbon flux estimates from inversions are subject to transport errors (Baker et al. 2006) and typically depend on TBM fluxes as priors. Inventory estimates of productivity and carbon stocks are another source for comparison with TBMs (Law et al. 2004, Pacala et al. 2007, Rogers et al. 2011, Hayes et al. 2012); however, the temporal resolution is too coarse to evaluate TBM skill at simulating seasonal and interannual carbon flux processes.

CO_2, water, and energy flux observations derived from eddy covariance flux towers (Baldocchi et al. 2001) offer another source of validation for TBMs. These observations have been used most commonly for the evaluation of site-level model simulations (e.g., Thornton et al. 2002, Hanson et al. 2004), but can also serve as a test for *regional model* performance (e.g., Hoffman et al. 2007, Potter et al. 2007, Randerson et al. 2009) or as a tool to evaluate regional flux maps (e.g., Ciais et al. 2005, Jung et al. 2011). Furthermore, flux tower data combined with statistical scaling approaches can produce regional flux maps (Xiao et al. 2008, 2010, Beer et al. 2010, Carvalhais et al. 2010).

More recently, the North American Carbon Program (NACP) Interim Synthesis Activity collected outputs from 34 TBMs both at the continental and site spatial scales representing the major biome types across North America. The variety of models, including data-driven, empirical, and process-based formulations, combined with flux data from 36 long-running eddy covariance flux towers make the scope of NACP Interim Synthesis Activity unprecedented. Flux towers (and biomass inventory [Hayes et al. 2012]) are valuable in that they provide a direct and independent estimate of carbon flux and a rare source of evaluation for model performance. In addition, the flux uncertainties are calculated for the integrated gross primary productivity (GPP), total ecosystem respiration (RE), and net ecosystem exchange (NEE). This model–data comparison effort was intended to diagnose the regional carbon fluxes across the continent (North American Carbon Program; information *available online*)[16] and has also led to a variety of site-level model performance evaluations. For example, Schwalm et al. (2010) examined model performance of monthly net ecosystem exchange (NEE) across gradients in dryness, seasonality, biome, site history, and model structure. They found model simulations were outside observed uncertainty, and that models performed the best for summer conditions, forested ecosystems, and with prescribed phenology. Dietze et al. (2011) studied model performance as a function of time scale and found that model errors in NEE were most pronounced during the annual, 20–120-day, and diurnal time scales. In addition, model performance was related to model time step, soil hydrology, and representation of photosynthesis and phenology. Richardson et al. (2012) found that models had an inadequate representation of phenology that led to inaccuracies in growing season timing, length, and the magnitude of photosynthesis for deciduous forests. Models performed better for evergreen forests. Schaefer et al. (2012) focused on modeled gross primary production (GPP), and found that daily averaged GPP could not be simulated within observed uncertainty. They concluded that simulated GPP could be most improved through better light-use-efficiency parameterization, representation of soil moisture, ecosystem response during dry conditions, and GPP inhibition during subfreezing conditions. Expanding the analysis to include NEE, GPP, and total ecosystem respiration (RE), Keenan et al. (2012) found that the models simulated interannual variability of each flux poorly. This finding was linked to shortcomings in the timing of spring phenology, onset of soil thaw, snowpack melting, and features due to extreme climate events.

The use of flux tower observations has also been used to evaluate *regional*-level model performance. For example, based upon four flux tower sites, Potter et al. (2007) found that NASA-CASA accurately simulated NPP at crop and deciduous sites, but not at coniferous and grassland sites. They concluded that the continental estimates of NPP for North America were unlikely to be underestimated. Based upon a larger sample of FLUXNET sites, Friend et al. (2007) identified that the carbon sink was under-predicted by the Sheffield Dynamic Global Vegetation Model. The under-prediction was attributed to the omission of disturbance in the simulation. Similarly, the Randerson et al. (2009) study found that regional simulations of CASA and CLM-CN underestimated the carbon uptake in boreal and temperate forest systems.

Our study is similar to previous regional model–data comparisons (e.g., Friend et al. 2007, Potter et al. 2007, Randerson et al. 2009), in that flux tower data are used as "ground-truth" for evaluating the model simulations. No previous study of this type, however, can rival the

[16] http://daac.ornl.gov

combination of models and flux tower observations organized within the NACP synthesis activity.

Here, we present a study that was an intersection of the NACP Site and Regional Synthesis activities and focused on the model performance of the *regional simulations,* 17 models that simulated carbon fluxes across all of North America. We used 36 flux tower observations from the NACP Interim Synthesis Activity to determine whether the regional model fluxes are consistent with the observations. Regional simulations are necessarily coarser in spatial resolution than site-level simulations because of the size of the domain (continent), input data, and computational limits associated with a multiyear simulation across that domain. Despite the obvious challenge of mismatch in spatial scales represented by the regional model runs vs. the flux tower measurements, it is reasonable to expect that these regional model runs demonstrate some consistency with 36 flux tower measurements spread across the continent. We judge consistency of model simulations with the observations based upon the gross and net fluxes in terms of magnitude, temporal and spatial correlation, magnitude of variability, seasonal timing, and shape of the seasonal cycle. Unlike any other model–data comparison to date, we used a combination of regional and site-level runs in order to assess the impact of (1) spatial mismatch; (2) model setup, including driver data, vegetation maps, model initialization choices; and (3) model structure upon the model–data misfit from the regional-level runs. Finally, we combined a suite of metrics that represent desirable model qualities and complete a model performance ranking. We evaluated performance in terms of time increment (i.e., annual, monthly), plant functional types, and model type (e.g., enzyme kinetic, light-use efficiency). We attempted to identify characteristics that are common to the best performing models. Some regional models are interpolations of flux tower data (e.g., EC-MOD [Xiao et al. 2008, 2010]), but have very simple representations of ecosystem processes. Others (ORCHIDEE [Ciais et al. 2005], CLM-CN [Thornton et al. 2009]) are more complex models with a mechanistic representation of processes such as photosynthesis, respiration, and disturbance. Though this makes the comparison sometimes difficult to interpret, it also represents a more realistic evaluation of the skill of TBMs currently employed for carbon cycle research.

Methods and Procedures

Model description and setup

The regional output of 17 TBMs (Table 1) was collected as part of the NACP Regional and Continental Interim Synthesis (RCIS; Huntzinger et al. 2012). The models vary in the level of complexity ranging from statistical representations to process-based biogeochemical descriptions of relevant ecosystem processes. The primary objective of the RCIS was to synthesize and compare TBMs to assess current understanding of the terrestrial carbon cycle in North America. Thus, the RCIS focused on "off-the-shelf" model simulations, or existing model results available from analyses that have been completed by ongoing NACP projects and other recently published studies. Consequently, the regional models used different meteorology, vegetation cover, prescribed phenology (for applicable models), and representation of disturbance (e.g., land use history, fire emissions). A description of the regional, gridded weather reanalysis, vegetation products, and disturbance for the regional models is provided in Table 1 and in Huntzinger et al. (2012). The majority of models considered here include some representation of disturbance. In these cases, the influence of disturbance on the carbon exchange is included in the modeled NEE.

The model driver data for the site-level runs were observed at the site locations and included air temperature, precipitation, wind speed, humidity, radiation, vegetation type, soil type, and elevation. For the site-level runs, gaps in the observed weather record were filled with observations either from a nearby flux tower or a National Climatic Data Center station. Additional details of the gap-filling methodology for meteorology data can be found in Ricciuto et al. (2009). Models that required prescribed phenology were provided a multi-year averaged satellite phenology product. The site-level model runs were initialized through a *spin-up* procedure that transitions an ecosystem to an equilibrium state. This was achieved by looping the weather data until the GPP and RE were nearly balanced.

A subset of the regional models, denoted by daggers (†) in Table 1, provide site-level simulations for all sites in Table 2 as part of the site synthesis. Here, these seven "crossover" models were used to help interpret the regional model results. The site-level model simulations shared a common simulation protocol, whereas the regional models did not share a consistent simulation protocol. As a result of the differences in the spatial resolution and model setup between the regional and site-level simulations, the crossover models provided an opportunity to evaluate the impact of spatial mismatch and the model setup (i.e., vegetation, climate, disturbance, initial conditions) on model performance. A listing of site vs. regional level differences amongst the crossover models is located in Appendix A: Table A1.

Flux tower observations

The flux observations for this analysis (Table 2) were obtained and processed within the NACP Site Interim Synthesis (Schwalm et al. 2010). We used observations from 36 sites across North America representing 10 different biomes for the years 2000–2005. These sites encompass a wide range of climate and vegetation types. Here, the NEE, measured from the towers, represents the difference between RE and GPP. A negative value of NEE indicates a net sink of carbon into the land. The NEE flux tower data were filtered on a site-by-site basis during low turbulence conditions to reduce uncertainties

TABLE 1. Description of regional model driver data and model formulation.

Regional models	Reference	Radiation	Temperature
BEPS†	Chen et al. (1999), Ju et al. (2006)	NCEP	NCEP
CASA-GFEDv2	van der Werf et al. (2004, 2006)	ISCCP, NCEP(R2)	IIASA, GISSTEMP
CASA-Trans	Randerson et al. (1997)	...	Leemans and Cramer (1991), Hansen et al. (1999)
CLM-CASA'	Randerson et al. (2009)	NCEP	NCEP
CLM-CN†	Thornton et al. (2009), Randerson et al. (2009)	NCEP	NCEP
Can-IBIS†	Kucharik et al. (2000), Foley et al. (1996)	CFS spatial data	CFS spatial data
DLEM†	Tian et al. (2010)	NARR	NARR, PRISM
EC-MOD	Xiao et al. (2008, 2010)	N/A	MODIS LST
ISAM†	Jain and Yang (2005), Yang et al. (2009)	N/A	Mitchell and Jones (2005)
LPJ-wsl†	Bondeau et al. (2007)	CRU05	CRU05
MC1	Bachelet et al. (2000)	N/A	PRISM
MOD17+	Beer et al. (2010)	ERA-INTERIM	ERA-INTERIM
NASA-CASA	Potter et al. (2007)	New et al. (2000)	DAYMET 1982–2000 NCEP 2001–2004
ORCHIDEE†	Krinner et al. (2005), Viovy et al. (2000)	CRU, NCEP	CRU, NCEP
SIB3.1	Baker et al. (2008)	NCEP	NCEP
TEM6	Hayes et al. (2011)	CRU, NCEP	CRU, NCEP
VEGAS2	Zeng et al. (2004, 2005)	NCEP	NASA, GISSTEMP

Note: More details of models and model driver data are provided in Huntzinger et al. (2012). Abbreviations are: EK, enzyme kinetic; LUE, light-use efficiency; DA, data assimilation; N/A, not applicable; and LAI, leaf area index. Zero-order kinetic models base the decomposition rate of soil carbon on moisture and temperature conditions only, and first-order kinetic models base the decomposition rate on moisture, temperature and the soil carbon pool; w/N stands for "with nitrogen" and indicates that the model also includes nitrogen limitation on the soil carbon decomposition rate calculation. Ellipsis indicates that data are not available.

† "Crossover" models that were run at both continental (regional) spatial domain and at individual sites.

in NEE associated with these conditions (Barr et al. 2009). Parameterized equations were used to gap-fill the NEE records, and to partition NEE into the gross fluxes of GPP and RE (Moffat et al. 2007, Desai et al. 2008, Barr et al. 2009). Strictly speaking, the GPP and RE values are not observed values, but products inferred from NEE observations. The uncertainties were calculated using a Monte Carlo approach for both monthly and annual flux increments (Barr et al. 2009). This approach accounts for the uncertainty from several sources: the low-turbulence filtering threshold (u^* threshold), the random error (Richardson and Hollinger 2007), and the gap-filling and partitioning algorithm.

Matching observations and model output in time and space

The model simulations and site observations underwent temporal and spatial aggregation to provide a uniform comparison. The site-level output and site observations were gap-filled and integrated to monthly and annual flux increments (Barr et al. 2009). Similarly, the regional model runs were aggregated to monthly and annual temporal resolution at one-degree spatial resolution (Huntzinger et al. 2012). The modeled carbon flux data were extracted from the grid cell that corresponded to the location of each flux tower site for comparison. While some of the regional models provide flux estimates at finer spatial resolution, we have chosen to evaluate all of the models at this common resolution. In this way, we evaluated all of the regional models on "equal footing" with regards to spatial resolution.

This comparison thus includes a significant mismatch in the spatial scales represented by the models ($\sim 10^4$-km^2 grid cells) and the flux tower observations ($\sim$1-km^2 flux footprint). This will degrade model performance relative to the flux tower observations, particularly in regions where climate and land cover are heterogeneous. We acknowledge that failure of the model to reproduce flux tower observations could be entirely due to this mismatch in spatial scales. Conversely, it is also possible that larger scale features that are coherent across grid cells (e.g., climate) dominate both the tower and model fluxes. We confront this issue by utilizing the crossover models to test for the influence of spatial mismatch and site-specific driver data on regional model performance.

Statistical measures

Given the many different types of models and sites within the analysis, we used the Taylor diagram (Taylor 2001), which provides a comprehensive representation of variability and correlation amongst the model output. Here, we compared modeled (f) to observed (r) fluxes of CO_2 at monthly and annual increments. Performance is determined by standard deviation (σ), Pearson correlation coefficient (R), and centered root mean-square deviation (E'). The Taylor diagram uses the biased form of σ as follows:

$$\sigma = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(f_n - \bar{f})^2} \quad (1)$$

TABLE 1. Extended.

Phenology	Photosynthesis	Soil decomposition
custom LAI	EK	first order, w/N
GIMMS NDVI-derived LAI	LUE	first order
prognostic	LUE	first order
prognostic	EK	first order
prognostic	EK	first order, w/N
prognostic	EK	first order
prognostic	EK	first order, w/N
MODIS EVI	statistical, DA	zero order
N/A	statistical	first order, w/N
prognostic	EK	first order
prognostic	statistical	first order, w/N
MODIS LAI	LUE, DA	zero order
MODIS EVI	LUE	first order, w/N
prognostic	EK	first order, w/N
GIMMSg	EK	zero order
prognostic	EK	first order, w/N
prognostic	LUE	first order

where N is the total number of data points. The correlation R is defined as

$$R = \frac{\frac{1}{N}\sum_{n=1}^{N}(f_n - \bar{f})(r_n - \bar{r})}{\sigma_f \sigma_r}. \quad (2)$$

Finally, E' is defined as

$$E' = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left((f_n - \bar{f}) - (r_n - \bar{r})\right)^2}. \quad (3)$$

We used several additional statistical criteria that are not included in the Taylor diagram. The bias ($\bar{E}$) is the difference in the average magnitude between the observations (r) and model output (f) and is defined as

$$\bar{E} = \frac{1}{N}\sum_{n=1}^{N}(r_n - f_n). \quad (4)$$

The total root mean square deviation (E) is the average deviation between each observation (r_n) and corresponding modeled value (f_n):

$$E = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(r_n - f_n)^2}. \quad (5)$$

The sigma ratio is defined as the logarithmic ratio between the modeled standard deviation and the observed standard deviation. This is the only metric that provides a direct comparison between simulated and observed magnitude of variation. A value close to zero indicates the model matches the observed standard deviation closely.

The chi-square (χ^2) statistic is a measure of how well modeled values match observed values considering observational uncertainty (ε) as defined:

$$\chi^2 = \frac{1}{N}\sum_{n=1}^{N}\left(\frac{r_n - f_n}{\varepsilon_n}\right)^2. \quad (6)$$

A chi-square <1 indicates the model matches the observations given the uncertainty inherent in the observations. The uncertainty in flux tower observations of NEE, RE, and GPP is due to random sampling error and uncertainty in filling missing observations. The random uncertainty is caused by limited sampling of the turbulence that transports CO_2 at the land–air interface (Richardson et al. 2006). Gap-filling uncertainty stems from the original measurement error, uncertainty in the low friction velocity, and the algorithms used to fill in missing data (Richardson and Hollinger 2007). A more complete discussion of the observed uncertainty calculation and filling technique is given in Barr et al. (2009).

We derived the mean monthly and annual flux observational uncertainty from the single-month and annual observational uncertainty. The "systematic" approach (worst-case scenario) assumes the single-month and annual uncertainty is entirely systematic, and therefore, the uncertainty is the same between the mean and single-month/annual fluxes. The "random" approach (best-case scenario), on the other hand, assumes the single-month/annual uncertainty is entirely random and independent; therefore, the relative uncertainty is reduced when the fluxes are averaged (Taylor 1997). The true flux uncertainty is composed of both random measurement and systematic error; therefore, the true uncertainty lies in between these estimates.

Partitioning statistics into time and space

We further diagnosed model performance by partitioning the temporal (within-site) and spatial (across-site) contributions to the annual correlation and magnitude of variation (sigma) statistics. This disaggregates the modeled year-to-year variations in flux at a single site from the modeled variations in fluxes across sites. To calculate the temporal contribution, the annual flux data were preprocessed by subtracting out the site-year mean for each of the modeled fluxes and observations. In order to calculate the spatial contribution, the site-years for each model and observational data set were averaged. The processed data in both cases then underwent the normal statistical calculation to obtain correlation and sigma (see subsection *Statistical measures*). For the monthly correlation and sigma statistics, both temporal and spatial contributions were considered simultaneously.

