A model-data intercomparison of CO_2 exchange during a large scale drought event: Results from the NACP Site Synthesis

Christopher Schwalm, Christopher Williams, Kevin Schaefer, NACP contributors/collaborators

Manuscript for submission to JGR-Biogeosciences
Original version: October 7, 2009
Current version: November 24, 2009
Last saved: November 24, 2009

Abstract

Our current understanding of terrestrial carbon processes is represented in various models that are routinely used to integrate and scale observations of CO_2 exchange with remote sensing and other data in time and space. Yet rigorous accuracy assessments are rarely conducted to determine how well they mimic carbon processes across a range of vegetation types and environmental conditions. Here we compare observed and simulated integrals of CO_2 exchange using 44 eddy covariance flux towers in North America and model runs from 22 terrestrial biosphere models. The observational time record spans ≈ 220 site-years over 10 biomes and overlaps with 2 large scale drought events in North America, providing a natural experiment to test model fidelity to observations as a function of drought and seasonality. We also evaluate the ability of the models to simulate the seasonal cycle of monthly CO_2 exchange using Taylor diagrams. Overall model performance was poor (< 0.50 predictive skill) with forested ecosystems better predicted than non-forested. Models achieved the highest levels of predictive skill in summer and winter in all biomes and in temperate evergreen forests during all climatic seasons. In contrast, model performance was consistently poor in spring and fall, especially in wetlands and mixed or deciduous forests, and during abnormally dry or wet periods. The best performing models were either highly specialized to specific biome types or used data assimilation techniques. Generalist models applied over a range of sites and biomes also performed well, with predictive skill in the upper quintile of all models.

Keywords: carbon modeling, ecosystem models, model validation, carbon exchange, drought, North American Carbon Program

Introduction

Drought is a reoccurring phenomenon in all climates (Larcher, 1995) and is characterized by a loss in plant function due to water limitation and heat stress. For terrestrial CO_2 exchange, drought typically reduces photosynthesis more than respiration (Baldocchi, 2008; Ciais et al., 2005; Schwalm et al., 2009), resulting in decreased net carbon uptake from the atmosphere. In the recent past drought conditions have become more prevalent globally (Dai et al., 2004) and in North America (Cook et al., 2004). Both incidence and severity of drought (Seager et al., 2007b) as well as heatwaves (Meehl et al., 2004) are expected to further increase in conjunction with an intensification of the hydrological cycle; itself linked to global warming (Houghton et al., 2001; Huntington, 2006; Sheffield & Wood, 2008; Trenberth et al., 2007).

Forecasting the onset, amplitude, and duration of drought (Hwang & Carbone, 2009; Mishra & Desai, 2006; Sheffield & Wood, 2008) as well as CO_2 exchange under more frequent drought events related to climate change requires the use of models. Past validation studies of terrestrial biosphere models focused only on few models and sites typically in close proximity and primarily in forested biomes with differing objectives, levels of calibration to site data, and parameter values (e.g., Amthor et al., 2001; Delpierre et al., 2009; Grant et al., 2005; Hanson et al., 2004; Granier et al. 2007; Ichii et al., 2009; Ito, 2008; Siqueira et al., 2006; Zhou et al., 2008). Furthermore, assessing

model performance relative to drought requires high quality observed CO_2 exchange data, a reliable drought metric as well as a natural experiment across sites and drought conditions.

In this study we evaluate model performance using terrestrial CO_2 flux data and simulated fluxes collected from 1991 to 2007. This timeframe included 2 widespread droughts in North America: (i) the turn-of-the-century drought from 1998 to 2004 that was centered in the interior West of North America (Seager et al., 2007a) and (ii) a smaller-scaled drought event in the southern continental Untied States from winter of 2005/2006 through October 2007 (Seager et al., 2009). During these events low Palmer Drought Severity Index values (Cook et al., 2007; Dai et al., 2004) and negative precipitation anomalies (Seager et al., 2007a; 2009) occurred over broad geographic areas which, in conjunction with ongoing eddy covariance measurements (Baldocchi et al., 2001), provided flux data across gradients of time, space, seasonality, and drought. We use this data to examine model performance relative to site-specific drought severity, climatic season, and time and to link model behavior to model architecture and site-specific attributes. Specifically, we address the following questions: Are current state-of-the-art terrestrial biosphere models capable of simulating CO_2 exchange subject to gradients in dryness and seasonality? Are these models able to reproduce the seasonal variation of observed CO_2 exchange across sites? Are certain characteristics of model structure coincident with better/worse model performance? Which biomes are simulated poorly/well?

Methods

Modeled and observed net ecosystem productivity (NEP, net carbon balance including soils where negative values indicate outgassing of CO_2 to the atmosphere) data were analyzed from 22 terrestrial biosphere models (Table 1) and 44 eddy covariance (EC) sites spanning ≈ 220 site-years and 10 biomes in North America (Table 2). Simulated NEP was based on model-specific runs using gap-filled observed weather at each site and locally observed values of soil texture according to a standard protocol (http://isynth-site.pbworks.com). Each model, except the mean model ensemble and an assimilated model, was spun up to steady state initial conditions with a target NEP of zero integrated over the simulation period.

Gaps in the meteorological data record occurred at EC sites due to data screening or instrument failure. Missing values of air temperature, humidity, shortwave radiation, and precipitation data, i.e., key model inputs, were filled using DAYMET (Thornton et al.,1997) before 2003 or the nearest available climate station in the National Climatic Data Center's Global Surface Summary of the Day (GSOD) database. Daily GSOD and DAYMET data were temporally downscaled to hourly or half-hourly using the phasing from observed mean diurnal cycles calculated from a 15-day moving window.

EC data were produced by AmeriFlux and Fluxnet Canada investigators and processed as a synthesis product of the North American Carbon Program (NACP) Site Level Interim Synthesis (http://www.nacarbon.org/nacp/). The observed NEP were corrected for storage, despiked, filtered to remove conditions of low turbulence (friction velocity

filtered), and the gaps filled to create a continuous time series (Barr et al., 2004). The time series included estimates of random uncertainty and uncertainty due to the friction velocity filtering (Barr et al., 2004). In this analysis NEP was aggregated to monthly integrals using only non-gap-filled data, i.e., observed values deemed spurious and subsequently infilled were not considered. Coincident modeled NEP values were similarly excluded. This removed the influence of gap-filling algorithms in the comparison of observed and modeled NEP.

Drought level was quantified using the 3-month Standard Precipitation Index (SPI, McKee et al., 1993). Monthly SPI values were taken from the U.S. Drought Monitor (http://drought.unl.edu/DM/) whereby each tower was matched to nearby meteorological station(s) indicative of local drought conditions given proximity, topography, and human impact. This study used 3 drought levels: Dry required SPI < -0.8, wet corresponded to SPI $> +0.8$, otherwise normal conditions existed. Climatic season was defined by 4 seasons of 3 months each with winter given by December, January, and February.

Model performance was evaluated using data groups defined by all possible intersections of site, model, climatic season, and drought level with at least 3 pairs of observed and simulated monthly NEP. All available months within each data group were simultaneously evaluated for the presence or absence of predictive skill based on correlation (ρ), root mean squared error (RMSE), and the standard error or observed NEP (SE):

$$\rho_{ijkd}^{o,m} = \frac{\left(NEP_{ijkd}^{o} - \overline{NEP_{ijkd}^{o}}\right)\left(NEP_{ijkd}^{m} - \overline{NEP_{ijkd}^{m}}\right)}{\sqrt{\sigma_{NEP_{ijkd}^{o}}\sigma_{NEP_{ijkd}^{m}}}},$$

$$RMSE_{ijkd}^{o,m} = \left(\frac{\sum\left(NEP_{ijkd}^{o} - NEP_{ijkd}^{m}\right)^{2}}{N_{ijkd}}\right)^{0.5},$$

where the overbar indicates averaging across all values, σ is the standard deviation across all months, N is the amount of paired observed and modeled monthly NEP integrals in a data group, subscript i is for site, j is for model, k is for climatic season, d is for drought level, and superscript o is for observations and m is for modeled estimates.

SE comprised 2 components calculated explicitly for monthly NEP integrals: random uncertainty and uncertainty associated with the friction velocity threshold (u_*^{Th}). Random uncertainty was estimated following Richardson & Hollinger (2007): (i) generate synthetic NEP data using the gap-filling model (Barr et al., 2004), (ii) introduce gaps as in the observed data with u_*^{Th} filtering, (iii) add noise, (iv) infill gaps in synthetic data, (v) repeat the process 1000 times. The random uncertainty SE component was then the standard deviation across all bootstrap replicates. The u_*^{Th} uncertainty component was also estimated using the bootstrap. Here 1000 bootstrap replicates of NEP were generated using a distribution of u_*^{Th} based on binning the raw flux data with respect to climatic season, temperature, and site-year. The standard deviation across all replicates gave the SE component associated with u_*^{Th}. These two components were then combined in quadrature, i.e., $z = \sqrt{x^2 + y^2}$, to estimate the total SE of monthly observed NEP.

The presence of predictive skill (π_{ijkd}) required (i) that modeled and observed flux data were directly related (not anti-correlated) with at least a weak correlation serving as a threshold, (ii) that model error was less than the mean observed flux, i.e., that a discrimination between source and sink was possible, and (iii) that the variability of model-data agreement was scaled to the uncertainty of observed and simulated fluxes. Mathematically, predictive skill for a single data group was present ($\pi_{ijkd} = 1$) if the following three conditions were satisfied:

$$(i) \quad \rho_{ijkd}^{o,m} \geq 0.2$$

$$(ii) \quad \frac{RMSE_{ijkd}^{o,m}}{\overline{NEP_{ijkd}^{o}}} \leq 1$$

$$(iii) \quad \overline{NEP_{ijkd}^{o}} \pm 2\overline{SE}_{ijkd}^{o} \cap \overline{NEP_{ijkd}^{m}} \pm 2RMSE_{ijkd}^{o,m}$$

Otherwise there was no predictive skill ($\pi_{ijkd} = 0$). This framework was used to aggregate across data groups as well. For example, calculating predictive skill for each model (π_{model}) required integration across all sites, climatic seasons, and drought levels by model. Similarly, biome-specific predictive skill (π_{biome}) was calculated by integrating across all sites, models, climatic seasons, and drought levels within a given biome. When aggregating in this manner the group-wise predictive skills were linearly combined weighted by *N*.

We defined the consistency of predictive skill by model (CV_π) as the coefficient of variation, in percent, of π across all sites simulated: $CV_n = \frac{\sigma_\pi}{\overline{\pi}}$, where σ_π is the standard deviation of π across site by model. Consistency addresses the degree to which model

performance was related to model parameter set (site or biome unique constants) and assumed that the effect of any errors in model structure and meteorological forcings on π was randomly distributed. High consistency (relatively low CV_{π}) indicated less error in model parameter set relative to a set level of predictive skill and vice versa.