TABLE 2. Location and vegetation description (plant functional type [PFT]) of flux tower sites.

Site code	Reference	State/province	Latitude, longitude (°N, °W)	PFT
Ca-Ca1	Schwalm et al. (2007)	British Columbia, Canada	49.87, −125.33	ENFT
Ca-Let	Flanagan and Adkinson (2011)	Alberta, Canada	49.71, −112.94	GRASS
Ca-Mer	Lafleur et al. (2003)	Ontario, Canada	45.41, −75.52	WET (MISC)
Ca-Oas	Barr et al. (2002)	Saskatchewan, Canada	53.63, −106.20	DBF
Ca-Obs	Kljun et al. (2006)	Saskatchewan, Canada	53.99, −105.12	ENFB
US-Ha1	Urbanksi et al. (2007)	Massachusetts, USA	42.54, −72.17	DBF
US-Ho1	Richardson et al. (2009)	Maine, USA	45.20, −68.74	ENFT
US-Me2	Thomas et al. (2009)	Oregon, USA	44.45, −121.56	ENFT
US-Ne3	Suyker and Verma (2008)	Nebraska, USA	41.18, −96.44	CROP
US-UMB	Gough et al. (2008)	Michigan, USA	45.56, −84.71	DBF
US-ARM	Fischer et al. (2007)	Oklahoma, USA	36.61, −97.49	CROP
US-Ne1	Suyker and Verma (2008)	Nebraska, USA	41.17, −96.48	CROP
US-Ne2	Suyker and Verma (2008)	Nebraska, USA	41.16, −96.47	CROP
US-IB1	Allison et al. (2005)	Illinois, USA	41.86, −88.22	CROP
US-Var	Ryu et al. (2008)	California, USA	38.41, −120.95	GRASS
US-Shd	Burba and Verma (2005)	Oklahoma, USA	36.93, −96.68	GRASS
US-IB2	Matamala et al. (2008)	Illinois, USA	41.84, −88.24	CROP
US-Dk2	Pataki and Oren (2003)	North Carolina, USA	35.97, −79.10	DBF
US-MMS	Schmid et al. (2000)	Indiana, USA	39.32, −86.41	DBF
US-WCr	Cook et al. (2004)	Wisconsin, USA	45.81, −90.08	DBF
US-Moz	Gu et al. (2006)	Missouri, USA	38.74, −92.20	DBF
Ca-Man	Goulden et al. (1997)	Manitoba, Canada	55.88, −98.48	ENFB
Ca-Ojp	Kljun et al. (2006)	Saskatchewan, Canada	53.92, −104.69	ENFB
Ca-Qfo	Bergeron et al. (2007)	Quebec, Canada	49.69, −74.34	ENFB
US-Dk3	Siqueira et al. (2006)	North Carolina, USA	35.98, −79.09	ENFT
US-NR1	Bradford et al. (2008)	Colorado, USA	40.03, −105.55	ENFT
Ca-TP4	Peichl and Arain (2007)	Ontario, Canada	42.71, −80.36	ENFT
US-Pfa	Davis et al. (2003)	Wisconsin, USA	45.95, −90.27	MF (DBF)
US-Syv	Desai et al. (2005)	Michigan, USA	46.24, −89.35	MF (DBF)
Ca-Gro	McCaughey et al. (2006)	Ontario, Canada	48.22, −82.16	MF (DBF)
US-Ton	Ma et al. (2007)	California, USA	38.43, −120.97	WSA (MISC)
US-So2	Luo et al. (2007)	California, USA	33.37, −116.62	SHR (MISC)
US-Brw	Harazono et al. (2003)	Alaska, USA	71.32, −156.63	TUN (MISC)
US-Atq	Oberbauer et al. (2007)	Alaska, USA	70.47, −157.41	TUN (MISC)
US-Los	Sulman et al. (2009)	Wisconsin, USA	46.08, −89.98	WET (MISC)
Ca-WP1	Flanagan and Syed (2011)	Alberta, Canada	54.95, −112.47	WET (MISC)

Plant functional types (PFT) are: ENF(T/B), evergreen needleleaf forest (temperate/boreal); GRASS, grassland; WET, wetland; DBF, deciduous broadleaf forest; CROP, cropland; SHR, shrubland; WSA, woody savannah; TUN, tundra; MF, mixed forest; and MISC, miscellaneous, a combination of WET, WSA, SHR, and TUN.

Grouping approach for model–data comparison

We conducted model–data comparisons by flux type (NEE, GPP, and RE), time increment (monthly, annual), plant functional type (PFT), and model formulation. All 36 sites were categorized into the following PFTs (Table 2): deciduous broadleaf forest (DBF), temperate evergreen forest (ENFT), boreal evergreen forest (ENFB), grassland (GRASS), crops (CROP), and miscellaneous (MISC). For our analysis, mixed-forest sites, a combination of deciduous and evergreen forest, were included under the DBF designation. Shrubland, tundra, woody savannah, and wetland sites, none of which were represented by more than two sites, were included in the MISC designation. In this way, we reduced the original 10 PFT groups based upon the IGBP classification into 6 PFTs to increase the sample size within the groupings. It is important to note that these classifications do not necessarily represent the PFTs used in the regional models. They represent a grouping according to the land cover representative of the flux tower footprints. The magnitude and variation of the observed fluxes grouped by the PFT of the flux tower site is provided in Appendix B: Tables B1–B3. The range of flux magnitude and variation across sites within a PFT grouping is provided in Appendix B: Figs. B1–B3. During a preliminary evaluation of results, the Can-IBIS model demonstrated outlier behavior that significantly changed the findings both when considering all regional models together or when grouped by model formulations. For example, when considering all sites, Can-IBIS averaged over twice as much magnitude in annual gross flux as compared to the observations. The vast majority of the other models were within $\pm 30\%$ of the observed gross carbon fluxes. Therefore, with the exception of the crossover model comparison, Can-IBIS was not included within any findings that required the grouping of models. Can-IBIS was retained for the crossover model comparison, however, because its model runs were part of both model groupings (site and regional models).

There are a variety of ways that biogeochemical models represent ecosystem function and processes as well as responses to environmental constraints. Here, we analyzed model performance based upon photosynthetic, phenological, and soil carbon decomposition formulations. The major photosynthetic groupings were

enzyme-kinetic (EK) and light-use-efficiency (LUE) models. EK models emphasize the light- and enzyme-limited constraints on photosynthesis and are generally considered more physiologically based than LUE models. In contrast, LUE models take an empirical approach to estimating photosynthesis, by combining the fraction of photosynthetically active radiation (fPAR), a measure or proxy of leaf area index (LAI), and a light-use-efficiency or conversion factor. Phenological groupings were divided between models that use prescribed vs. internally predicted LAI. LAI can be estimated over large regions using remote measurements (e.g., Cook et al. 2008), but there can be considerable variability in performance when using different LAI products (Garrigues et al. 2008). Nevertheless, prescribed LAI should be more accurate and reduce computational costs, but limits the model's prognostic capability. The phenological and photosynthetic groupings in the models represented here are very similar in that the EK models mostly use prognostic LAI, whereas the LUE models primarily prescribe LAI. For this reason, the photosynthetic grouping was used in place of the phenological grouping because we could not separate these factors with the suite of models available. Finally, the soil carbon decomposition grouping was divided between first-order soil carbon decomposition rate models that include the influence of nitrogen dynamics upon respiration processes and those that did not (Huntzinger et al. 2012). First-order decomposition models include the size of the soil carbon pool when determining the rate of decomposition. The inclusion of nitrogen dynamics should have an impact on overall ecosystem respiration (Waring and Running 2007). Although not a formal model grouping, the models EC-MOD and MOD17+ are unique in that they are based on data-driven or data assimilation (data fusion) methods. Data-driven methods make use of flux observations and statistical approaches (e.g., ensemble of regression models) to develop flux models (Xiao et al. 2008). Data assimilation indicates that model parameters are estimated or optimized with the observed flux tower data (e.g., Braswell et al. 2005). Clearly, incorporating flux observations provides EC-MOD and MOD17+ with an advantage over the remaining models.

Model ranking

The model ranking developed for this study took into account five metrics at once in order to quantify a model's overall performance. These five metrics measured (1) the bias in magnitude between the average modeled and observed fluxes, (2) the average difference in flux magnitude between the modeled and observed fluxes (RMSD), (3) the temporal correlation between modeled and observed fluxes (R), (4) the similarity in temporal magnitude of variation between modeled and observed fluxes (sigma ratio), and (5) the agreement of modeled and observed fluxes considering the observed uncertainty (chi-square). The final model rankings (Tables 6–10) were created by first calculating a statistical value for every combination of metric ($n = 5$), model grouping (PFTs and all sites) ($n = 7$), flux ($n = 3$), and time resolution ($n = 2$), where the values in parentheses are the number of groups in each category. For example, a correlation value was calculated for every model (17 total), for annual NEE, for DBF sites only. The correlation values are sorted from best (smallest value) to worst (largest value). Next, the correlation value for each model is replaced by a ranking value of 1 through 17 (1 = best, 17 = poor). This process was completed a total of 210 times to cover every combination within each category. The values in Tables 6–10 are the average model ranking within the respective grouping.

Diagnostic roadmap

Our approach was to present the performance of all regional models based upon one metric at a time, followed by a diagnosis of the main result (e.g., NEE bias). The diagnosis was accomplished by comparison of performance based upon grouping of the models by individual PFTs, site vs. regional model runs, model formulations, and meteorology data as needed. The ability to compare site and regional model performance allows us to identify the significance of spatial mismatch and model setup upon the results. Furthermore, the crossover models can help identify if shortcomings in regional model performance are inherent to the model structure or a result of spatial mismatch and setup differences.

Results and Discussion

Diagnosis of bias

Annual bias.—Overall, the annual NEE of the 16 regional models had a positive bias ranging from 52% to 71% across all PFTs (Table 3). A positive bias in annual NEE indicates that overall, the models are systematically underestimating the net carbon uptake. The models also overestimate the annual gross fluxes (GPP and RE) by 5% and 15%, respectively, when considering all sites together. These biases are outside of the 1-sigma observed uncertainty range (worst case) for NEE (±14%), GPP (±4 %), and RE (±6%). The models tend to overestimate RE more than GPP, which contributes to the overall underestimation of the carbon sink. On a seasonal basis, the regional models underestimated GPP during the growing season, but overestimated the GPP during the transition seasons (see Appendix A: Fig. A1). This is consistent with the findings from the site-level analyses where models tended to over-predict GPP during cold conditions (Schaefer et al. 2012) and overestimate leaf area during transition seasons (Richardson et al. 2012). Nevertheless, over the entire year, the regional models underestimated the net carbon uptake (Keenan et al. 2012), whereas they overestimated gross fluxes. The range of

TABLE 3. Regional modeled fluxes with bias (model output − observations), excluding Can-IBIS.

PFT	NEE			GPP			RE		
	Annual modeled flux ($g\ C \cdot m^{-2} \cdot yr^{-1}$)	Bias ($g\ C \cdot m^{-2} \cdot yr^{-1}$)	Bias (%)	Annual modeled flux ($g\ C \cdot m^{-2} \cdot yr^{-1}$)	Bias ($g\ C \cdot m^{-2} \cdot yr^{-1}$)	Bias (%)	Annual modeled flux ($g\ C \cdot m^{-2} \cdot yr^{-1}$)	Bias ($g\ C \cdot m^{-2} \cdot yr^{-1}$)	Bias (%)
ENFT	−55	135	71	1187	−472	−28	1133	−337	−23
DBF	−73	110	60	1262	66	5	1179	164	16
CROP	−86	184	68	1394	169	14	1288	332	35
ENFB	−17	18	52	914	225	33	898	244	37
GRASS	−38	56	60	828	242	41	786	296	61
MISC	−39	45	54	1089	351	48	1048	396	61
All sites	−54	96	64	1142	51	5	1082	140	15

Notes: The percentage bias is calculated in relation to the observed flux values. Abbreviations are: NEE, net ecosystem exchange; GPP, gross primary production; and RE, total ecosystem respiration. See Table 2 for PFT abbreviations.

the overestimation of the gross fluxes varied greatly across the PFTs (Table 3). The models significantly overestimated the gross fluxes of the grass sites during the growing season. The site-level analysis (Schaefer et al. 2012) attributed this to the inability of the models to properly simulate soil moisture, drought, and humidity stress. The ENFT sites are unique in that the models systematically underestimate the gross fluxes for the entire year (see Appendix A: Fig. A1). Individual model performance is listed in Appendix A: Table A2.

Crossover models.—We diagnosed the flux bias of the regional models by comparing the site- and regional-level simulations of the crossover models. The gross fluxes for the regional level runs are at least 30% higher in magnitude than the site level runs (Table 4). From a seasonal vantage point (Fig. 1; Appendix A: Fig. A2), the regional runs approximate the magnitude of NEE and GPP better than the site runs during the growing season months. Outside of the growing season, the regional runs overestimated the GPP by roughly 20%. This does not appear to be due to an error in the seasonal timing of GPP, but to a persistent overestimate of the GPP. Although the site-level model mean matches the observed annual sum of GPP well, this is the result

TABLE 4. Comparison of annual flux bias (model output − observations) between the site-level and regional-level (crossover) model runs.

PFT and level	NEE			GPP			RE		
	Annual bias ($g\ C \cdot m^{-2} \cdot yr^{-1}$)	Bias (%)	Δ†	Annual bias ($g\ C \cdot m^{-2} \cdot yr^{-1}$)	Bias (%)	Δ†	Annual bias ($g\ C \cdot m^{-2} \cdot yr^{-1}$)	Bias (%)	Δ†
ENFT									
Region	132	68	8	−146	−9	142	−23	−2	192
Site	124	62		−288	−17		−215	−15	
DBF									
Region	101	55	−47	425	35	314	521	51	307
Site	147	76		111	9		214	21	
CROP									
Region	182	68	12	545	45	881	720	75	886
Site	170	55		−336	−26		−166	−16	
ENFB									
Region	−4	−11	−13	486	70	252	481	74	254
Site	9	24		233	33		227	34	
GRASS									
Region	54	59	−20	314	54	371	366	75	353
Site	74	67		−57	−10		13	3	
MISC									
Region	43	51	−16	657	90	351	701	108	375
Site	59	68		306	43		326	52	
All sites									
Region	88	59	−14	357	33	342	441	47	356
Site	102	66		15	1		86	9	

Notes: "Region" denotes regional-level model runs, and "site" denotes site-level model runs. See Table 2 for PFT abbreviations.
† Delta (Δ) is the difference in bias between the regional simulations (first row) and the site simulations (second row), therefore there is only one Δ value for each PFT.

of compensating biases between the growing and transition seasons. RE was systematically overestimated during the entire year for the regional runs, whereas the site-level runs approximated the RE magnitude and seasonal pattern very well. The crossover model performance thus shows that: (1) the difference in model setup (site vs. regional) did not change the overall annual biases in modeled carbon uptake (NEE); however, (2) the gross fluxes of the regional model runs were significantly higher than the site-level runs. Thus, even though many of the regional models include spin-up and a representation of disturbance history, they, like the site model runs, underestimated the observed carbon sink. On the other hand, it appears that the model setup does play a role in the gross flux biases. The cause of the degraded performance of the regional crossover runs has important implications for interpretation of the validity of the regional model results as a whole. If the differences in the crossover runs were caused simply by vegetation map mismatch, for example, then we might expect the regional flux biases to be random by site, and the regional models to perform well for a continental average. The persistent overestimation of GPP and RE by the regional crossover models suggests a systematic cause other than vegetation cover resolution.

Meteorology data.—We examined the influence of meteorology upon the simulated fluxes by parsing the model output into groupings of shortwave radiation driver data. This grouping was chosen because two of the shortwave radiation products used for the regional models were found to be 39% (NCEP) and 28% (NARR) positively biased with respect to site observations (Oak Ridge National Laboratory Distributed Active Archive Center [ORNL DAAC]; data *available online*).[17] Another radiation product, CRU-NCEP, was found to be slightly negatively biased (−4%; see footnote 17). These radiation products were used in eight out of the 17 regional models (Table 1), and only these eight models were considered within the radiation groupings. The radiation bias is most pronounced during cloudy conditions, in which leaves were not saturated with light, unlike in full-sun conditions. Consequently, shortwave radiation positive bias during cloudy conditions should promote vegetation growth. On the contrary, the models using positively biased radiation gave lower GPP and net carbon uptake than the models run with negatively biased radiation (Appendix A: Table A3). To test if other confounding model factors masked the expected signal upon GPP, we examined the three crossover models that used known positively biased regional radiation products and compared the fluxes between the site and regional-level runs. BEPS and CLM-CN demonstrate flux biases consistent with positively biased radiation, i.e., high net carbon uptake and high gross fluxes; however, DLEM shows a negative bias (Appendix A: Table A4).