The existence of predictive skill was then related to biome, climatic season, drought level, site history and model structure using regression tree analysis (RTA) as a supervised classification algorithm (Table 3). RTA is a form of binary recursive partitioning (Breiman et al., 1984) that successively splits the data into subsets (nodes) by minimizing within-subset variation. The result is a pruned tree-like topology whereby predicted values (presence/absence of predictive skill) are derived by a top-to-bottom traversal following the rules (branches) that govern subset membership until a predicted value is reached (terminal node). The splitting rules at each node as well as its position allow for a calculation of relative variable importance (Breiman et al., 1984) with the most important variable given a score of 100.

A second characterization of model performance was done using Taylor diagrams (Taylor, 2001); visual displays based on pattern matching, i.e., the degree to which simulations matched the temporal evolution of monthly NEP. Taylor plots are polar coordinate displays of ρ, RMSE, and σ of NEP where all 3 quantities were calculated using the full data record by site and model (ranging from 7 to 178 months). Taylor diagrams were constructed for the mean model ensemble and across-site mean model performance. Each point representing a model was subsequently scored using a skill

metric: $S = 2(1+\rho)/(\sigma_{norm} + 1/\sigma_{norm})^2$, where σ_{norm} is the ratio of simulated to observed σ (Taylor, 2001), bound by zero and unity where unity indicated perfect agreement. This scoring allowed a direct comparison with predictive skill (π), also bound by zero and unity. In sum, the calculation of predictive skill evaluated model performance relative to ecological controls on CO_2 exchange independent of time whereas the Taylor displays examined models' ability to mimic the monthly trajectory of observed NEP.

Results

Model performance based on predictive skill

Predictive skill was related to drought level, climatic season, model, and site ($p = 0.01$). Of the 26747 paired data-model months (across 3479 data groups) available for analysis 10530 ($\pi_{overall} = 0.39$) showed predictive skill. Overall performance was higher in forested than non-forest biomes with the highest performance levels occurring during periods of peak biological activity (climatic summer) under non-drought conditions (Table 4). While drought was a significant control on predictive skill the aggregated contrast was only significant for non-normal (abnormally wet and dry) vs. normal conditions.

Integrated across all models and sites predictive skill was, both within and across drought level, higher in summer and winter than in spring and fall (Table 4). A similar pattern, except for ENFT, was present at the biome level (Table 5) with severe losses in predictive skill occurring during climatic spring and fall, especially in DBF, MF, and WET. Across the 3 levels of dryness π_{biome} declined for non-normal hydrological conditions apart from CRO and ENFB where abnormally dry and wet conditions

respectively were better predicted. Overall biome-level predictive skill was loosely ranked in 4 tiers: ENFT > ENFB, MF, DBF, CRO > GRA, WET, WSA > SHR. These rankings were robust across models used in the majority of biomes although some divergence was apparent for MF and WET (Figure 1).

By model, predictive skill showed, similar to the biome level, a sharp decline in spring and fall with the best model performance achieved through data assimilation (LoTEC). Also, most models showed the highest performance in normal hydrological conditions (Figure 2). Overall performance by model (π_{model}) ranged from 0.04 to 0.68 (Figure 3); CV_{π} had larger amplitude, ranging from 34% to 174%. These 2 characterizations of model performance were inversely related ($r = -0.8$; $p < 0.05$) i.e., higher levels of predictive skill were associated with higher levels of consistency in predictive skill.

Among crop models, SiBCrop and AgroIBIS performed well, especially in climatic spring and during dry conditions. In contrast, DNDC exhibited 0.04 predictive skill overall, $\pi_{model} = 0.55$ in climatic winter and zero otherwise. All crop only models ($n = 4$) showed $\pi_{model} = 0$ for wet conditions. While the best model performance across all CRO sites was achieved by a specialist model (AgroIBIS: $\pi_{biome} = 0.56$), the generalist Ecosys model, used on 26 sites in 9 biomes, achieved a similar level of performance ($\pi_{biome} = 0.54$). Overall, i.e., across all sites, dryness levels, and climatic seasons, LoTEC had the highest performance (0.68). This platform was optimized using a data assimilation technique, unique among model runs evaluated here, and was applied at 10 sites. Performance by individual models ($n = 10$) applied to a wider range of sites (at least 24

sites) was highest for SibCASA (0.51) and lowest for TECO (0.24). Model performance in excess of 0.50 was achieved only with the mean model ensemble (MEAN), a generalist model (SibCASA), an optimized platform (LoTEC), crop models (AgroIBIS, SiBcrop) and 1 model used on 9 sites (ISOLSM) (Figure 3).

Site-level predictive skill also showed a high degree of variability. In 2 instances (1 forest and croplands site) predictive skill across all models for a given site ranged from zero to unity. Even for the best predicted site on average (US-Ho1) π_{model} ranged from 0.43 to 1 (Figure 4). Alternatively, no one site was predicted well ($\pi_{model} > 0.5$) by all models; whereas 4 sites (CA-Let, US-ARM, US-SO2, and US-Ton) were consistently poorly predicted ($\pi_{model} < 0.5$). In contrast, mean model performance was notably high at some sites as a function of drought level and seasonality: US-Dk3 ($\pi_{site} = 0.88$), US-Ho1 (0.81), and US-Ho1 (0.85) for dry, normal, and wet conditions respectively as well as US-Me2 (0.79), US-Ho1 (0.98), US-PFa and US-WCr (0.90), and US-Ho1 (0.94) for winter, spring, summer, and fall respectively. Despite the wide range in model performance, π_{model} was not related to the number of sites ($p = 0.9$) nor biomes ($p = 0.7$) simulated, i.e., using a more general rather than a specialized model did not result in a loss in model performance. Similarly, model performance was not better at sites with longer term records on average ($p > 0.16$), although ORCHIDEE did perform marginally better on sites with more data ($p < 0.09$).