[17] http://daac.ornl.gov

The variability of meteorology data may also contribute to the positive bias in modeled GPP. Medvigy et al. (2010) demonstrated that lower variation (lower standard deviation) in radiation and precipitation driver data boosts modeled GPP, RE, and NEP fluxes as compared to modeled output derived from higher variation in driver data at Harvard Forest. In particular, they found that two of the regional meteorology products that are used in the NACP Interim Synthesis, NCEP and ISCCP, exhibited lower overall variability than the site meteorology observations. We could not identify a strong link between the regional models that used these regional meteorology products and provided a positive bias in GPP.

In summary, only when we considered a subset of crossover models that use positively biased shortwave radiation data, do the majority of these models produce a positively biased GPP as compared to the site-level runs. Otherwise, the positive bias GPP signal is lost, indicating that the biases in the shortwave radiation products were not the dominant influence on the overall regional model positive flux biases. The regional meteorology products are also subject to temperature biases; however, a separate analysis (not shown here) did not reveal any obvious linkages between biases in temperature and biases in GPP.

Model structure formulations.—We now examine the influence of model structure upon the flux biases. EK-based models simulated higher magnitude fluxes compared to LUE-based models for both GPP and RE (Fig. 2, Table 5; Appendix A: Fig. A3). The EK models also simulated more net carbon uptake (more negative NEE) relative to LUE models. These results are consistent with those of Huntzinger et al. (2012), who performed a similar model formulation comparison of continental fluxes for North America. The flux tower data suggest that the EK models are simulating high GPP (15% bias) that is offset by even higher RE (26% bias). This results in a substantial underestimation of carbon uptake (61% bias). The LUE models, on the other hand, underestimate GPP (−15%), leading to larger underestimations of net carbon uptake (81% bias in NEE).

It is not clear what drives the difference in estimation of GPP (and RE) between EK and LUE models. The EK models overestimate the length of the growing season (Fig. 2) consistent with Richardson et al. (2012), who found the modeled growing-season length was overestimated for deciduous sites within the NACP site synthesis. The overestimated growing season (positive bias in GPP), however, is at least partially compensated for by the underestimation of GPP during the peak of the growing season. Therefore, phenology is unlikely the main driver for the gross flux bias in EK models. Another explanation for the differences in GPP relates to the extent of parameter estimation used within the model groupings. VEGAS2, an LUE model, optimizes a key photosynthetic parameter that forces the continental GPP to fall within an accepted range (N. Zeng, *personal communication*).

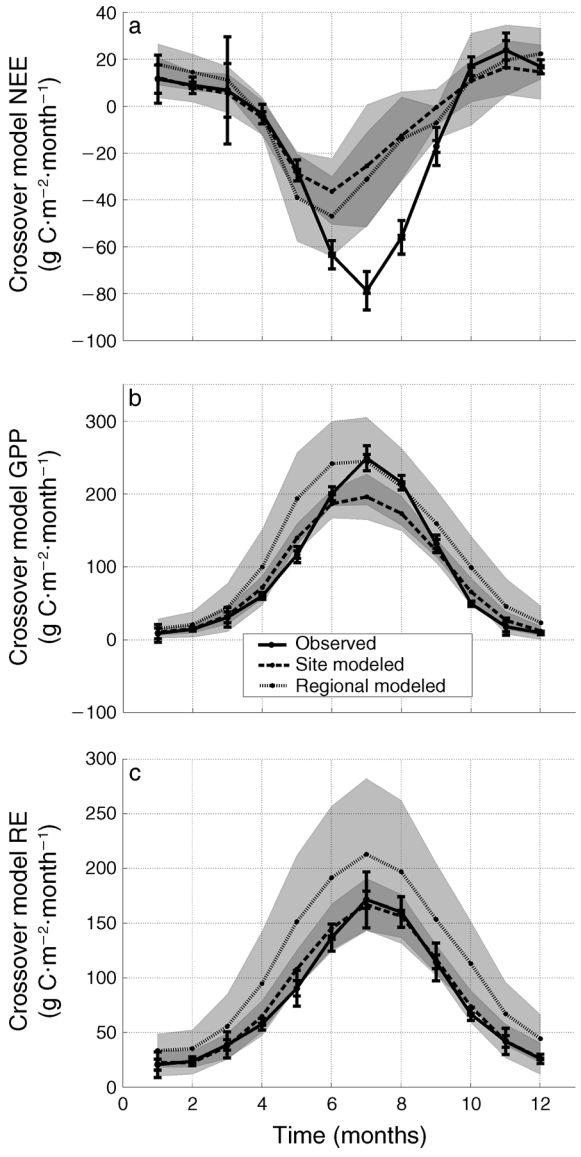

FIG. 1. Mean monthly fluxes for all sites for crossover models only (including Can-IBIS) for (a) net ecosystem exchange (NEE), (b) gross primary productivity (GPP), and (c) total ecosystem respiration (RE). The error bars on the observations are ±1 sigma values (best- and worst-case scenarios) calculated from monthly modeled uncertainty. The shaded regions represent the ±1 sigma values of the across-model spread for each model grouping (darker gray indicates overlap of shaded areas).

VEGAS2 is not unique in its calibration methodology as other types of models perform similar calibrations based upon accepted regional flux values (Cramer et al. 1999, Ruimy et al. 1999). In this way parameter optimization might help explain the minimal positive GPP biases for LUE and EK models, even when confronted with positively biased radiation data.

When grouped by soil carbon decomposition type, only the no-nitrogen model mean overestimates the magnitude of respiration (25% bias), resulting in an underestimation of the overall carbon uptake (84%; Fig. 3; Appendix A: Table A5). The increased amount of RE for the no-nitrogen grouping is consistent with the findings of Huntzinger et al. (2012), who found a similar trend between the two formulations when comparing heterotrophic respiration results for all of North America. In this study, nitrogen-inclusive models were, on average, more consistent with the observations (lower RE and NEE bias).

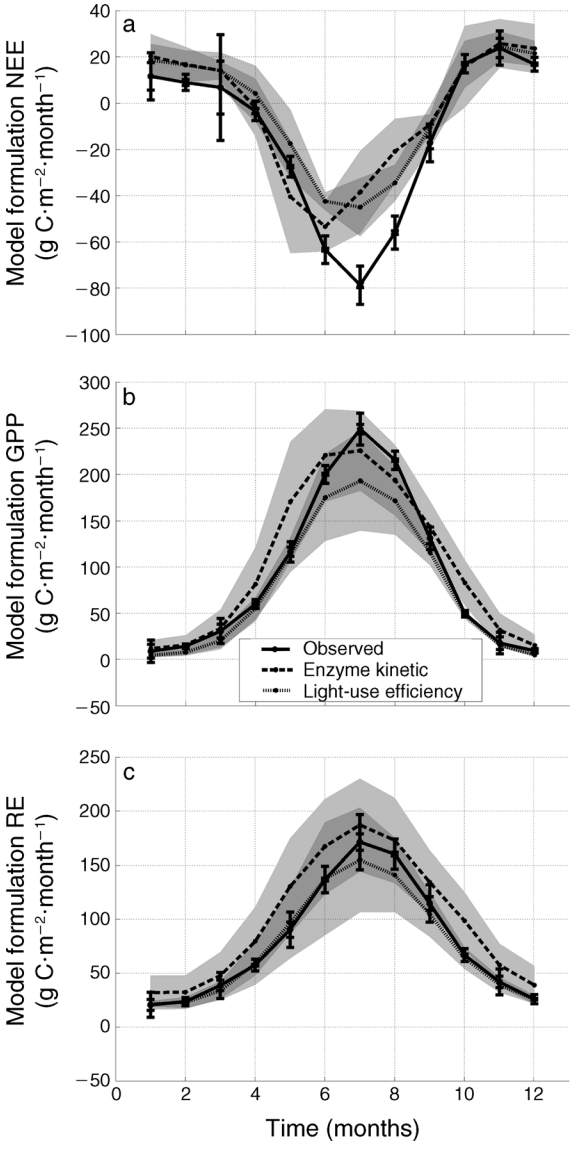

FIG. 2. Mean monthly fluxes for regional models at all sites categorized by photosynthetic model formulation for (a) NEE, (b) GPP, and (c) RE. The error bars on the observations are ±1 sigma values (best- and worst-case scenarios) calculated from monthly modeled uncertainty. The shaded regions represent the ±1 sigma values of the across-model spread for each model grouping (darker gray indicates overlap of shaded areas).

TABLE 5. Comparisons of annual flux bias between enzyme-kinetic (EK) and light-use-efficiency (LUE) models.

PFT and photosynthetic formulation	NEE			GPP			RE		
	Annual bias ($g\ C{\cdot}m^{-2}{\cdot}yr^{-1}$)	Bias (%)	Δ†	Annual bias ($g\ C{\cdot}m^{-2}{\cdot}yr^{-1}$)	Bias (%)	Δ†	Annual bias ($g\ C{\cdot}m^{-2}{\cdot}yr^{-1}$)	Bias (%)	Δ†
ENFT									
EK	133	70	−14	−355	−21	334	−222	−15	292
LUE	147	80		−690	−42		−515	−36	
DBF									
EK	105	59	−19	166	14	293	271	27	285
LUE	124	67		−126	−10		−13	−1	
CROP									
EK	184	71	−35	295	24	412	479	50	389
LUE	219	80		−118	−9		90	9	
ENFB									
EK	−3	−9	−72	336	49	320	333	51	246
LUE	69	198		16	2		86	13	
GRASS									
EK	58	60	−26	276	48	161	334	70	137
LUE	84	90		115	19		197	39	
MISC									
EK	47	55	−34	465	63	358	512	79	320
LUE	81	97		107	15		192	29	
All sites									
EK	89	61	−34	157	15	325	246	26	286
LUE	123	81		−168	−15		−40	−4	

† Delta (Δ) is the difference in bias between the regional simulations (EK; first row) and the site simulations (LUE; second row); therefore there is only one Δ value for each PFT.

In summary, both the regional- and site-level model runs substantially underestimated the net carbon sink, despite that the majority of the regional model runs include a disturbance history and are not run from a state of equilibrium. The underestimate of the carbon sink is persistent regardless of the implementation of disturbance history (e.g., prescribed land use, harvest, fire) for the regional models. For the regional models that include fire, the average carbon emissions from fire only accounted for 6% of the NEE and are too small to explain the NEE bias. This indicates either the initial conditions do not represent the state of the ecosystem or that the models underestimate ecosystem productivity as a result of inaccurate parameterization/structure. Although the flux tower observations indicated a larger carbon sink than the regional models (64% lower NEE), they are consistent with atmospheric inversion estimates, which predict an 80% larger carbon sink than the regional models (Hayes et al. 2012). On the other hand, the inventory approach by Hayes et al. (2012) provides a net carbon sink that is more consistent with the regional models.

Annual variability

Temporal (within-site) and spatial (across-site) correlation.—Overall, the regional model runs demonstrated negligible temporal correlation (range in R was −0.2 to +0.2 across all sites) with observed interannual variation in NEE, GPP, and RE, with the exception of the correlation of GPP and RE at ENFB sites ($R = 0.4$) and grass sites ($R = 0.3$; Fig. 4; Appendix A: Table A6). The correlation between modeled and observed gross fluxes is consistently higher than that of NEE (Appendix A: Table A6), most likely a reflection of the high sensitivity of NEE to small relative errors in the large gross fluxes. No single model has a higher R value than 0.4 for any annual flux when all sites are considered (Fig. 4). The interannual correlation improved from regional- to site-level runs across all sites and models (Fig. 5). This improvement is most pronounced for GPP ($R = 0.09$–0.46; Appendix A: Table A7). This suggests that the models possess some skill at predicting interannual variations for all fluxes, but this skill is diluted when regional driver data (meteorology and vegetation cover) were used. For the site level models, the correlation was lowest for RE, suggesting this is the main contributor to poor interannual variability in NEE. The crop sites demonstrated the most improvement between region and site runs (Appendix A: Table A7). Such a large improvement in site-level performance is likely due to the specification of changes in vegetation cover (i.e., corn/soybean crop rotation), whereas the regional runs were likely provided an unchanging, generic land cover type.

In terms of photosynthetic formulation, LUE models tended to outperform EK models for temporal flux correlation (Fig. 4; Appendix A: Table A8). Almost all of the LUE models used prescribed phenology, whereas EK models used mostly prognostic phenology. The prescribed phenology products used within the RCIS

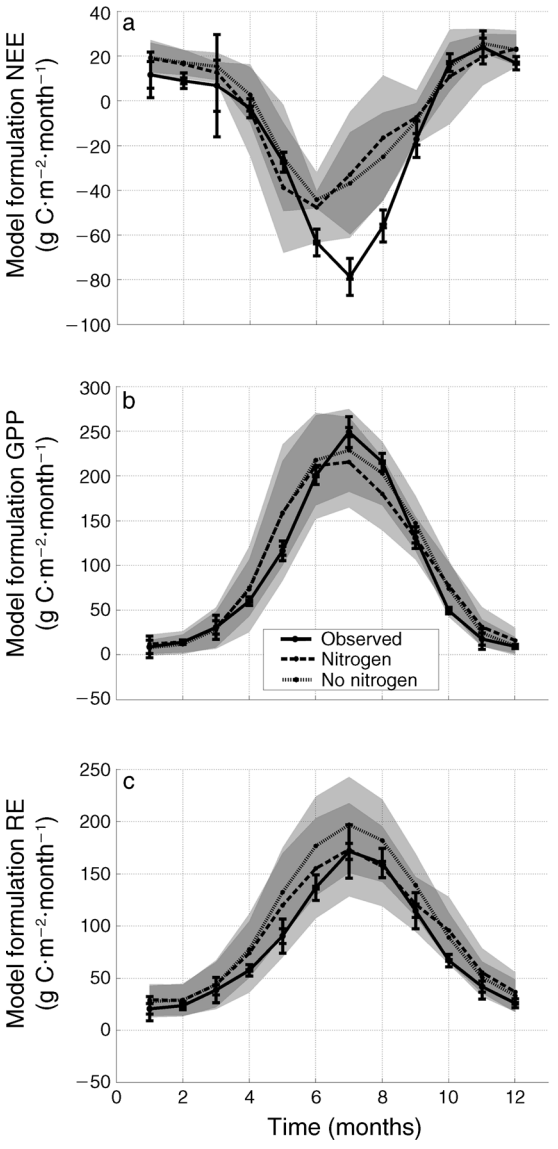

FIG. 3. Mean monthly fluxes for regional models at all sites categorized by soil carbon decomposition formulation for (a) NEE, (b) GPP, and (c) RE. The error bars on the observations are ±1 sigma values (best- and worst-case scenarios) calculated from monthly modeled uncertainty. The shaded regions represent the ±1 sigma values of the across model spread for each model grouping (darker gray indicates overlap of shaded areas).

include interannual variation and are capable of improving the modeled leaf onset and senescence processes that control the length of the growing season, which in turn influences the year-to-year flux variation (correlation). The seasonal fluxes in Fig. 2 appear to support this idea, as the gross fluxes are well captured by the LUE models during the transition seasons where leaf onset and senescence timing are important. The improvement in correlation for the LUE models is most significant for the crop, grass, and ENFB sites. For the soil carbon decomposition formulation, there were no significant correlation differences between the model groupings for all fluxes.

Overall, the regional models possessed some skill at capturing spatial (across-site) correlation for the annual gross fluxes (for GPP, $R = 0.36$; for RE, $R = 0.29$), but less skill for NEE ($R = 0.18$; Appendix A: Table A9). The spatial correlation was generally stronger than the temporal correlation. We attribute this to the fact that annual flux has more variation across space than time. Therefore, the models are more likely to identify flux patterns where there exist larger flux gradients across sites (i.e., larger signal), hence higher correlation values. The overall performance for individual models is provided in the appendices (Appendix A: Fig. A5).

Temporal (within-site) and spatial (across-site) magnitude of variability (sigma).—Overall, the regional models captured approximately half of the magnitude of year-to-year temporal variability for all fluxes ([normalized sigma] σ_{NEE}, 0.51; σ_{GPP}, 0.63; σ_{Re}, 0.56; Fig. 4; Appendix A: Table A10). The models tended to underestimate the variability for ENFT and crop sites the most. This suggests that the models do not capture the influence of crop rotation (soybean, corn), which could limit the magnitude of year-to-year flux variability (Lokupitiya et al. 2009). For crop sites, this implies that specific planting type/schedules, not climate, is the main driver of interannual variability. Conversely, the models predict twice as much variability than is observed for the ENFB sites for all fluxes. This could be a symptom of the models' tendency to overestimate fluxes during cold conditions (Schaefer et al. 2012), due to inaccurate temperature inhibition functions, or as a result of presumed instantaneous recovery from frost days during the growing season. The true recovery timing is dependent upon the number of frost days and may take weeks to months (Strand and Oquist 1985). On a per model basis, all models underestimated the temporal magnitude of variability in NEE and most models underestimated the variability in GPP and RE (Fig. 4). Next, we used the model groupings to help identify the source of the underestimation of variability. For the crossover models, overall, the regional runs showed slightly less temporal variability than the site-level runs for all fluxes (Appendix A: Table A11). Differences in temporal variability were dependent upon the PFT type, however, as the regional runs were much more variable for the ENFB sites, but less variable for the DBF sites for all fluxes. Unlike the case for interannual correlation, however, the site level runs did not substantially increase or improve the interannual variability as compared to the regional level runs.

In terms of photosynthetic formulation, the EK models were more variable (higher temporal annual variability) for all flux types compared to the LUE models (Fig. 4; Appendix A: Table A12). Whereas the EK models captured 62%, 61%, and 57% of the variability for NEE, GPP, and RE, respectively, the LUE models only captured 23%, 45%, and 36% of the

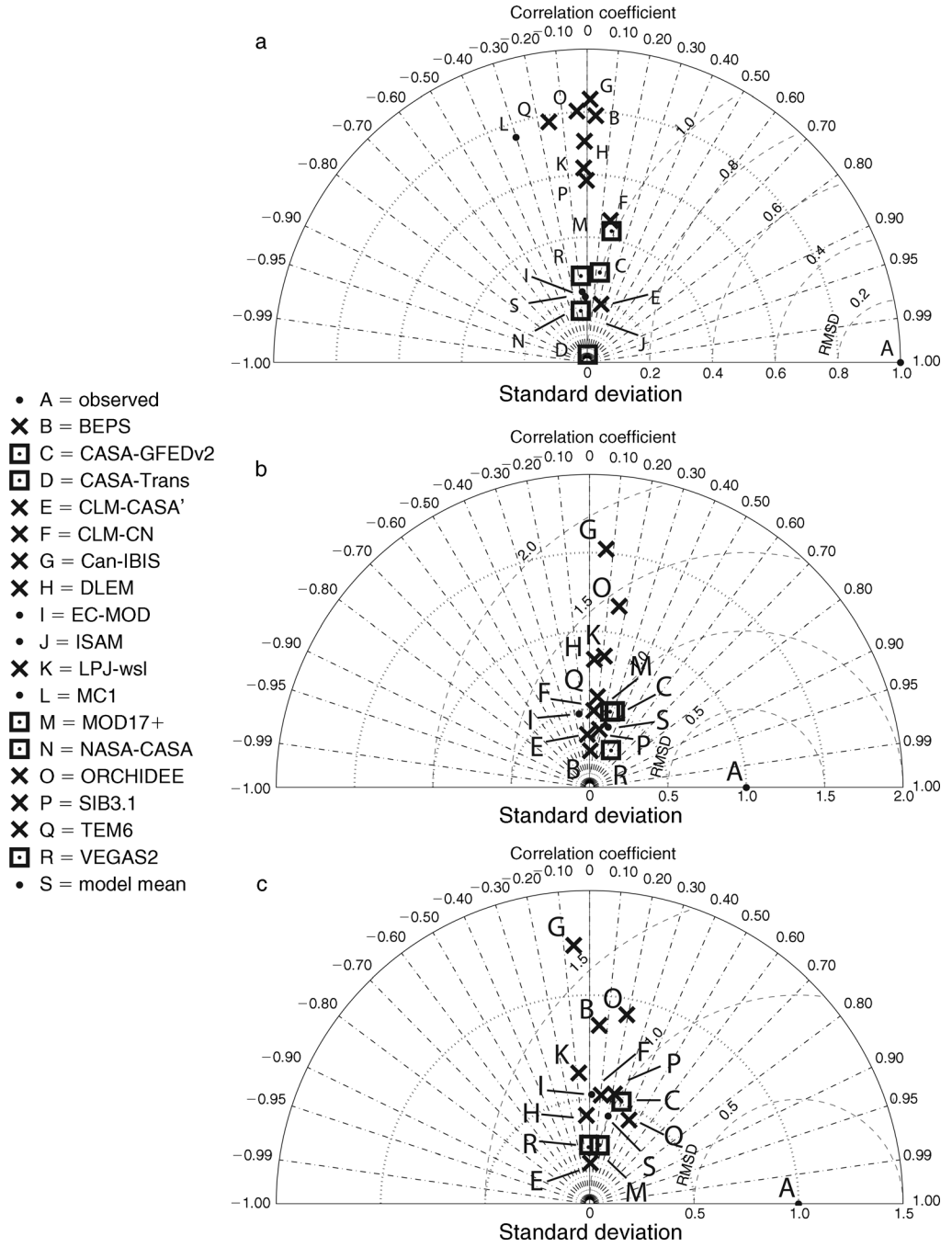

FIG. 4. Annual fluxes for all sites for (a) NEE, (b) GPP, and (c) RE. The statistics of correlation coefficient (black dotted-dashed axis lines), average difference in flux magnitude between the modeled and observed fluxes (RMSD; gray dashed axis lines), and standard deviation (gray dotted axis lines) are calculated from temporal (within-site) modeled variability. Squares represent light-use-efficiency models, X's represent enzyme-kinetic models, and dots represent statistical models (observed and model mean).

variability. The significant underestimation by the LUE models likely reflects the highly empirical nature of these models that are driven predominantly by radiation and LAI, and are less capable of capturing the temperature and soil moisture stresses that influence year-to-year changes in flux magnitude. For soil carbon decomposition formulation, the nitrogen inclusive models showed consistently higher annual variability as compared to the no-nitrogen models (Appendix A: Table A13).

In summary, all models showed a tendency to underestimate the magnitude of interannual variability for NEE, GPP, and RE. This tendency was reinforced in

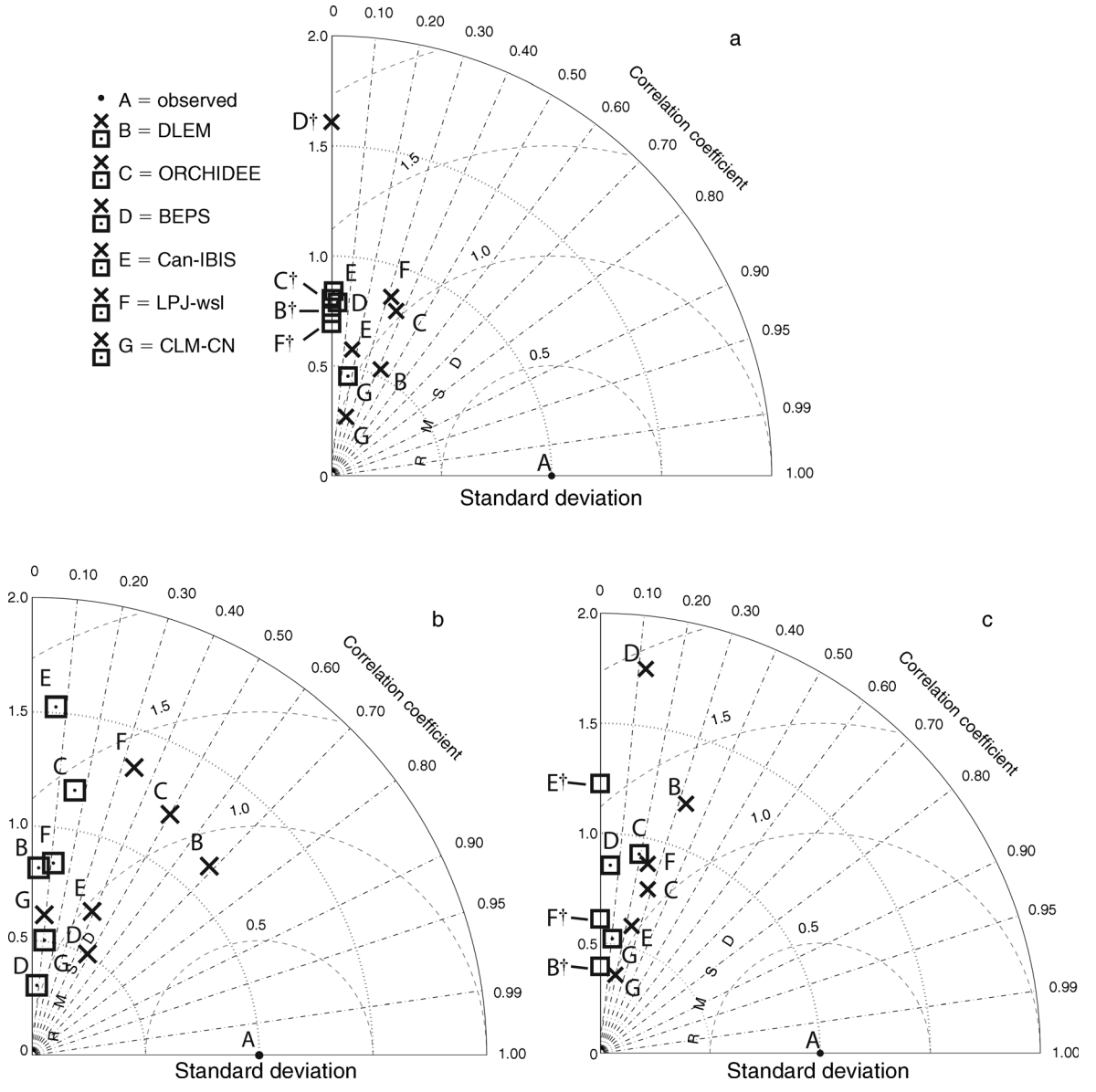

FIG. 5. Crossover models for all sites for annual (a) NEE, (b) GPP, and (c) RE. The statistics of correlation coefficient, RMSD, and standard deviation are calculated from within-site modeled variability. The site-level run for each model type is represented by X's, and the regional-level run is represented by squares. The daggers (†) indicate the model runs that were slightly negatively correlated with the observations but were moved to a correlation of zero in order to fit all the data on a single quadrant display for better viewing. Lines are defined as in Fig. 4.

the cases of LUE and no-nitrogen soil decomposition models. The crossover model analysis suggests that the magnitude of interannual variability is a function of model structure and parameters, yet mostly independent of model setup. In contrast to the temporal variability, the models captured the spatial (across-site) magnitude of variability for annual gross fluxes well, suggesting that input data such as vegetation cover, soil type, and climate forcing are sufficient for simulating spatial variability in these fluxes. The spatial variability is listed in Appendix A: Fig. A5 and Table A14.

Seasonal variability

Monthly correlation coefficient (R).—When considering all models and sites, the monthly fluxes correlated best with the observations of GPP ($R = 0.70$), followed by RE ($R = 0.63$) and NEE ($R = 0.43$; Appendix A: Table A15). More specifically, all modeled fluxes correlated best with the observations for the forested sites, including both deciduous and evergreen vegetation types. On the other hand, the modeled fluxes for the grass and crop sites tended to correlate most poorly with the observations. On a per model basis, the correlation

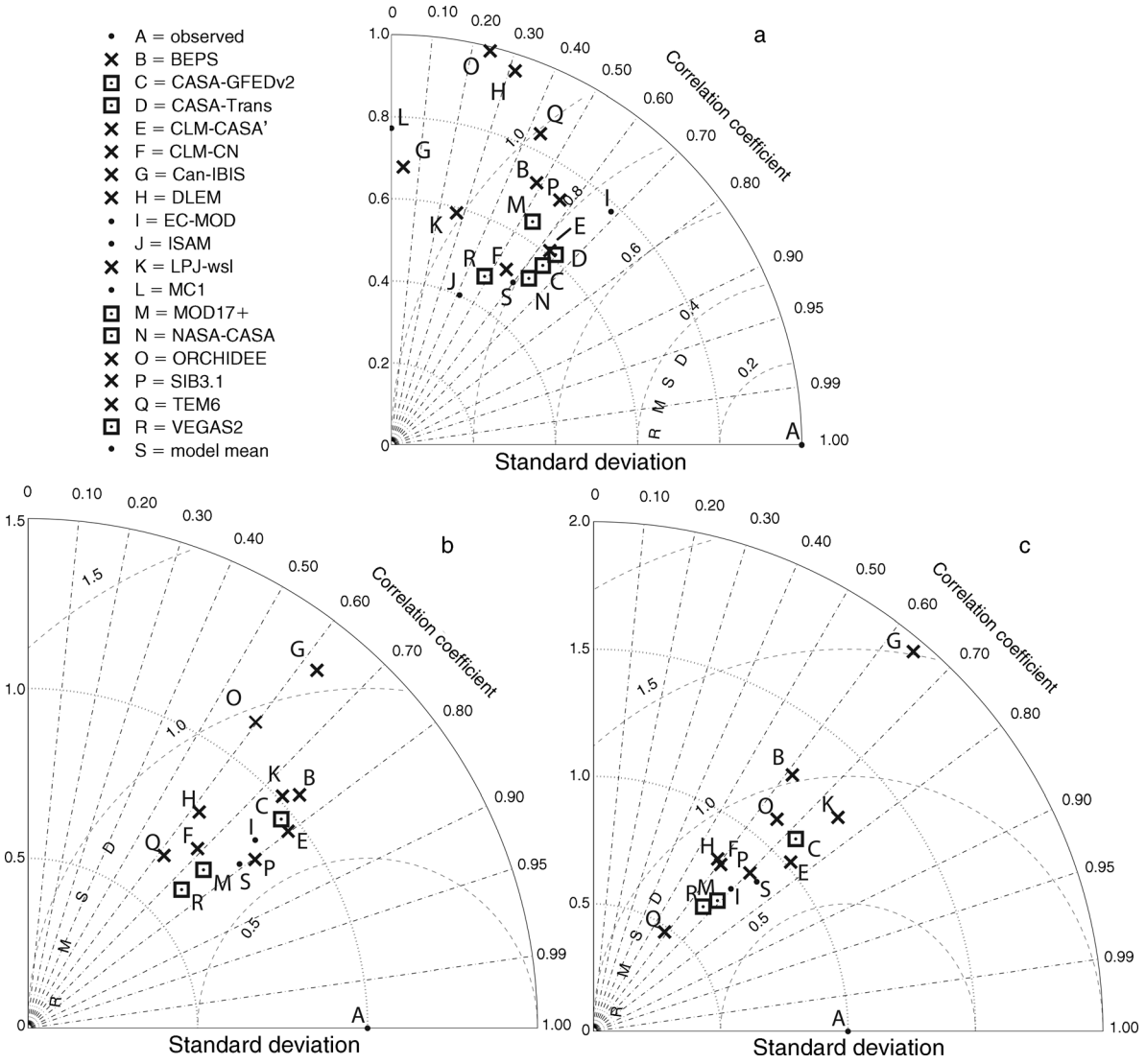

FIG. 6. Taylor diagrams grouped by photosynthetic formulation for monthly (a) NEE, (b) GPP, and (c) RE. All site–model pairs are grouped, and then one statistical value is calculated for all (includes both within- and across-site contributions). Squares represent light-use-efficiency models, X's represent enzyme-kinetic models, and dots represent statistical models (observed and model mean). Lines are defined as in Fig. 4.

performance was consistent for the gross fluxes, as the models ranged between an R value of 0.6 to 0.8 (Fig. 6). The modeled NEE, however, has a wider range of correlation ($R = 0.0$–0.7).

For the crossover models, the site-level runs consistently correlated better with the observations than the regional runs for all fluxes (Appendix A: Table A16). The most improvement between the crossover models was concentrated within the crop and grass sites. The improvement was just as persistent on a per model basis as nearly every site-level model run improved in comparison to its regional model counterpart in correlation for all fluxes (Fig. 7). The consistent improvement for site-level runs indicates that a significant amount of valuable site-specific driver data is lost in the regional model simulations. In particular, the improved performance of the grass and crop site level runs indicate that the correct vegetation characterization and initial condition are critical factors for capturing seasonal variation.