Factors governing the distribution of predictive skill

Biome and climatic season were the most important factors in the distribution of predictive skill (Figure 5) followed by 2 model structural attributes: the number of dynamic vegetation pools and overall model complexity (Table 3). Stand age was the most important site-specific attribute whereas only 3 of the 13 evaluated site disturbances achieved an importance score of at least 10 and were of tertiary importance overall. Models with lower complexity (the lowest 2 quintiles of overall complexity) performed better at ENFT sites and models with at least 5 dynamic vegetation pools faired better at non-ENFT sites. For SHR and WSA no model structural attribute was coincident with the presence of predictive skill; these biomes were always poorly predicted. While the regression tree algorithm achieved an accuracy of 79% for predicting the presence or absence of predictive skill, the site and model characteristics considered here did not explain the underlying cause of biome and seasonal differences in predictive skill.

Overall model performance using Taylor diagrams

Average model performance (both across-site and across-model) was evaluated using Taylor diagrams based on all simulated and observed NEP monthly integrals. Better model performance was indicated by higher S (Figure 3) values and proximity to the benchmark. The benchmark was normalized by observed standard deviation such that the distance of σ and RMSE from the benchmark was in observed σ units. Similar to predictive skill, forested sites were better predicted than non-forested ones. The MEAN model, i.e., average model performance across all models, at all sites (excluding CA-SJ2 and US-Atq) showed $\rho \geq 0.2$ but generally (33 of 44 sites) underpredicted the variability associated with monthly NEP at forested (Figure 6) and non-forested (Figure 7) sites.

Similarly, 40 of 44 sites were predicted with RMSE < σ_{obs}. Also 8 (6 forested and 2 croplands sites: CA-Obs, CA-Qfo, CA-TP4, US-Ho1, US-IB1, US-MMS, US-Ne3, US-UMB) of the 44 sites were predicted with $\rho \geq 0.95$ and RMSE < 1. The worst predicted site was CA-SJ2 with $\rho = -0.67$, $\sigma = 4.3$, and RMSE = 5.1.

Overall model performance, aggregated across sites, was similar (Figure 8). Most models underpredicted variability and showed RMSE < σ_{obs}. Of all 22 models only DNDC exhibited $\rho < 0.2$. Based on proximity to the benchmark, i.e., a high *S* value (Figure 3), the best models were: EPIC (crop only model used one 1 site), ISOLSM (used on 9 sites), LoTEC (data assimilation model), SiBcrop and AgroIBIS (crop only models), EDCM (used on 10 sites), ecosys and SiBCASA (more generalist models used on 39 and 35 sites respectively), and MEAN (mean model ensemble). All of these "best" models had $\rho >$ 0.8, RMSE < 0.8 and slightly underpredicted variability; except the crop only models and Ecosys where variability was overpredicted. Models whose average behavior was furthest away from the benchmark were DNDC followed by BEPS. Finally, the general model rankings with regards to any single measure of model performance were preserved, i.e., higher π was coincident with higher *S* and lower CV_{π} ($p < 0.05$; $|r| \approx 0.5$).

Discussion

As expected no single model performed well across the full spectrum of dryness, seasonality, and sites simulated. Exact agreement between simulations and observations is not however a realistic expectation: models can match observations only within the observational uncertainty. While flux measurements are useful for validating and

improving the process-based simulation of CO_2 exchange (Friend et al., 2007), accounting for the weaknesses of the observational record is central to model validation (Grant et al., 2005; Hanson et al., 2004). In this study the uncertainty associated with monthly NEP integrals was used to define predictive skill. This created an observational interval, as opposed to a point estimate, allowing for a more accurate estimate of predictive skill. Nonetheless model performance was low (there were few instances where $\pi > 0.50$), underscoring the continued need for models to improve fidelity to observations (Friedlingstein et al., 2006).

Contrasting predictive skill and the skill metric based on the Taylor diagrams (S vs. π) highlighted difficulties in validating process-based models. Taylor diagrams were indexed to time only whereas predictive skill was evaluated in the context of a factorial design with 3 levels of dryness and 4 levels of seasonality. Overall, models were better in capturing the temporal dynamics of monthly NEP than observed relationships between carbon flux, dryness, and seasonality. However averaging over these ecologically relevant factors can cause under- and overpredictions to cancel (Medlyn et al., 2005). Alternatively, aggregate results can overestimate model fidelity to observations at shorter timescales. It is noteworthy that metrics of performance skill were closely related. This implies that evaluating model performance, especially relative to other models, is possible using temporal evolution only but that doing so has limited diagnostic ability in terms of highlighting model weaknesses and overestimates how well a model captures key processes of carbon metabolism.

Effect of model structure and parameterization on model performance

The model characteristics considered here could not explain the underlying causes of the biome and seasonal differences in predictive skill (Figure 5). Lower variable importance scores revealed only a subsidiary link between predictive skill and the model structural attributes evaluated here: temporal resolution, types of algorithms for carbon balance terms, overall model complexity, number of vegetation and soil pools, soil layers, and diagnostic vs. prognostic leaf area index. In contrast, the effect of model parameterization was larger than model structure. Model parameter sets are a large source of variability in terms of model performance (Jung et al., 2007b). They influence output and accuracy (Grant et al., 2005) and are more important for simulating CO_2 exchange than interannual climatic variability (Amthor et al., 2001). This is related to the use of biome-specific parameters relative to within-biome variability (Purves & Pacala, 2008). A corollary occurs in the context of validating EC observations as tower footprints can exhibit heterogeneity, particularly in soils, that is not reproduced in model site-specific parameters (Amthor et al., 2001).