The LUE model groupings correlated slightly better than the EK grouping for gross fluxes and moderately better for NEE (Appendix A: Table A17). This result is most pronounced for the crop sites, which show the largest improvement in correlation for the LUE models. In general, LUE models have the advantage of using prescribed LAI likely improving the correlation. On a per model basis, the individual LUE models perform similarly in terms of correlation (and variability) for all fluxes, whereas the EK models have a much larger

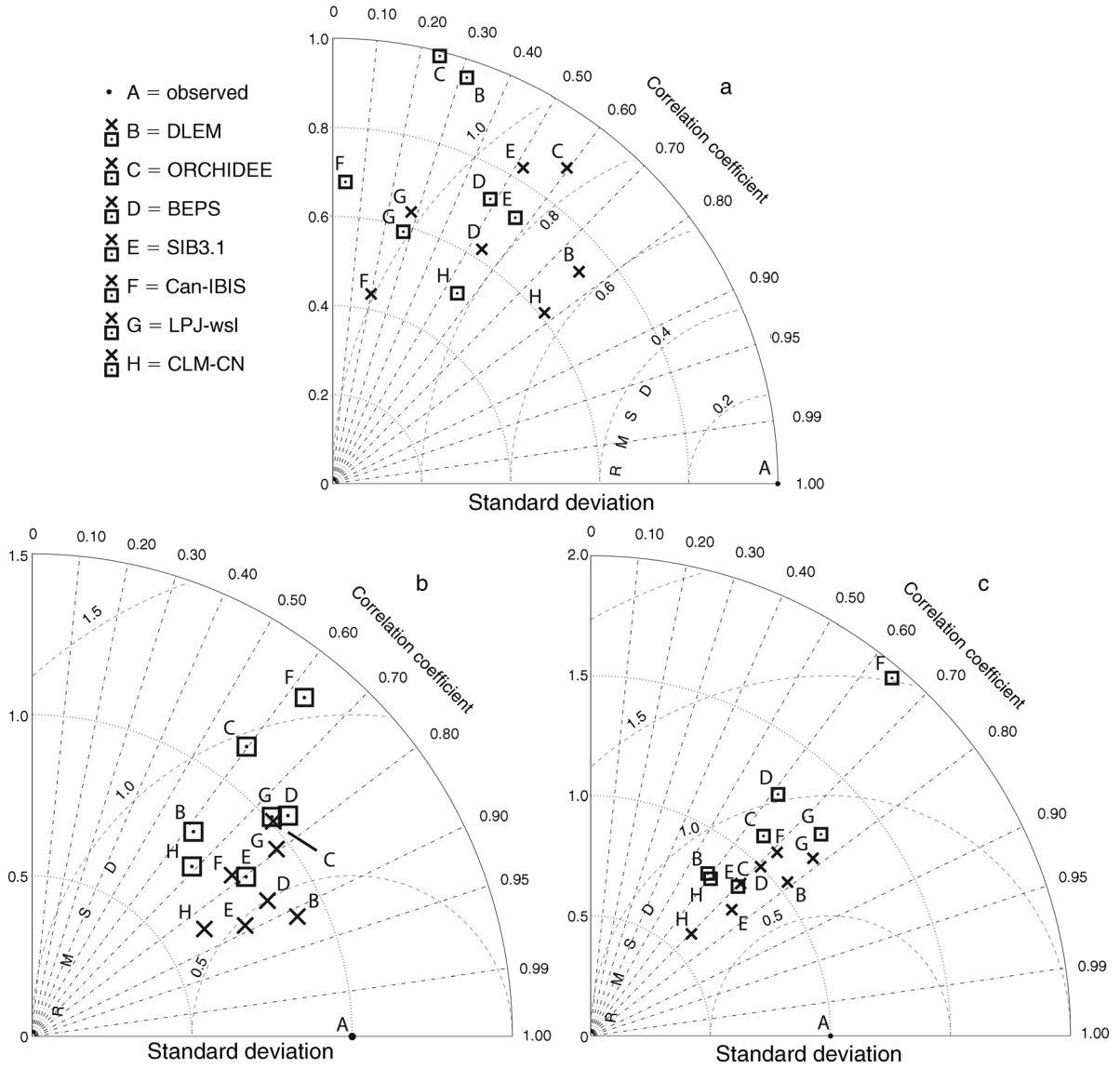

FIG. 7. Taylor diagrams for crossover models only for monthly (a) NEE, (b) GPP, and (c) RE. All site–model pairs are grouped and then one statistical value is calculated for all (includes both within- and across-site contributions). The site-level run for each model type is represented by X's; the regional-level run is represented by squares. Lines are defined as in Fig. 4.

spread in performance presumably influenced by the variety of phenological sub-models used to approximate bud-burst and senescence (Fig. 6). When grouped by soil carbon decomposition formulations, the no-nitrogen grouping performed consistently better across all sites in general, but in particular for crop sites (Appendix A: Table A18).

Monthly magnitude of variability (sigma).—On average, the modeled gross fluxes captured almost all of the observed magnitude of seasonal variation, whereas the modeled NEE captured only 70%, based upon the normalized standard deviation (Appendix A: Table A19). The models overestimated the variability for the ENFB and miscellaneous sites the most, whereas the crops sites were typically underestimated. Individual model performance is listed in Fig. 6.

For the crossover models, the regional runs were more highly variable across all gross fluxes for each PFT as compared to the site-level runs (Fig. 7; Appendix A: Table A20). Overall, both types of model runs were equally variable for NEE flux, but the regional runs demonstrated much higher variability for ENFB sites than the site-level models.

For the photosynthetic formulation, the EK models displayed higher variability and were more consistent with the observed seasonal variability for all flux types (Fig. 6; Appendix A: Table A21). Similar to the crossover model comparison, the EK models' increased

seasonal variability was a result of the consistent monthly positive bias for the gross fluxes as discussed in the subsection *Model structure formulations* (Fig. 2). The model groupings performed more similarly in terms of the NEE, but the ability of EK models to simulate a larger carbon sink during the growing season gave them an advantage in modeling seasonal variability.

The soil carbon decomposition model groupings did not express any significant differences in seasonal variability for GPP, but the no-nitrogen grouping had more variability for RE (Fig. 3; Appendix A: Table A22). The difference in the seasonal variability of RE likely stems from the high respiration for the no-nitrogen model grouping during the growing season (Fig. 3). The nitrogen model grouping captured more variability than the no-nitrogen grouping for NEE. The individual model behavior for the magnitude of seasonal variability is given in Appendix A: Fig. A6.

Seasonal timing maps.—The regional model mean of the gross fluxes predicted an elongated growing season (defined as the sharp ascent and descent of the monthly GPP) as compared to the observations (Appendix A: Fig. A1). This finding is consistent with that of the site-level model runs within the site synthesis analysis (Richardson et al. 2012). In addition, the models tended to predict a peak in uptake (NEE) that is approximately one month earlier than the observations. The deviations of growing-season length or peak uptake timing from the observations, however, were also PFT dependent. DBF and crop sites both displayed extended growing seasons with premature peak carbon storage, while ENFT and ENFB sites both showed a shortened growing season with a late peak carbon uptake. For the grass sites, the models depicted a late start to the growing season and also extended the growing season approximately two to three months too late into the fall. The modeled timing of the maximum carbon uptake for the grass sites matches the observations well, although the models were unable to capture the suppression of carbon uptake during the late spring and early summer, likely a result of soil moisture constraints (Schaefer et al. 2012).

For the crop sites, the models tended to predict an early onset of growth and a late senescence. The peak of the carbon uptake was modeled two months earlier than that of the observations. Most striking for the crop sites was the models' inability to capture the narrow and sharp peak of the growing-season fluxes. Whereas the observations indicated an intense growing season lasting from June to September, the models predicted a longer and gradual growing season extending roughly from April to October.

When examining seasonal flux characteristics (all sites) between crossover models, there was no discernible difference between the model groupings in terms of seasonal timing or maximum carbon uptake timing (Appendix A: Fig. A2). In fact, the seasonal flux patterns mimicked each other quite closely, and flux biases alone (most pronounced during the growing season) were the only differences between the model groupings (see *Results and discussion: Crossover models*). Similar differences from biases alone are observed for the ENFT, DBF, ENFB, and MISC sites. The crop and grass sites, however, displayed fundamental differences in seasonal flux shape and timing between the two model groupings. For the crop sites, the regional models predicted a muted and elongated growing season, whereas the site-level runs match the observations better. The site-level runs predicted a far more modest increase in photosynthesis during the spring that held steady into the late summer, similar to the observations. The differences are likely due to inaccuracy in the land cover maps used for the regional runs.

The photosynthetic formulation model groupings displayed differences in both seasonal timing and maximum carbon uptake when considering all sites (Appendix A: Fig. A3). The EK models predicted an early growing season and later senescence than the observations, whereas the LUE models were synchronous with the observations. The LUE models also matched the observations in terms of the maximum carbon uptake (July), whereas the EK model grouping maximum uptake was one month earlier. This is most likely because LUE models take advantage of remotely sensed LAI proxies, whereas EK models rely on internal mechanisms to predict leaf onset and senescence. Also of note, is that neither model grouping offers significant advantages over the other when considering grass and crop sites. This reinforces the assertion that vegetation cover issues are responsible for the inaccurate seasonal flux representation, and not issues of phenology or photosynthetic sub-modules in the case of these sites. The soil carbon decomposition model groupings offer similar seasonal fluxes (Appendix A: Fig. A4). Biased meteorology products (e.g., radiation, temperature) did not impose a discernible influence on the seasonal timing.

Individual model performance: individual and combined weighting of metrics

In this final section, individual model behavior is evaluated in terms of bias, RMSD, correlation, sigma ratio, and chi-square. MOD17+, a data assimilated model, consistently performed the best overall across all statistics, flux types, and time increments, performing especially well for the gross fluxes (Table 6). Other top performing models include CASA-GFEDv2, SIB3.1, and EC-MOD. With the exception of EC-MOD (data-driven), models that perform well for gross fluxes tend to perform more poorly for NEE. Conversely, models that best captured NEE, such as CLM-CASA' and CLM-CN, performed more poorly for the gross fluxes.

When evaluating model performance for all statistics and fluxes based upon annual time increments only, CLM-CN, a process-based model, performed the best,

TABLE 6. Average model ranking and model-ranking standard deviation (in parentheses), showing photosynthetic formulation.

NEE				GPP			
Annual		Monthly		Annual		Monthly	
Rank	Model	Rank	Model	Rank	Model	Rank	Model
5.8 (0.4)	CLM-CASA'	6.0 (4.2)	**EC-MOD†**	5.3 (3.5)	DLEM	4.4 (2.5)	SIB3.1
6.3 (3.8)	CLM-CN	6.6 (4.6)	*NASA-CASA*	5.6 (2.8)	*MOD17+†*	5.2 (3.4)	*MOD17+†*
6.9 (4.1)	**ISAM**	7.1 (4.4)	MOD17+†	5.8 (3.2)	*CASA-GFEDv2*	5.4 (3.2)	*CASA-GFEDv2*
7.6 (4.6)	ORCHIDEE	7.2 (4.3)	CLM-CASA'	5.9 (2.6)	SIB3.1	5.9 (4.1)	*VEGAS2*
8.0 (5.2)	**EC-MOD†**	7.9 (4.0)	*CASA-GFEDv2*	6.4 (3.1)	CLM-CN	6.2 (3.1)	BEPS
8.0 (3.5)	LPJ-wsl	8.0 (4.6)	BEPS	6.4 (3.4)	**EC-MOD†**	6.3 (3.2)	**EC-MOD†**
8.5 (5.2)	*NASA-CASA*	8.0 (4.8)	*CASA-Trans*	6.5 (3.9)	*VEGAS2*	6.6 (4.3)	TEM6
8.7 (5.4)	*MOD17+†*	8.3 (4.3)	CLM-CN	7.1 (3.6)	LPJ-wsl	7.3 (3.1)	CLM-CN
8.8 (3.3)	*CASA-GFEDv2*	8.3 (4.5)	SIB3.1	7.2 (4.5)	TEM6	7.4 (2.8)	LPJ-wsl
9.4 (5.6)	BEPS	8.9 (5.6)	TEM6	7.2 (3.3)	BEPS	7.4 (3.3)	DLEM
9.7 (4.4)	Can-IBIS	9.0 (4.5)	**ISAM**	8.4 (4.3)	ORCHIDEE	7.5 (3.8)	CLM-CASA'
10.0 (5.9)	*CASA-Trans*	9.4 (4.2)	*VEGAS2*	8.7 (3.4)	CLM-CASA'	9.4 (3.7)	ORCHIDEE
10.1 (5.6)	TEM6	9.6 (3.9)	LPJ-wsl	9.6 (4.1)	Can-IBIS	10.9 (2.6)	Can-IBIS
10.1 (4.3)	SIB3.1	9.6 (6.2)	ORCHIDEE				
10.2 (3.4)	DLEM	10.6 (4.2)	DLEM				
10.6 (5.8)	**MC1**	12.3 (4.6)	**MC1**				
10.8 (2.9)	*VEGAS2*	12.6 (4.0)	Can-IBIS				

Notes: The average model ranking is calculated from the individual model rankings (1–17) for every model-PFT-metric combination based upon bias, the average difference in flux magnitude between the modeled and observed fluxes (RMSD), correlation coefficient, sigma ratio, and chi-square. The photosynthetic formulation is displayed as enzyme kinetic in normal type, light-use efficiency in italic type, and statistical in boldface type. Statistical models did not use a mechanistic formulation to simulate photosynthesis (GPP), but rather a statistical fitting routine or statistical estimation for calculating photosynthesis (GPP). A dagger (†) denotes that the model underwent data assimilation. Empty cells indicate that data were not available for this model inter-comparison, meaning either that the models did not simulate GPP or RE, or that they did not simulate specific sites.

followed by EC-MOD and MOD17+ (Table 7). At least one of those models ranked in the top three overall for each PFT, except for ENFT and ENFB, where none of the overall top model performers ranked in the top three. The top performers were ISAM, DLEM, and SIB3.1 for ENFT sites, and ORCHIDEE, LPJ-wsl, and CLM-CASA' for ENFB sites. This type of behavior is consistent with the idea that models are designed or parameterized to simulate specific sites, making it rare for a single model to perform well across all sites. It should be noted that ISAM only simulated annual NEE, and the sample size of performance is limited.

TABLE 7. Average model ranking and model-ranking standard deviation (in parentheses) for annual fluxes only, showing photosynthetic formulation.

CROPS		DBF		ENFB		ENFT	
Rank	Model	Rank	Model	Rank	Model	Rank	Model
5.2 (4.2)	*VEGAS2*	5.3 (4.9)	**EC-MOD†**	2.8 (4.0)	**ISAM**	4.3 (2.2)	ORCHIDEE
5.5 (3.3)	CLM-CN	5.9 (4.3)	*MOD17+†*	4.4 (4.0)	DLEM	5.3 (4.1)	LPJ-wsl
5.9 (3.5)	*MOD17+†*	6.0 (2.7)	*CASA-GFEDv2*	4.5 (1.6)	SIB3.1	5.5 (3.7)	CLM-CASA'
6.1 (3.7)	*CASA-GFEDv2*	6.2 (2.7)	LPJ-wsl	5.7 (3.6)	TEM6	6.4 (2.9)	CLM-CN
6.1 (3.7)	**EC-MOD†**	6.6 (3.3)	DLEM	6.0 (4.2)	**EC-MOD†**	6.5 (3.9)	**EC-MOD†**
6.9 (4.2)	DLEM	6.7 (3.3)	CLM-CN	6.2 (3.1)	*VEGAS2*	7.3 (4.6)	SIB3.1
7.1 (2.1)	LPJ-wsl	6.9 (4.5)	ORCHIDEE	6.6 (2.6)	*CASA-GFEDv2*	7.4 (5.5)	*CASA-Trans*
7.5 (4.7)	SIB3.1	7.1 (4.4)	TEM6	6.8 (4.5)	*MOD17+†*	7.8 (4.3)	BEPS
7.6 (4.3)	BEPS	7.2 (4.7)	**ISAM**	7.4 (4.7)	*CASA-Trans*	8.0 (3.7)	*MOD17+†*
8.2 (3.9)	CLM-CASA'	7.8 (4.5)	BEPS	7.9 (4.1)	CLM-CASA'	8.2 (4.8)	*NASA-CASA*
8.4 (4.3)	*NASA-CASA*	8.1 (4.3)	*VEGAS2*	8.0 (3.5)	CLM-CN	8.5 (3.6)	*CASA-GFEDv2*
8.5 (4.4)	ORCHIDEE	9.7 (3.2)	CLM-CASA'	8.3 (2.9)	BEPS	8.6 (3.6)	DLEM
8.8 (3.6)	**ISAM**	9.8 (3.0)	SIB3.1	10.3 (2.8)	LPJ-wsl	9.1 (4.5)	Can-IBIS
9.3 (4.2)	Can-IBIS	11.0 (4.2)	*NASA-CASA*	10.9 (2.6)	Can-IBIS	9.2 (4.1)	**ISAM**
11.2 (6.1)	*CASA-Trans*	11.2 (4.2)	Can-IBIS	11.7 (2.7)	ORCHIDEE	10.8 (5.4)	TEM6
11.4 (3.4)	**MC1**	11.4 (5.1)	*CASA-Trans*			10.9 (2.0)	*VEGAS2*
		12.7 (6.0)	**MC1**			13.4 (7.0)	**MC1**

Notes: The average model ranking is calculated from the individual model rankings (1–17) for every model-PFT-metric combination based upon bias, RMSD, correlation coefficient, sigma ratio, and chi-square. The photosynthetic formulation is displayed as enzyme kinetic in normal type, light-use efficiency in italic type, and statistical in boldface type. A dagger (†) denotes that the model underwent data assimilation. Empty cells indicate that data were not available for this model inter-comparison, meaning either that the models did not simulate GPP or RE, or that they did not simulate specific sites.

TABLE 6. Extended.