The importance of model parameter sets was visible in this intercomparison in two ways. Firstly, biome had the highest variable importance score. Insomuch as models rely on biome-specific parameter values this finding indicates model parameter sets as a key factor in the distribution of predictive skill. Furthermore, the high degree of variability in predictive skill across sets of biome-specific constants underscores that biomes may be too heterogeneous in time (Stoy et al., 2005; 2009) and space to assume constant parameter values relative to within-biome climate partitions (Hargrove et al. 2003).

Secondly, given that some model constants are site-specific scoring models for consistency (calculating the coefficient of variation of predictive skill across all sites for each model, Figure 1) highlighted the uniqueness of model results related to model parameter sets using site-specific runs as a proxy. That is, the high degree of within-model variation in model skill suggested that model parameter sets may need to be refined to capture local, site-specific realities.

Links between model performance and environmental factors

Drought level was significantly related to predictive skill with the largest effect occurring during climatic summer when dry (and wet) conditions were associated with a decline in model performance. As summer corresponds to peak biological activity, the ability of models to simulate NEP during anomalous climatic conditions influences carbon budgets on longer time scales, i.e., process uncertainty with regards to dry (Sitch et al., 2008) and wet conditions must be reduced to enhance model utility. While normal hydrological conditions were generally better predicted exceptions to this rule occurred during climatic fall as well as in ENFB and CRO where non-normal conditions showed the highest degree of predictive skill. This indicated that the variability around normal conditions was large and models' behavior insensitive, i.e., changes in driving variables within the range of normal conditions were too subtle for the model but not for the simulated sites. Finally, ecosystem response to longer-term drought can exhibit lags and positive feedbacks (Arnone et al., 2008; Granier et al., 2007; Thomas et al., 2009; Williams et al., 2009) that were not explicitly included in the drought metric used in this study.

In spring and fall, corresponding to leaf initiation and senescence for biomes with a significant deciduous component, models showed the least predictive skill (see also Morales et al., 2005). Springtime phenological cues influence the annual carbon balance at point to continental scales (Barr et al., 2007; Delpierre et al., 2009; Keeling et al., 1996). Similarly how well simulated fluxes fit observations has been linked to correctly reproducing leaf initiation and senescence (Hanson et al., 2004). However, the observed trend toward earlier leaf out does not appear coincident with an increase in the terrestrial sink due to warmer/drier summers (Angert et al., 2005; Piao et al, 2008). This suggests that a possible reason for the general loss of predictive skill in spring and fall was related to how the initiation and cessation of canopy photosynthesis were modeled relative to soil moisture and soil temperature (Waring & Running, 2007). In wetlands a similar decrease in predictive skill was observed in the shoulder season. This was likely associated with the general inability of models to simulate assimilation and respiration of ericaceous shrubs and mosses common to wetlands (St-Hilaire et al., 2008).

ENFT and ENFB diverged in performance during fall and winter although these biomes differ in climatic zone only. A similar divergence was reported using Biome-BGC, LPJ and ORCHIDEE to simulate gross CO_2 uptake across a temperature gradient in Europe (Jung et al., 2007a); average relative RMSE was higher for ENFB. This was linked to an overestimation of leaf area index at the boreal sites and relationships between resource availability and leaf area (Friedlingstein et al., 2006; Jung et al., 2007a; Stich et al., 2008). Additionally, recent observations in the circumboreal region, where all ENFB sites are located, suggest that transient effects of climate change, e.g., increased severity and

intensity of natural disturbances (fire, pest outbreaks) and divergence from climate normals in temperature, have already occurred (Soja et al., 2007). We speculate the loss of predictive skill in ENFB relative to ENFT was linked to insufficient characterization of cold temperature sensitivity of metabolic processes and water flow in plants as well as freeze-thaw dynamics (Schaefer et al., 2007; 2009) and that this was exacerbated by the effects of transient climate change.

Site history and constraints on model evaluation

Disturbance regime and how a model treats disturbance are known to impact model performance (Ito, 2008). In this study stand age and, to a lesser extent, planting, grazing, and fire (but not harvest) were controls on predictive skill. However comparing sites with and without fire activity showed no significant difference in mean predictive skill ($p > 0.10$), i.e., disturbance and age were confounded. Furthermore, sites with a recent stand replacement disturbance were poorly predicted, e.g., CA-SJ2, the worst predicted site, was harvested in 2000 and scarified in 2002, and US-SO2, a second poorly predicted site, suffered catastrophic wildfire during the analyzed data record. The poor model performance for recently disturbed sites follows from the steady state spinup used in simulation and the absence of an explicit disturbance layer. However, the distribution of site history metrics was skewed; only few sites were in the early stages of recovery from disturbance when NEP is more nonlinear relative to established stands. Furthermore, age class was biased toward older stands; of the 17 forested sites only 1 was classified as a young stand. Other site characteristics were also unbalanced; of the 29 sites evaluated for predictive skill 17 were forested with 4 or less CRO, GRA, SHR, WET, and WSA sites.

While regression trees are inherently robust additional observed and simulated fluxes in rapidly growing young forested stands and undersampled biomes are desirable to better characterize model performance.