RE			
Annual		Monthly	
Rank	Model	Rank	Model
5.2 (2.6)	DLEM	4.9 (3.3)	*MOD17+*†
5.3 (2.4)	SIB3.1	5.0 (4.3)	TEM6
5.5 (4.2)	*VEGAS2*	5.1 (2.9)	**EC-MOD**†
5.8 (2.8)	*MOD17+*†	5.1 (3.7)	*VEGAS2*
5.9 (2.4)	*CASA-GFEDv2*	5.8 (2.3)	*CASA-GFEDv2*
6.0 (3.7)	**EC-MOD**†	6.2 (2.7)	SIB3.1
6.3 (4.2)	TEM6	6.8 (3.8)	CLM-CN
6.4 (3.4)	CLM-CN	6.9 (3.1)	DLEM
7.9 (3.8)	LPJ-wsl	7.4 (3.3)	LPJ-wsl
7.9 (2.6)	BEPS	7.7 (3.2)	BEPS
8.3 (4.4)	ORCHIDEE	8.3 (3.4)	CLM-CASA'
9.6 (2.8)	CLM-CASA'	9.2 (4.0)	ORCHIDEE
9.8 (4.1)	Can-IBIS	11.4 (1.8)	Can-IBIS

The best model performers for all statistics and fluxes based upon monthly time increments only were EC-MOD, CASA-GFEDv2, and SIB3.1 (Table 8). The higher performance of EC-MOD and CASA-GFEDv2 overall is largely influenced by their high performance within the DBF sites. These same models yielded only average performance for ENFT and grass sites. The high performance of EC-MOD is likely, in part, due to the incorporation of the flux tower data in the statistical simulation methodology (Xiao et al. 2008).

When evaluating model performance by individual statistics for annual fluxes (Table 9) the models, in general, segregate into two groupings: those that perform well for bias, RMSD, and chi-square, and those that perform well for correlation (R) and sigma ratio. Although the metrics in the first statistical grouping are related and likely to rank models in a similar order, it is unclear why these same models tend to perform poorly in terms of correlation and sigma ratio. The reverse is also generally true. The best model performers across all statistics include CLM-CN and EC-MOD. It is notable that CLM-CN did not assimilate flux tower data like EC-MOD, yet performed equally as well. On the other hand, it is important to recognize that a relatively simple data-driven model such as EC-MOD does not include the level of detail of a process-oriented CLM-CN. Both CLM-CN and EC-MOD excel at all statistics with the exception of the temporal correlation (R value), where they finish near the bottom of all models. This indicates that these models are most capable of capturing the magnitude and range of variability of annual fluxes, but are mostly incapable of capturing the interannual variation consistent with Fig. 5. Other models that perform consistently well are MOD17+, ISAM, and DLEM, although MOD17+ and ISAM performed poorly for RMSD and sigma ratio, respectively (Table 9).

The best model performers across all statistics based on monthly fluxes were EC-MOD and MOD17+,

TABLE 7. Extended.

GRASS		MISC		All sites	
Rank	Model	Rank	Model	Rank	Model
4.9 (4.0)	CLM-CN	4.8 (1.8)	**ISAM**	5.8 (2.4)	CLM-CN
6.1 (2.6)	SIB3.1	6.1 (5.7)	TEM6	5.9 (4.3)	**EC-MOD**†
6.3 (3.1)	*CASA-GFEDv2*	6.6 (3.0)	CLM-CN	6.4 (4.7)	*MOD17+*†
6.6 (3.6)	*MOD17+*†	6.6 (3.8)	**MC1**	6.6 (3.1)	*CASA-GFEDv2*
6.7 (2.9)	DLEM	6.7 (5.0)	*VEGAS2*	6.6 (4.8)	SIB3.1
6.9 (4.6)	Can-IBIS	6.9 (4.9)	*MOD17+*†	6.7 (4.7)	BEPS
7.6 (8.2)	*NASA-CASA*	7.2 (5.3)	*NASA-CASA*	7.4 (5.1)	**ISAM**
7.7 (2.8)	**EC-MOD**†	7.5 (4.1)	SIB3.1	7.4 (4.0)	TEM6
8.4 (3.4)	**ISAM**	7.6 (2.5)	*CASA-GFEDv2*	7.5 (4.2)	DLEM
8.5 (4.2)	LPJ-wsl	7.8 (4.4)	CLM-CASA'	7.9 (2.7)	LPJ-wsl
8.6 (4.9)	*VEGAS2*	8.1 (4.7)	DLEM	8.0 (5.0)	ORCHIDEE
9.1 (4.4)	BEPS	8.1 (3.8)	ORCHIDEE	8.9 (4.1)	*VEGAS2*
9.1 (3.4)	CLM-CASA'	8.5 (4.4)	BEPS	9.3 (3.9)	CLM-CASA'
9.1 (5.1)	ORCHIDEE	8.7 (3.3)	LPJ-wsl	9.6 (4.8)	*NASA-CASA*
9.2 (7.2)	**MC1**	9.1 (5.0)	**EC-MOD**†	10.4 (4.1)	Can-IBIS
9.4 (4.3)	TEM6	9.2 (7.6)	*CASA-Trans*	11.2 (5.0)	*CASA-Trans*
13.2 (6.9)	*CASA-Trans*	10.6 (3.5)	Can-IBIS	14.0 (5.2)	**MC1**

TABLE 8. Average model ranking and model-ranking standard deviation (in parentheses) for monthly fluxes only, showing photosynthetic formulation.

CROPS		DBF		ENFB		ENFT	
Rank	Model	Rank	Model	Rank	Model	Rank	Model
3.5 (2.6)	**EC-MOD†**	3.7 (2.7)	**EC-MOD†**	3.4 (2.6)	**ISAM**	4.4 (3.9)	CLM-CASA'
3.7 (2.7)	*CASA-GFEDv2*	4.9 (4.0)	*CASA-GFEDv2*	4.1 (3.5)	*MOD17+†*	4.9 (3.1)	ORCHIDEE
5.7 (3.3)	SIB3.1	5.3 (2.6)	*MOD17+†*	4.7 (2.5)	SIB3.1	5.4 (4.1)	LPJ-wsl
6.0 (2.8)	BEPS	6.6 (3.8)	*NASA-CASA*	4.8 (2.4)	*VEGAS2*	5.8 (3.3)	*MOD17+†*
6.1 (2.6)	CLM-CASA'	6.7 (3.3)	BEPS	5.5 (3.8)	**EC-MOD†**	7.1 (4.9)	SIB3.1
6.2 (6.2)	*CASA-Trans*	7.2 (6.4)	*CASA-Trans*	6.4 (3.5)	BEPS	7.4 (4.9)	BEPS
6.5 (4.2)	*VEGAS2*	7.3 (4.9)	*VEGAS2*	6.5 (4.8)	TEM6	7.4 (3.5)	*NASA-CASA*
6.8 (3.4)	*NASA-CASA*	7.6 (6.7)	DLEM	6.7 (3.9)	DLEM	7.5 (3.5)	CLM-CN
7.5 (3.9)	*MOD17+†*	7.7 (3.4)	SIB3.1	7.5 (3.0)	CLM-CN	7.5 (2.3)	**EC-MOD†**
7.9 (3.5)	CLM-CN	7.8 (3.5)	CLM-CASA'	7.9 (2.0)	*CASA-GFEDv2*	7.6 (4.0)	*CASA-GFEDv2*
8.5 (2.8)	LPJ-wsl	8.2 (4.0)	TEM6	8.9 (2.6)	LPJ-wsl	8.9 (3.1)	DLEM
8.8 (4.7)	DLEM	8.5 (4.4)	ORCHIDEE	9.8 (1.6)	*CASA-Trans*	9.0 (4.8)	**ISAM**
9.9 (4.3)	ORCHIDEE	8.9 (4.0)	LPJ-wsl	10.3 (3.9)	CLM-CASA'	9.1 (5.5)	TEM6
11.2 (2.1)	Can-IBIS	9.2 (3.8)	CLM-CN	11.0 (2.5)	Can-IBIS	9.6 (3.5)	*CASA-Trans*
11.2 (2.9)	**ISAM**	11.6 (3.9)	**ISAM**	12.0 (3.4)	ORCHIDEE	10.5 (3.1)	*VEGAS2*
12.0 (4.5)	**MC1**	12.6 (6.1)	**MC1**			12.1 (2.5)	Can-IBIS
		13.3 (2.3)	Can-IBIS			14.6 (4.3)	**MC1**

Notes: The average model ranking is calculated from the individual model rankings (1–17) for every model-PFT-metric combination based upon bias, RMSD, correlation coefficient, sigma ratio, and chi-square. The photosynthetic formulation is displayed as enzyme kinetic in normal type, light-use efficiency in italic type, and statistical in boldface type. A dagger (†) denotes that the model underwent data assimilation. Empty cells indicate that data were not available for this model inter-comparison, meaning either that the models did not simulate GPP or RE, or did not simulate specific sites.

although EC-MOD performed relatively poorly for chi-square and MOD17+ for correlation (*R* value; Table 10). Other models that performed well overall were NASA-CASA, CASA-GFEDv2, and SIB3.1; however, they each performed relatively poorly for sigma ratio and bias, respectively. Although the goal of the model ranking was to identify the best performing models, the ultimate choice of model also depends upon the application. The strength of highly statistical, data-driven models resides in their ability to extrapolate carbon fluxes to create spatial flux maps. Process-based, prognostic models, on the other hand, have additional capacity for prediction and attribution of fluxes.

CONCLUSIONS

We compared the performance of 17 regional TBMs across North America against observations from 36 flux tower observations. Here, we condense our broad range of results into six major findings. First, the regional models significantly underestimated the net carbon sink

TABLE 9. Average model ranking and model-ranking standard deviation (in parentheses) for annual fluxes only, showing photosynthetic formulation.

Bias		RMSD		*R*	
Rank	Model	Rank	Model	Rank	Model
5.4 (3.1)	CLM-CN	5.0 (2.5)	**ISAM**	4.4 (3.7)	*CASA-GFEDv2*
5.4 (4.2)	**EC-MOD†**	5.7 (2.7)	CLM-CN	5.8 (3.6)	*MOD17+†*
5.8 (3.1)	BEPS	5.7 (4.4)	**EC-MOD†**	6.3 (3.5)	ORCHIDEE
5.8 (3.9)	*MOD17+†*	5.8 (4.1)	*NASA-CASA*	6.4 (3.1)	SIB3.1
6.3 (3.7)	**ISAM**	6.2 (4.1)	*VEGAS2*	6.5 (6.9)	*CASA-Trans*
7.1 (4.3)	DLEM	6.3 (4.3)	DLEM	7.3 (4.8)	Can-IBIS
7.7 (5.3)	*VEGAS2*	6.7 (4.5)	TEM6	7.4 (3.4)	DLEM
7.8 (3.1)	*NASA-CASA*	6.7 (4.2)	CLM-CASA'	7.5 (3.4)	**ISAM**
7.8 (6.0)	TEM6	7.1 (2.6)	*CASA-GFEDv2*	7.8 (4.5)	*VEGAS2*
7.9 (3.0)	*CASA-GFEDv2*	7.3 (4.4)	SIB3.1	7.9 (3.4)	BEPS
8.2 (3.6)	CLM-CASA'	7.3 (5.5)	*CASA-Trans*	8.0 (5.7)	TEM6
8.2 (3.0)	LPJ-wsl	7.4 (4.7)	*MOD17+†*	8.5 (3.2)	CLM-CASA'
8.2 (3.7)	SIB3.1	7.8 (3.1)	LPJ-wsl	8.5 (3.4)	LPJ-wsl
8.7 (4.9)	ORCHIDEE	10.1 (4.0)	ORCHIDEE	8.7 (4.5)	CLM-CN
9.0 (7.3)	**MC1**	10.7 (4.0)	BEPS	9.9 (3.7)	**EC-MOD†**
11.2 (2.3)	Can-IBIS	11.8 (1.6)	Can-IBIS	12.2 (6.4)	*NASA-CASA*
12.2 (4.5)	*CASA-Trans*	12.6 (5.7)	**MC1**	13.2 (4.1)	**MC1**

Notes: The average model ranking is calculated from the individual model rankings (1–17) for every model-PFT-metric combination. The photosynthetic formulation is displayed as enzyme kinetic in normal type, light-use efficiency in italic type, and statistical in boldface type. A dagger (†) denotes that the model underwent data assimilation.

TABLE 8. Extended.

GRASS		MISC		ALL SITES	
Rank	Model	Rank	Model	Rank	Model
4.1 (2.6)	*MOD17+†*	3.8 (3.8)	*VEGAS2*	4.7 (3.9)	**EC-MOD†**
5.3 (3.6)	SIB3.1	4.0 (4.5)	TEM6	5.1 (3.2)	*CASA-GFEDv2*
5.8 (4.8)	CLM-CN	5.0 (6.8)	*NASA-CASA*	5.4 (3.7)	SIB3.1
6.5 (5.1)	TEM6	6.3 (2.8)	*CASA-GFEDv2*	6.2 (2.7)	BEPS
7.0 (6.4)	*NASA-CASA*	6.7 (4.3)	**EC-MOD†**	6.5 (3.1)	*MOD17+†*
7.4 (5.9)	*CASA-Trams*	6.9 (3.4)	CLM-CN	6.9 (4.5)	CLM-CASA'
7.7 (2.4)	*CASA-GFEDv2*	7.2 (3.5)	**ISAM**	7.2 (6.1)	*CASA-Trans*
7.9 (2.7)	**EC-MOD†**	7.3 (3.6)	SIB3.1	7.9 (4.2)	TEM6
7.9 (3.5)	CLM-CASA'	7.4 (5.4)	*MOD17+†*	8.1 (3.2)	CLM-CN
7.9 (4.3)	*VEGAS2*	7.6 (5.1)	*CASA-Trans*	8.1 (4.6)	*VEGAS2*
8.5 (3.5)	LPJ-wsl	8.0 (3.0)	BEPS	8.1 (4.1)	DLEM
8.7 (4.1)	DLEM	8.5 (2.7)	LPJ-wsl	8.2 (3.7)	LPJ-wsl
9.3 (4.1)	BEPS	9.1 (3.8)	DLEM	10.1 (5.1)	ORCHIDEE
9.7 (4.6)	Can-IBIS	9.6 (2.6)	CLM-CASA'	11.4 (3.7)	**ISAM**
10.1 (5.6)	ORCHIDEE	11.0 (4.0)	ORCHIDEE	11.4 (4.6)	*NASA-CASA*
10.6 (4.2)	**MC1**	11.8 (5.0)	**MC1**	12.0 (2.3)	Can-IBIS
11.6 (3.6)	**ISAM**	12.5 (2.3)	Can-IBIS	13.0 (5.7)	**MC1**

observed at the flux towers. This finding is reminiscent of previous regional model vs. flux tower comparisons (Friend et al. 2007, Randerson et al. 2009). Though the scale mismatch is large, the flux tower observations are more consistent with the atmospheric inversion models' estimate of the magnitude of the North American carbon sink (Hayes et al. 2012). Applying this finding to Huntzinger et al. (2012), where the same regional models in this study simulated the net ecosystem productivity across North America (–0.7 to 2.2 Pg C/yr), gives credibility to only the models simulating a large carbon sink (positive number is carbon sink). A few regional models in our study accurately estimated the magnitude of the carbon sink; however, this resulted from compensating positive biases in the gross fluxes. Similarly, the models that accurately estimated the annual gross fluxes accomplish this through the underestimation of gross fluxes during the growing season and the overestimation of gross fluxes during the transition seasons (e.g., Richardson et al. 2012). Although biases in regional meteorology data can influence modeled fluxes (Ito and Sasai 2006, Poulter et al. 2011, and Zhao et al. 2011), they do not seem to be the main driver of the flux biases here. The state of the soil and vegetation (choice of spin-up and initialization) are known to influence carbon dynamics (Stoy et al. 2008, Carvalhais et al. 2010) and we hypothesize that this had a significant role in the flux biases. Identifying whether the modeled flux biases are a result of insufficient representation of initial conditions or inherent flaws within the model parameterization/structure (e.g., soil carbon decomposition) should be a focus of future studies.

TABLE 9. Extended.

Sigma ratio		Chi-square	
Rank	Model	Rank	Model
5.5 (4.6)	LPJ-wsl	4.6 (4.3)	*NASA-CASA*
5.6 (3.7)	ORCHIDEE	5.2 (3.5)	**EC-MOD†**
6.2 (4.8)	Can-IBIS	5.7 (3.0)	CLM-CN
6.2 (3.3)	SIB3.1	5.7 (2.9)	**ISAM**
6.2 (2.8)	CLM-CN	6.4 (4.3)	DLEM
7.2 (3.1)	*MOD17+†*	6.8 (2.8)	*CASA-GFEDv2*
7.2 (5.2)	**MC1**	6.9 (4.0)	*VEGAS2*
7.2 (3.8)	DLEM	7.3 (4.9)	*MOD17+†*
7.6 (3.8)	**EC-MOD†**	7.4 (4.1)	CLM-CASA'
7.7 (4.6)	BEPS	7.4 (4.5)	SIB3.1
7.9 (3.3)	*CASA-GFEDv2*	7.7 (4.8)	*CASA-Trans*
8.7 (4.1)	TEM6	8.1 (5.2)	TEM6
9.4 (4.2)	CLM-CASA'	8.4 (3.3)	LPJ-wsl
9.5 (3.4)	*VEGAS2*	8.8 (3.9)	BEPS
9.8 (6.4)	**ISAM**	9.8 (4.5)	ORCHIDEE
12.0 (4.1)	*NASA-CASA*	10.6 (6.7)	**MC1**
16.2 (1.6)	*CASA-Trans*	11.8 (2.4)	Can-IBIS

Second, the models were most successful at simulating seasonal patterns and variability in flux, but unable to simulate the year-to-year flux variability (Braswell et al. 2005, Ricciuto et al. 2008, Dietze et al. 2011, Keenan et al. 2012) and year-to-year magnitude of variation in flux. A single reason for this remains elusive (e.g., Siqueira et al. 2006, Urbanski et al. 2007, Stoy et al. 2009). The fact that the models' performance improves when using local site driver data indicates that the models have some inherent skill in simulating interannual variability. On the other hand, the systemic underestimation of the magnitude of annual variability (sigma) seems to be an inherent property of the model

TABLE 10. Average model ranking and model-ranking standard deviation (in parentheses) for monthly fluxes only, showing photosynthetic formulation.

Bias		RMSD		*R*	
Rank	Model	Rank	Model	Rank	Model
5.3 (2.9)	CLM-CN	4.3 (2.5)	**EC-MOD†**	4.7 (4.8)	TEM6
5.8 (3.1)	BEPS	4.6 (2.5)	*NASA-CASA*	5.3 (5.4)	*CASA-Trans*
5.9 (4.3)	**EC-MOD†**	5.3 (3.1)	*MOD17+†*	5.5 (3.7)	*CASA-GFEDv2*
6.0 (4.2)	*MOD17+†*	5.4 (3.4)	SIB3.1	5.6 (3.6)	SIB3.1
6.3 (3.6)	**ISAM**	6.1 (3.7)	*VEGAS2*	5.6 (4.0)	*NASA-CASA*
6.4 (4.2)	*NASA-CASA*	6.3 (5.0)	TEM6	5.7 (2.9)	CLM-CASA'
7.0 (4.2)	DLEM	6.4 (3.2)	*CASA-GFEDv2*	6.4 (2.7)	**EC-MOD†**
7.5 (5.7)	TEM6	6.4 (2.8)	CLM-CN	7.1 (4.1)	LPJ-wsl
7.6 (5.2)	*VEGAS2*	6.5 (4.0)	CLM-CASA'	7.1 (4.1)	*MOD17+†*
7.9 (3.1)	*CASA-GFEDv2*	6.8 (3.3)	*CASA-Trans*	7.1 (4.4)	ORCHIDEE
8.0 (3.6)	CLM-CASA'	8.3 (3.6)	BEPS	7.3 (4.0)	*VEGAS2*
8.1 (3.2)	LPJ-wsl	8.5 (4.5)	**ISAM**	8.6 (3.5)	CLM-CN
8.3 (3.9)	SIB3.1	9.0 (3.5)	DLEM	9.3 (4.2)	BEPS
8.8 (4.9)	ORCHIDEE	9.3 (3.2)	LPJ-wsl	10.0 (3.1)	**ISAM**
11.4 (2.9)	Can-IBIS	11.3 (4.3)	ORCHIDEE	10.8 (2.2)	DLEM
11.4 (4.6)	**MC1**	13.1 (1.6)	Can-IBIS	12.2 (2.8)	Can-IBIS
12.0 (4.7)	*CASA-Trans*	16.0 (1.2)	**MC1**	15.4 (0.9)	**MC1**

Notes: The average model ranking is calculated from the individual model rankings (1–17) for every model-PFT-metric combination. The photosynthetic formulation is displayed as enzyme kinetic in normal type, light use efficiency in italic type, and statistical in boldface type. A dagger (†) denotes that the model underwent data assimilation.

structure based upon the crossover model analysis. This finding reinforces the need for model structural improvements identified during the site-level synthesis work including improved phenology (Richardson et al. 2012), soil moisture (Schaefer et al. 2012), and vegetation responses to heat and stress (Keenan et al. 2012).

Third, the use of prescribed phenology improves the models' ability to simulate seasonal fluxes, but offers no advantage for annual fluxes. In general, the LUE models are superior in terms of monthly correlation and RMSD statistics. Nevertheless, EK models with prognostic phenology perform better when capturing the (growing-season) magnitude for gross fluxes and are more capable of capturing the magnitude of annual variation for all fluxes. This result is encouraging because models with prognostic phenology are critical for the prediction of fluxes under future climate or disturbance scenarios.

Fourth, the models simulated fluxes the best for deciduous forests, but were poor at simulating crops, grasslands, and evergreen forests. This finding likely reflects the influence of DBF sites upon model development and the strong correlation between phenology and seasonal flux. Simulations of crop and grass sites likely suffer from inaccurate vegetation cover.

Fifth, the site-level model runs performed better than the region-level runs for annual and monthly flux correlation only. The fact that the models' performance improves when using local site driver data indicates that the models have some inherent skill in simulating interannual variability. Nevertheless, even with spatial resolution, vegetation cover, disturbance history, and meteorology data designed specifically to capture site-level fluxes, the site-level runs offered minimal improvement for a majority of statistical metrics. This implies that much of the model–observation mismatch for the regional runs is attributable to shortcomings in model structure, parameters, and setup. As the skill of the models improve, however, we anticipate that spatial mismatch will become a primary source of the overall model–data mismatch. Our ability to diagnose the causes of the model–data divergence is limited because multiple factors could explain these differences. While it is important to document the range of fluxes obtained from an unconstrained model comparison, more limited experiments will be required to diagnose the particular causes of the divergent model performance documented here.

Finally, highly statistical, data-driven approaches can perform better than process-based TBMs built upon detailed ecosystem processes if the goal is to quantify past and present fluxes. We base this conclusion on the finding that EC-MOD and MOD17+ performed the best overall. The performance of these data-driven models is rivaled or exceeded in some cases by process-based models CLM-CN, CASA-GFEDv2, and SIB3.1, perhaps because of their more mechanistically precise descriptions of ecosystem carbon cycling. This finding is important if the goal is not only to quantify past and present fluxes, but for attribution and prediction of fluxes. It seems likely that if flux tower data are assimilated into process-based ecosystem models, model performance should improve even further.

ACKNOWLEDGMENTS

This research was supported by the U.S. Department of Energy's Office of Science through the Northeastern Regional Center of the National Institute for Climatic Change Research and through NASA's Terrestrial Ecology Program. We also acknowledge the DOE Office of Science for support of the three

TABLE 10. Extended.

Sigma ratio		Chi-square	
Rank	Model	Rank	Model
4.9 (4.7)	**EC-MOD†**	3.8 (2.3)	*VEGAS2*
6.1 (4.2)	*MOD17+†*	4.1 (3.2)	*MOD17+†*
6.3 (4.4)	DLEM	5.1 (2.6)	*CASA-GFEDv2*
6.4 (3.8)	**MC1**	5.7 (3.0)	BEPS
6.4 (3.1)	SIB3.1	5.8 (3.9)	SIB3.1
6.8 (3.9)	*CASA-GFEDv2*	6.0 (6.8)	*NASA-CASA*
7.2 (3.8)	CLM-CASA'	6.3 (4.1)	TEM6
7.4 (3.7)	BEPS	6.3 (4.8)	*CASA-Trans*
7.8 (4.7)	ORCHIDEE	7.4 (4.0)	**EC-MOD†**
7.9 (3.9)	LPJ-wsl	8.3 (3.0)	LPJ-wsl
8.4 (4.6)	CLM-CN	8.4 (3.5)	DLEM
9.2 (4.6)	*VEGAS2*	8.6 (3.6)	CLM-CN
9.3 (3.8)	*CASA-Trans*	8.7 (4.3)	**ISAM**
9.3 (4.7)	TEM6	11.1 (2.8)	CLM-CASA'
9.7 (3.9)	Can-IBIS	11.8 (2.5)	Can-IBIS
10.2 (4.5)	*NASA-CASA*	11.9 (3.4)	ORCHIDEE
11.5 (6.1)	**ISAM**	12.4 (4.2)	**MC1**

NACP Interim Synthesis workshops. NASA provided support for the Modeling and Synthesis Thematic Data Center that processed the model output. Ameriflux measurement and data protocols, QA, and coordination of data activities were supported by the U.S. Department of Energy's Office of Science (Science Team Research, Grant Number DE-FG02-04ER63911). J. Xiao was partly supported by National Science Foundation (NSF) through MacroSystems Biology (award number 1065777) and National Aeronautics and Space Administration (NASA) through Carbon Monitoring System (award number NNX11AL32G). B. M. Raczka is grateful to his Ph.D. committee and all those involved in the NACP Interim Synthesis Activity, including the site PIs and modelers who made this work possible.

LITERATURE CITED

Allison, V. J., R. M. Miller; J. D. Jastrow; R. Matamala; D. R. Zak. 2005. Changes in soil microbial community structure in a tallgrass prairie chronosequence. Soil Science Society of America Journal 69:1412–1421.

Bachelet, D., J. M. Lenihan, C. Daly, and R. P. Neilson. 2000. Interactions between fire, grazing and climate change at Wind Cave National Park, SD. Ecological Modelling 134:229–244.

Baker, D. F., et al. 2006. TransCom 3 inversion intercomparison: Impact of transport model errors on the interannual variability of regional CO_2 fluxes, 1988–2003. Global Biogeochemical Cycles 20:GB1002.

Baker, I. T., L. Prihodko, A. S. Denning, M. Goulden, S. Miller, and H. R. da Rocha. 2008. Seasonal drought stress in the Amazon: Reconciling models and observations. Journal of Geophysical Research Biogeosciences 113:G00B01.

Baldocchi, D., et al. 2001. FLUXNET: A new tool to study the temporal and spatial variability of ecosystem-scale carbon dioxide, water vapor, and energy flux densities. Bulletin of the American Meteorological Society 82:2415–2434.

Barr, A. G., T. J. Griffis, T. A. Black, X. Lee, R. M. Staebler, J. D. Fuentes, Z. Chen, and K. Morgenstern. 2002. Comparing the carbon budgets of boreal and temperate deciduous forest stands. Canadian Journal of Forest Research 32:813–822.

Barr, A., D. Hollinger, and A. D. Richardson. 2009. CO_2 flux measurement uncertainty estimates for NACP. EOS Transactions, American Geophysical Union (Fall Meeting Supplement) 90(52):B54A-04.

Beer, C., et al. 2010. Terrestrial gross carbon dioxide uptake: Global distribution and covariation with climate. Science 329:834–838.

Bergeron, O., H. A. Margolis, T. A. Black, C. Coursolle, A. L. Dunn, A. G. Barr, and S. C. Wofsy. 2007. Comparison of carbon dioxide fluxes over three boreal black spruce forests in Canada. Global Change Biology 13:89–107.

Bondeau, A., P. C. Smith, S. Zaehle, S. Schaphoff, W. Lucht, W. Cramer, and D. Gerten. 2007. Modelling the role of agriculture for the 20th century global terrestrial carbon balance. Global Change Biology 13:679–706.

Bradford, J. B., R. A. Birdsey, L. A. Joyce, and M. G. Ryan. 2008. Tree age, disturbance history, and carbon stocks and fluxes in subalpine Rocky Mountain forests. Global Change Biology 14:2882–2897.

Braswell, B. H., W. J. Sacks, E. Linder, and D. S. Schimel. 2005. Estimating diurnal to annual ecosystem parameters by synthesis of a carbon flux model with eddy covariance net ecosystem exchange observations. Global Change Biology 11:335–355.

Burba, G. G., and S. B. Verma. 2005. Seasonal and interannual variability in evapotranspiration of native tallgrass prairie and cultivated wheat ecosystems. Agricultural and Forest Meteorology 135:190–201.

Carvalhais, N., M. Reichstein, G. J. Collatz, M. D. Mahecha, M. Migliavacca, C. S. R. Neigh, E. Tomelleri, A. A. Benali, D. Papale, and J. Seixas. 2010. Deciphering the components of regional net ecosystem fluxes following a bottom-up approach for the Iberian Peninsula. Biogeosciences 7:3707–3729.

Chen, J. M., J. Liu, J. Cihlar, and M. L. Goulden. 1999. Daily canopy photosynthesis model through temporal and spatial scaling for remote sensing applications. Ecological Modelling 124:99–119.

Ciais, P., et al. 2005. Europe-wide reduction in primary productivity caused by the heat and drought in 2003. Nature 437:529–533.

Cook, B. D., P. V. Bolstad, J. G. Martin, F. A. Heinsch, K. J. Davis, W. G. Wang, A. R. Desai, and R. M. Teclaw. 2008. Using light-use and production efficiency models to predict photosynthesis and net carbon exchange during forest canopy disturbance. Ecosystems 11:26–44.

Cook, B. D., et al. 2004. Carbon exchange and venting anomalies in an upland deciduous forest in northern Wisconsin, USA. Agricultural and Forest Meteorology 126:271–295.

Cramer, W., D. W. Kicklighter, A. Bondeau, B. Moore, G. Churkina, B. Nemry, A. Ruimy, A. L. Schloss, and the Participants of the Potsdam NPP Model Intercomparison. 1999. Comparing global models of terrestrial net primary productivity (NPP): overview and key results. Global Change Biology 5:1–15.

Davis, K. J., P. S. Bakwin, C. X. Yi, N. W. Berger, C. L. Zhao, R. M. Teclaw, and J. G. Isebrands. 2003. The annual cycles of CO_2 and H_2O exchange over a northern mixed forest as observed from a very tall tower. Global Change Biology 9:1278–1293.

Desai, A. R., P. V. Bolstad, B. D. Cook, K. J. Davis, and E. V. Carey. 2005. Comparing net ecosystem exchange of carbon dioxide between an old-growth and mature forest in the upper Midwest, USA. Agricultural and Forest Meteorology 128:33–55.

Desai, A. R., et al. 2008. Influence of vegetation and surface forcing on carbon dioxide fluxes across the Upper Midwest, USA: Implications for regional scaling. Agricultural and Forest Meteorology 148:288–308.

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.