Several aspects of the simulation protocol influenced the interpretation of our performance analysis. Firstly, this analysis focused solely on non-gap-filled data to allow the model-data intercomparison to inform model devolvement. However, the low turbulence (friction velocity) filtering removes many more data at night than during the day, so our analysis may be skewed towards daytime conditions. Secondly, each model that used remotely sensed inputs (such as leaf area index) repeated an average seasonal cycle calculated from site-specific time series based on all pixels within 1 km of the tower site. This likely deflated relevant variable importance scores (Figure 5) and precluded a full comparison of diagnostic vs. prognostic LAI. While only 6 models used such inputs, including the best performing generalist model SiBCASA, removing the inherent bias in using an invariant seasonal cycle over multiple years may improve model performance. Incorporating disturbance information to recreate historical land use and disturbance, especially for recent site entries, could also improve model performance. Lastly, the models used here, except LoTEC and the mean model ensemble, were spun up to steady state initial conditions. However, few, if any, of the sites are actually at steady state, resulting in an inherent bias between simulated and observed NEP. Relaxing the steady state assumption (Carvalhais et al., 2008) or initializing using observed wood biomass and the quasi-steady state assumption (Schaefer et al., 2008) could improve model performance.

Despite these constraints the distribution of predictive skill and overall skill metrics indicated that the performance of comprehensive generalist models can equal that of specialist site-specific models (see also Grant et al., 2005). This has important implications for model choice and recasts model utility as a tradeoff between intensive biome and site-specific parameterizations vs. generic parameter sets linked to gridded datasets or remotely sensed data. Lastly the use of data assimilation techniques and mean model ensembles greatly increased model predictive skill.

Conclusion

We used observed CO_2 exchange from 44 eddy covariance towers in North America with simulations from 22 terrestrial biosphere models to validate model performance during a large scale drought event. Overall forested sites were better predicted than non-forested sites. Weaknesses in model performance concerned model parameter sets, growing season length, especially in biomes with a significant deciduous component, and abnormally dry or wet conditions. Undersampled biomes (grasslands, shrublands, wetlands, woody savannah, and tundra) also showed a large divergence between observations and simulations. The best model performance occurred in temperate evergreen forests across all climatic seasons and during winter and summer across all biomes. Models' ability to match observed monthly integrals of net ecosystem productivity across levels of dryness and seasonality was generally poor (< 0.50 predictive skill). In contrast, performance was higher when evaluating month-to-month trajectories. This indicated that the temporal evolution of NEP is better modeled than

responses to finer scaled changes in driving variables, albeit at the cost of under and overpredictions that cancel out over longer-term simulations. Models with relatively high performance metrics included generalist models applied over a wide range of sites and biomes as well as highly specialized (used on croplands or a single site only), optimized (parameters tuned using data assimilation), or ensembles (linear combinations of several models).

Acknowledgements

CRS, CAW, and KS were supported by the U.S. National Science Foundation grant ATM-0910766. We would like to thank the North American Carbon Program Site-Level Interim Synthesis team, the Modeling and Synthesis Thematic Data Center, and the Oak Ridge National Laboratory Distributed Active Archive Center for collecting, organizing, and distributing the model output and flux observations required for this analysis. This study was in part supported by the U.S. National Aeronautics and Space Administration (NASA) grant NNX06AE65G, the U.S. National Oceanic and Atmospheric Administration (NOAA) grant NA07OAR4310115, and the U.S. National Science Foundation (NSF) grant OPP-0352957 to the University of Colorado at Boulder.

References

Amthor, JS et al. (2001) Boreal forest CO2 exchange and evapotranspiration predicted by nine ecosystem process models: Intermodel comparisons and relationships to field measurements. J. Geophys. Res., 106(D24), 33,623-33,648.

Angert A, Biraud S, Bonfils C et al. (2005) Drier summers cancel out the CO_2 uptake enhancement induced by warmer springs. Proceedings of the National Academy of Sciences of the United States of America, 102, 10823-10827.

Arain MA, Yaun F, Black TA (2006) Soil-plant nitrogen cycling modulated carbon exchanges in a western temperate conifer forest in Canada. Agricultural and Forest Meteorology, 140, 171-192.

Arnone, JA, Verburg PSJ, Johnson DW et al. (2008) Prolonged suppression of ecosystem carbon dioxide uptake after an anomalously warm year. Nature 455:383-386.

Baldocchi D, Falge E, Gu LH 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.

Baldocchi, D (2008) Breathing of the terrestrial biosphere: lessons learned from a global network of carbon dioxide flux measurement systems. Australian Journal of Botany, 56, 1-26.

Baker IT, Prihodko L, Denning AS, Goulden M, Miller S, da Rocha HR (2008) Seasonal drought stress in the Amazon: Reconciling models and observations. J. Geophys. Res., 113, G00B01, doi:10.1029/2007JG000644.

Barr AG, Black TA, Hogg EH et al. 2004. Inter-annual variability in the leaf area index of a boreal aspen-hazelnut forest in relation to net ecosystem production. Agricultural Forest Meteorology, 126: 237-255.

Barr AG, Black TA, Hogg EH et al. (2007) Climatic controls on the carbon and water balances of a boreal aspen forest, 1994–2003. Global Change Biology, 13, 561-576.

Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Wadsworth, Belmont, CA, 358 pp.

Carvalhais N et al. (2008) Implications of the carbon cycle steady state assumption for biogeochemical modeling performance and inverse parameter retrieval. Global Biogeochem. Cycles, 22, GB2007, doi:10.1029/2007GB003033.

Causarano HJ, Shaw JN, Franzluebbers AJ, Reeves DW, Raper RL, Balkcom KS, Norfleet ML, Izaurralde RC (2007) Simulating field-scale soil organic carbon dynamics using EPIC. Soil Science Society of America Journal, 71, 1174-1185.

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

Cook ER, Woodhouse CA, Eakin CM, Meko DM, Stahle DW (2004) Long-term aridity changes in the western United States. Science, 306, 1015-1018.

Cook ER, Seager R, Cane MA, Stahle DW (2007) North American droughts: reconstructions, causes and consequences. Earth-Sci. Rev., 81, 93-134.

Dai A, Trenberth KE, Qian T (2004) A global data set of Palmer Drought Severity Index for 1870-2002: Relationship with soil moisture and effects of surface warming. J. Hydrometeorol., 5, 1117-1130.

Delpierre N, Soudani K, Francois C et al. (2009) Exceptional carbon uptake in European forests during the warm spring of 2007: a data–model analysis. Glob. Change Biol., 15, 1455-1474.

Friedlingstein P, Cox PM, Betts R et al. (2006) Climate-carbon cycle feedback analysis, results from the C^4MIP model intercomparison. Journal of Climate, 19, 3337-3353.

Friend AD, Arneth A, Kiang NY et al. (2007) FLUXNET and modelling the global carbon cycle. Glob. Change Biol., 13, 610-633.

Granier A, Reichstein M, Breda N et al. (2007) Evidence for soil water control on carbon and water dynamics in European forests during the extremely dry year: 2003. Agricultural and Forest Meteorology, 143, 123–145.

Grant RF et al. (2005) Intercomparison of techniques to model high temperature effects on CO_2 and energy exchange in temperate and boreal coniferous forests. Ecol. Model., 188, 217–252.

Hanson PJ, Amthor JS, Wullschleger SD et al. (2004) Oak forest carbon and water simulations: model intercomparisons and evaluations against independent data. Ecological Monographs, 74, 443-489.

Hargrove WW, Hoffman FM, Law BE (2003) New Analysis Reveals Representativeness of AmeriFlux Network. Earth Observing System Transactions, American Geophysical Union 84(48):529.

Higuchi K, Shashkov A, Chan D, Saigusa N, Murayama S, Yamamoto S, Kondo H, Chen JM, Liu J, Chen B (2005) Simulations of seasonal and inter-annual variability of net CO_2 flux at Takayama with BEPS Ecosystem Model. Agricultural and Forest Meteorology, 134, 143-150.

Houghton JT, Ding Y, Griggs DJ, Noguer M, van der Linden PJ, Xia D, Maskell K, Johnson CA (eds) (2001) Climate Change 2001: The Scientific Basis: Contributions of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, New York, 881 pp.

Huntington TG (2006) Evidence for intensification of the global water cycle: Review and synthesis. J. Hydrol., 319, 83-95.

Hwang Y, Carbone GJ (2009) Ensemble forecasts of drought indices using a conditional residual resampling technique. Journal of Applied Meteorology and Climatology, 48, 1289-1301.

Ichii K, Suzuki T, Kato T et al (2009) Multi-model analysis of terrestrial carbon cycles in Japan: reducing uncertainties in model outputs among different terrestrial biosphere models using flux observations, Biogeosciences Discuss., 6, 8455-8502.

Ito A (2008) The regional carbon budget of East Asia simulated with a terrestrial ecosystem model and validated using AsiaFlux data. Agr. Forest Meteorol., 148, 738-747.

Jung M, Le Maire G, Zaehle S et al. (2007a) Assessing the ability of three land ecosystem models to simulate gross carbon uptake of forests from boreal to Mediterranean climate in Europe. Biogeosciences, 4, 647-656.

Jung M, Vetter M, Herold M et al. (2007b) Uncertainties of modeling gross primary productivity over Europe: A systematic study on the effects of using different drivers and terrestrial biosphere models. Global Biogeochem. Cycles, 21, GB4021, doi:10.1029/2006GB002915.

Keeling CD, Chin JF, Whorf TP (1996) Increased activity of northern vegetation inferred from atmospheric CO_2 measurements. Nature, 382, 146-149.

King AW, Post WM, Wullschleger SD, Ricciuto DM (2009) Sensitivity of modeled net ecosystem exchange to uncertainty in initial ecosystem carbon stocks (in preparation).

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

Kucharik CJ, Twine TE (2007) Residue, respiration, and residuals: Evaluation of a dynamic agroecosystem model using eddy flux measurements and biometric data. Agricultural and Forest Meteorology, 146, 134-158.

Larcher W (1995) Physiological Plant Ecology. Springer Verlag, Berlin, 506 pp.

Li et al. (2009) Modeling impacts of alternative farming management practices on greenhouse gas emissions from a winter wheat-maize rotation system in China. Agriculture, Ecosystems and Environment, 135, 24-33.

Liu S, Bliss N, Sundquist E, Huntington TG (2003) Modeling carbon dynamics in vegetation and soil under the impact of soil erosion and deposition. Global Biogeochemical Cycles, 17, doi:10.1029/2002GB002010.

Lokupitiya E, Denning S, Paustian K, Baker I, Schaefer K, Verma S, Meyers T. Bernacchi CJ, Suyker A, Fischer M (2009) Incorporation of crop phenology in Simple Biosphere Model (SiBcrop) to improve land-atmosphere carbon exchanges from croplands. Biogeosciences, 6, 969-986.