Journal of Geophysical Research Biogeosciences 116: G04029.
Fischer, M. L., D. P. Billesbach, W. J. Riley, J. A. Berry, and M. S. Torn. 2007. Spatiotemporal variations in growing season exchanges of CO_2, H_2O, and sensible heat in agricultural fields of the southern Great Plains. Earth Interactions 11:1–21.
Flanagan, L. B., and A. C. Adkinson. 2011. Interacting controls on productivity in a northern Great Plains grassland and implications for response to ENSO events. Global Change Biology 17:3293–3311.
Flanagan, L. B., and K. H. Syed. 2011. Stimulation of both photosynthesis and respiration in response to warmer and drier conditions in a boreal peatland ecosystem. Global Change Biology 17:2271–2287.
Foley, J. A., I. C. Prentice, N. Ramankutty, S. Levis, D. Pollard, S. Sitch, and A. Haxeltine. 1996. An integrated biosphere model of land surface processes, terrestrial carbon balance, and vegetation dynamics. Global Biogeochemical Cycles 10:603–628.
Friedlingstein, P., et al. 2006. Climate-carbon cycle feedback analysis: Results from the (CMIP)-M-4 model intercomparison. Journal of Climate 19:3337–3353.
Friend, A. D., et al. 2007. FLUXNET and modelling the global carbon cycle. Global Change Biology 13:610–633.
Garrigues, S., et al. 2008. Validation and intercomparison of global leaf area index products derived from remote sensing data. Journal of Geophysical Research Biogeosciences 113:G02028.
Gough, C. M., C. S. Vogel, H. P. Schmid, H. B. Su, and P. S. Curtis. 2008. Multi-year convergence of biometric and meteorological estimates of forest carbon storage. Agricultural and Forest Meteorology 148:158–170.
Goulden, M. L., B. C. Daube, S. M. Fan, D. J. Sutton, A. Bazzaz, J. W. Munger, and S. C. Wofsy. 1997. Physiological responses of a black spruce forest to weather. Journal of Geophysical Research Atmospheres 102:28987–28996.
Gu, L., T. Meyers, S. G. Pallardy, P. J. Hanson, B. Yang, M. Heuer, K. P. Hosman, J. S. Riggs, D. Sluss, and S. D. Wullschleger. 2006. Direct and indirect effects of atmospheric conditions and soil moisture on surface energy partitioning revealed by a prolonged drought at a temperate forest site. Journal of Geophysical Research Atmospheres 111:D16102.
Hansen, J., R. Ruedy, J. Glascoe, and M. Sato. 1999. GISS analysis of surface temperature change. Journal of Geophysical Research Atmospheres 104:30997–31022.
Hanson, P. J., et al. 2004. Oak forest carbon and water simulations: model intercomparisons and evaluations against independent data. Ecological Monographs 74:443–489.
Harazono, Y., M. Mano, A. Miyata, R. C. Zulueta, and W. C. Oechel. 2003. Inter-annual carbon dioxide uptake of a wet sedge tundra ecosystem in the Arctic. Tellus Series B Chemical and Physical Meteorology 55:215–231.
Hayes, D. J., A. D. McGuire, D. W. Kicklighter, K. R. Gurney, T. J. Burnside, and J. M. Melillo. 2011. Is the northern high latitude land-based CO_2 sink weakening? Global Biogeochemical Cycles 25:GB3018.
Hayes, D. J., et al. 2012. Reconciling estimates of the contemporary North American carbon balance among terrestrial biosphere models, atmospheric inversions, and a new approach for estimating net ecosystem exchange from inventory-based data. Global Change Biology 18:1282–1299.
Hoffman, F. M., et al. 2007. Results from the Carbon-Land Model Intercomparison Project (C-LAMP) and Availability of the Data on the Earth System Grid (ESG). Journal of Physics Conference Series 78:012026.
Huntzinger, D. N., et al. 2012. North American Carbon Project (NACP) regional interim synthesis: terrestrial biospheric model intercomparison. Ecological Modelling 232:144–157.
Ito, A., and T. Sasai. 2006. A comparison of simulation results from two terrestrial carbon cycle models using three climate data sets. Tellus Series B Chemical and Physical Meteorology 58:513–522.
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 Biogeochemical Cycles 19:GB2015.
Ju, W. M., J. M. Chen, T. A. Black, A. G. Barr, J. Liu, and B. Z. Chen. 2006. Modelling multi-year coupled carbon and water fluxes in a boreal aspen forest. Agricultural and Forest Meteorology 140:136–151.
Jung, M., et al. 2011. Global patterns of land-atmosphere fluxes of carbon dioxide, latent heat, and sensible heat derived from eddy covariance, satellite, and meteorological observations. Journal of Geophysical Research Biogeosciences 116:G00J07.
Keenan, T. F., et al. 2012. Terrestrial biosphere model performance for inter-annual variability of land-atmosphere CO_2 exchange. Global Change Biology 18:1971–1987.
Kljun, N., T. A. Black, T. J. Griffis, A. G. Barr, D. Gaumont-Guay, K. Morgenstern, J. H. McCaughey, and Z. Nesic. 2006. Response of net ecosystem productivity of three boreal forest stands to drought. Ecosystems 9:1128–1144.
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 Biogeochemical Cycles 19:GB1015.
Kucharik, C. J., J. A. Foley, C. Delire, V. A. Fisher, M. T. Coe, J. D. Lenters, C. Young-Molling, N. Ramankutty, J. M. Norman, and S. T. Gower. 2000. Testing the performance of a Dynamic Global Ecosystem Model: Water balance, carbon balance, and vegetation structure. Global Biogeochemical Cycles 14:795–825.
Lafleur, P. M., N. T. Roulet, J. L. Bubier, S. Frolking, and T. R. Moore. 2003. Interannual variability in the peatland-atmosphere carbon dioxide exchange at an ombrotrophic bog. Global Biogeochemical Cycles 17:1036.
Law, B. E., D. Turner, J. Campbell, O. J. Sun, S. Van Tuyl, W. D. Ritts, and W. B. Cohen. 2004. Disturbance and climate effects on carbon stocks and fluxes across western Oregon USA. Global Change Biology 10:1429–1444.
Leemans, R., and W. Cramer. 1991. The IIASA database for mean monthly values of temperature, precipitation and cloudiness of a global terrestrial grid. Report RR-91-18. International Institute for Applied Systems Analysis (IIASA), Laxenburg, Austria.
Lokupitiya, E., S. Denning, K. Paustian, I. Baker, K. Schaefer, S. Verma, T. Meyers, C. J. Bernacchi, A. Suyker, and M. Fischer. 2009. Incorporation of crop phenology in Simple Biosphere Model (SiBcrop) to improve land-atmosphere carbon exchanges from croplands. Biogeosciences 6:969–986.
Luo, H., W. C. Oechel, S. J. Hastings, R. Zulueta, Y. Qian, and H. Kwon. 2007. Mature semiarid chaparral ecosystems can be a significant sink for atmospheric carbon dioxide. Global Change Biology 13:386–396.
Ma, S. Y., D. D. Baldocchi, L. K. Xu, and T. Hehn. 2007. Inter-annual variability in carbon dioxide exchange of an oak/grass savanna and open grassland in California. Agricultural and Forest Meteorology 147:157–171.
Matamala, R., J. D. Jastrow, R. M. Miller, and C. T. Garten. 2008. Temporal changes in C and N stocks of restored prairie: implications for C sequestration strategies. Ecological Applications 18:1470–1488.
McCaughey, J. H., M. R. Pejam, M. A. Arain, and D. A. Cameron. 2006. Carbon dioxide and energy fluxes from a boreal mixedwood forest ecosystem in Ontario, Canada. Agricultural and Forest Meteorology 140:79–96.
Medvigy, D., S. C. Wofsy, J. W. Munger, and P. R. Moorcroft. 2010. Responses of terrestrial ecosystems and carbon budgets to current and future environmental variability. Proceedings of the National Academy of Sciences USA 107:8275–8280.

Mitchell, T. D., and P. D. Jones. 2005. An improved method of constructing a database of monthly climate observations and associated high-resolution grids. International Journal of Climatology 25:693–712.

Moffat, A. M., et al. 2007. Comprehensive comparison of gap-filling techniques for eddy covariance net carbon fluxes. Agricultural and Forest Meteorology 147:209–232.

New, M., M. Hulme, and P. Jones. 2000. Representing twentieth-century space-time climate variability. Part II: Development of 1901–96 monthly grids of terrestrial surface climate. Journal of Climate 13:2217–2238.

Oberbauer, S. F., et al. 2007. Tundra CO_2 fluxes in response to experimental warming across latitudinal and moisture gradients. Ecological Monographs 77:221–238.

Pacala, S., et al. 2007. The North American carbon budget past and present. Chapter 3 *in* A. W. King, L. Dilling, G. P. Zimmerman, D. M. Fairman, R. A. Houghton, G. Marland, A. Z. Rose, T. J. Wilbanks, editors. The first State of the Carbon Cycle Report (SOCCR): The North American carbon budget and implications for the global carbon cycle. A report by the U.S. Climate Change Science Program and the Subcommittee on Global Change Research, Washington, D.C., USA.

Pataki, D. E., and R. Oren. 2003. Species differences in stomatal control of water loss at the canopy scale in a mature bottomland deciduous forest. Advances in Water Resources 26:1267–1278.

Peichl, M., and M. A. Arain. 2007. Allometry and partitioning of above- and belowground tree biomass in an age-sequence of white pine forests. Forest Ecology and Management 253:68–80.

Peters, W., et al. 2007. An atmospheric perspective on North American carbon dioxide exchange: CarbonTracker. Proceedings of the National Academy of Sciences USA 104:18925–18930.

Potter, C., S. Klooster, A. Huete, and V. Genovese. 2007. Terrestrial carbon sinks for the United States predicted from MODIS satellite data and ecosystem modeling. Earth Interactions 11:013.

Poulter, B. D., C. Frank, E. L. Hodson, and N. E. Zimmermann. 2011. Impacts of land cover and climate data selection on understanding terrestiral carbon dynamics and the CO_2 airborne fraction. Biogeosciences Discussion 8:1617–1642.

Randall, D. A., et al. 2007. Climate models and their evaluation. Pages 589–662 *in* S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. B. Averyt, M. Tignor, and H. L. Miller, editors. Climate Change 2007: the physical science basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, UK.

Randerson, J. T., et al. 2009. Systematic assessment of terrestrial biogeochemistry in coupled climate-carbon models. Global Change Biology 15:2462–2484.

Randerson, J. T., M. V. Thompson, T. J. Conway, I. Y. Fung, and C. B. Field. 1997. The contribution of terrestrial sources and sinks to trends in the seasonal cycle of atmospheric carbon dioxide. Global Biogeochemical Cycles 11:535–560.

Rayner, P. J., et al. 2008. Interannual variability of the global carbon cycle (1992–2005) inferred by inversion of atmospheric CO_2 and delta(CO_2)-C-13 measurements. Global Biogeochemical Cycles 22(3):GB3008.

Ricciuto, D. M., M. P. Butler, K. J. Davis, B. D. Cook, P. S. Bakwin, A. E. Andrews, and R. M. Teclaw. 2008. Causes of interannual variability in ecosystem-atmosphere CO_2 exchange. Agricultural and Forest Meteorology 148:309–327.

Ricciuto, D. M., P. E. Thornton, K. Schaefer, R. B. Cook, and K. J. Davis. 2009. How uncertainty in gap-filled meteorological input forcing at eddy covariance sites impacts modeled carbon and energy flux. EOS Transactions, American Geophysical Union (Fall Meeting Supplement) 90(52):B54A–03.

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 Biology 18:566–584.

Richardson, A. D., and D. Y. Hollinger. 2007. A method to estimate the additional uncertainty in gap-filled NEE resulting from long gaps in the CO_2 flux record. Agricultural and Forest Meteorology 147:199–208.

Richardson, A. D., et al. 2006. A multi-site analysis of random error in tower-based measurementsof carbon and energy fluxes. Agricultural and Forest Meteorology 136:1–18.

Richardson, A. D., D. Y. Hollinger, D. B. Dail, J. T. Lee, W. Munger, and J. O'Keefe. 2009. Influence of spring phenology on seasonal and annual carbon balance in two contrasting New England forests. Tree Physiology 29:321–331.

Rogers, B. M., R. P. Neilson, R. Draypek, J. M. Lenihan, J. R. Wells, D. Bachelet, and B. E. Law. 2011. Impacts of climate change on fire regimes and carbon stocks of the U.S. Pacific Northwest. Journal of Geophysical Research 116:G03037.

Ruimy, A., L. Kergoat, A. Bondeau, and the Participants of the Potsdam NPP Model Intercomparison. 1999. Comparing global models of terrestrial net primary productivity (NPP): analysis of differences in light absorption and light-use efficiency. Global Change Biology 5:56–64.

Ryu, Y., D. D. Baldocchi, S. Ma, and T. Hehn. 2008. Interannual variability of evapotranspiration and energy exchange over an annual grassland in California. Journal of Geophysical Research Atmospheres 113:D09104.

Schaefer, K., et al. 2012. A model-data comparison of gross primary productivity: Results from the North American Carbon Program site synthesis. Journal of Geophysical Research Biogeosciences 117:G03010.

Schmid, H. P., C. S. B. Grimmond, F. Cropley, B. Offerle, and H. B. Su. 2000. Measurements of CO_2 and energy fluxes over a mixed hardwood forest in the mid-western United States. Agricultural and Forest Meteorology 103:357–374.

Schwalm, C. R., T. A. Black, K. Morgenstern, and E. R. Humphreys. 2007. A method for deriving net primary productivity and component respiratory fluxes from tower-based eddy covariance data: a case study using a 17-year data record from a Douglas-fir chronosequence. Global Change Biology 13:370–385.

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. Journal of Geophysical Research Biogeosciences 115:G00H05.

Siqueira, M. B., G. G. Katul, D. A. Sampson, P. C. Stoy, J. Y. Juang, H. R. McCarthy, and R. Oren. 2006. Multiscale model intercomparisons of CO_2 and H_2O exchange rates in a maturing southeastern US pine forest. Global Change Biology 12:1189–1207.

Stoy, P. C., G. G. Katul, M. B. S. Siqueira, J.-Y. Juang, K. A. Novick, H. R. McCarthy, A. C. Oishi, and R. Oren. 2008. Role of vegetation in determining carbon sequestration along ecological succession in the southeastern United States. Global Change Biology 14:1409–1427.

Stoy, P. C., et al. 2009. Biosphere-atmosphere exchange of CO_2 in relation to climate: a cross-biome analysis across multiple time scales. Biogeosciences 6:2297–2312.

Strand, M., and G. Oquist. 1985. Inhibition of photosynthesis by freezing temperatures and high light levels in cold-acclimated seedlings of Scots pine (*Pinus sylvestris*). I. Effects on the light-limited and light-saturated rates of CO_2 assimilation. Physiologia Plantarum 64:425–430.

Sulman, B. N., A. R. Desai, B. D. Cook, N. Saliendra, and D. S. Mackay. 2009. Contrasting carbon dioxide fluxes between a drying shrub wetland in Northern Wisconsin, USA, and nearby forests. Biogeosciences 6:1115–1126.

Suyker, A. E., and S. B. Verma. 2008. Interannual water vapor and energy exchange in an irrigated maize-based agroecosystem. Agricultural and Forest Meteorology 148:417–427.

Taylor, J. R. 1997. An introduction to error analysis. Second edition. University Science Books, Herndon, Virginia, USA.

Taylor, K. E. 2001. Summarizing multiple aspects of model performance in a single diagram. Journal of Geophysical Research Atmospheres 106:7183–7192.

Thomas, C. K., B. E. Law, J. Irvine, J. G. Martin, J. C. Pettijohn, and K. J. Davis. 2009. Seasonal hydrology explains interannual and seasonal variation in carbon and water exchange in a semiarid mature ponderosa pine forest in central Oregon. Journal of Geophysical Research Biogeosciences 114:G04006.

Thornton, P., et al. 2002. Modeling and measuring the effects of disturbance history and climate on carbon and water budgets in evergreen needleleaf forests. Agricultural and Forest Meteorology 113:185–222.

Thornton, P. E., S. C. Doney, K. Lindsay, J. K. Moore, N. Mahowald, J. T. Randerson, I. Fung, J. F. Lamarque, J. J. Feddema, and Y. H. Lee. 2009. Carbon-nitrogen interactions regulate climate-carbon cycle feedbacks: results from an atmosphere-ocean general circulation model. Biogeosciences 6:2099–2120.

Tian, H. Q., X. Xu, M. Liu, W. Ren, C. Zhang, G. Chen, and C. Lu. 2010. Spatial and temporal patterns of CH_4 and N_2O fluxes in terrestrial ecosystems of North America during 1979–2008: application of a global biogeochemistry model. Biogeosciences 7:2673–2694.

Urbanski, S., C. Barford, S. Wofsy, C. Kucharik, E. Pyle, J. Budney, K. McKain, D. Fitzjarrald, M. Czikowsky, and J. W. Munger. 2007. Factors controlling CO_2 exchange on timescales from hourly to decadal at Harvard Forest. Journal of Geophysical Research Biogeosciences 112:G02020.

van der Werf, G. R., J. T. Randerson, G. J. Collatz, L. Giglio, P. S. Kasibhatla, A. F. Arellano, S. C. Olsen, and E. S. Kasischke. 2004. Continental-scale partitioning of fire emissions during the 1997 to 2001 El Nino/La Nina period. Science 303:73–76.

van der Werf, G. R., J. T. Randerson, L. Giglio, G. J. Collatz, P. S. Kasibhatla, and A. F. Arellano. 2006. Interannual variability in global biomass burning emissions from 1997 to 2004. Atmospheric Chemistry and Physics 6:3423–3441.

Viovy, N., C. Francois, A. Bondeau, G. Krinner, J. Polcher, L. Kergoat, G. Dedieu, N. De noblet, P. Ciais, and P. Friedlingstein. 2000. Assimilation of remote sensing measurements into the ORCHIDEE/STOMATE DGVM biosphere model. Pages 713–716 *in* Proceedings of the 8th International Symposium on Physical Measurements and Signatures in Remote Sensing. 8–12 January 2001, Aussois, France.

Waring, R. H., and S. W. Running. 2007. Forest ecosystems: analysis at multiple scales. Third edition. Elsevier Academic, Burlington, Massachusetts, USA.

Xiao, J., et al. 2008. Estimation of net ecosystem carbon exchange for the conterminous United States by combining MODIS and AmeriFlux data. Agricultural and Forest Meteorology 148:1827–1847.

Xiao, J., et al. 2010. A continuous measure of gross primary production for the conterminous U.S. derived from MODIS and AmeriFlux data. Remote Sensing of Environment 114:576–591.

Yang, X. J., 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 Biogeochemical Cycles 23:GB4029.

Zeng, N., A. Mariotti, and P. Wetzel. 2005. Terrestrial mechanisms of interannual CO_2 variability. Global Biogeochemical Cycles 19:GB1016.

Zeng, N., H. F. Qian, E. Munoz, and R. Iacono. 2004. How strong is carbon cycle-climate feedback under global warming? Geophysical Research Letters 31:L20203.

Zhao, Y., P. Ciais, P. Peylin, N. Viovy, B. Longdoz, and J. M. Bonnefond, et al. 2011. How errors on meteorological variables impact simulated ecosystem fluxes: a case study for six French sites. Biogeosciences Discussion 8:5467–2522.

Supplemental Material

Appendix A

Tables and figures providing statistical support including representation of the differences between the regional and site level protocols and of model bias, correlation, and variability statistics (*Ecological Archives* M083-018-A1).

Appendix B

Tables and figures providing a statistical representation of the observed flux tower carbon fluxes grouped by the plant function type of the site locations (*Ecological Archives* M083-018-A2).

Data Availability

Data associated with this paper have been deposited in the Oak Ridge National Laboratory Distributed Active Archive Center for Biogeochemical Dynamics: http://dx.doi.org/10.3334/ORNLDAAC/1183