McKee TB, Doeskin NJ, Kleist J (1993) The relationship of drought frequency and duration to time scales. Proc. 8th Conf. on Applied Climatology, January 17-22, 1993, American Meteorological Society, Boston, Massachusetts, 179-184.

Medlyn BE, Robinson AP, Clement R, McMurtrie RE (2005) On the validation of models of forest CO_2 exchange using eddy covariance data: some perils and pitfalls. Tree Physiology, 25, 839-857.

Medvigy D, Wofsy SC, Munger JW, Hollinger DY, Moorcroft PR (2009) Mechanistic scaling of ecosystem function and dynamics in space and time: Ecosystem Demography model version 2. J. Geophys. Res., 114, G01002, doi:10.1029/2008JG000812.

Meehl GA, Tebaldi C (2004) More intense, more frequent, and longer lasting Heat waves in the 21st Century. Science, 305, 994-997.

Mishra VR, Desai AK (2006) Drought forecasting using feed-forward recursive neural network. Ecol. Modell., 198, 127-138.

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

Morales P, Sykes MT, Prentice IC et al (2005) Comparing and evaluating process-based ecosystem model predictions of carbon and water fluxes in major European forest biomes. Global Change Biol., 11, 2211-2233.

Papale D, Reichstein M, Aubinet M et al. (2006) Towards a standardized processing of net ecosystem exchange measured with eddy covariance technique: algorithms and uncertainty estimation. Biogeosciences, 3, 571-583.

Piao SL, Friedlingstein P, Peylin P, Reichstein M, Luyssaert S, Margolis H, Fang JY, Barr AG, Chen A, Grelle A, Hollinger DY, Laurila T, Lindroth A, Richardson A, Vesala T (2008) Net carbon dioxide losses of northern ecosystems in response to autumn warming. Nature 451, 49-53.

Purves DW, Pacala S (2008) Predictive models of forest dynamics. Science, 320, 1452-1453.

Richardson AD and Hollinger DY. 2007. A method to estimate the additional uncertainty in gap-filled NEE resulting from long gaps in the CO2 flux record. Agricultural and Forest Meteorology, 147, 199-208.

Riley WJ, Still CJ, Torn MS, Berry JA (2002) A mechanistic model of H218O and C18OO fluxes between ecosystems and the atmosphere: Model description and sensitivity analyses, Global Biogeochem. Cycles, 16, 1095, doi:10.1029/2002GB001878.

Schaefer K, Zhang T, Tans P, Stöckli R (2007) Temperature anomaly reemergence in seasonally frozen soils. J. Geophys. Res., 112, D20102, doi:10.1029/2007JD008630.

Schaefer K, Collatz GJ, Tans P et al. (2008) Combined Simple Biosphere/Carnegie-Ames-Stanford Approach terrestrial carbon cycle model, J. Geophys. Res., 113, G03034, doi:10.1029/2007JG000603.

Schaefer K, Zhang T, Slater AG, Lu L, Etringer A, Baker I (2009), Improving simulated soil temperatures and soil freeze/thaw at high-latitude regions in the Simple Biosphere/Carnegie-Ames-Stanford Approach model. J. Geophys. Res., 114, F02021, doi:10.1029/2008JF001125.

Schwalm CR, Williams CA, Schaefer KS et al. (2009) Assimilation exceeds respiration sensitivity to drought: A FLUXNET synthesis. Global Change Biology, doi: 10.1111/j.1365-2486.2009.01991.x

Seager R (2007a) The turn of the century drought across North America: global context, dynamics and past analogues. Journal of Climate, 20, 5527-5552.

Seager R, Ting MF, Held IM (2007b) Model projections of an imminent transition to a more arid climate in Southwestern North America, Science, 316, 1181-1184.

Seager R, Tzanova A, Nakamura J (2009) Drought in the Southeastern United States: Causes, variability over the last millennium and the potential for future hydroclimate change, Journal of Climate, 22, 5021-5045.

Sheffield J, Wood EF (2008) Projected changes in drought occurrence under future global warming from multi-model, multi-scenario, IPCC AR4 simulations. Climate Dynamics, 13, 79-105.

Siqueira MB, Katul GG, Sampson DA, Stoy PC, Juang J-Y, McCarthy HR, Oren R (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.

Sitch S, Smith B, Prentice IC, Arneth A, Bondeau A, Cramer W, Kaplan JO, Levis S, Lucht W, Sykes MT, Thonicke K, Venevsky S (2003) Evaluation of ecosystem dynamics, plant geography and terrestrial carbon cycling in the LPJ dynamic global vegetation model. Global Change Biology, 9, 161-185.

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

Soja AJ, Tchebakova NM, French NHF, Flannigan MD, Shugart HH, Stocks BJ, Sukhinin AI, Varfenova EI, Chapin FS, Stackhouse PW (2007). Climate induced boreal forest change: Predictions versus current observations. Global and Planetary Change, 56, 274-296.

St-Hilaire F, Wu J, Roulet NT, Frolking S, Lafleur PM, Humphreys ER, Arora V (2008) McGill Wetland Model: evaluation of a peatland carbon simulator developed for global assessments. Biogeosciences Discussions, 5, 1689-1725.

Stoy P, Katul G, Siqueira M, Juang J, McCarthy H, Kim H, Oishi A, Oren R (2005) Variability in net ecosystem exchange from hourly to inter-annual time scales at adjacent pine and hardwood forests: a wavelet analysis. Tree Physiology, 25, 887-902.

Stoy P, Richardson A, Baldocchi D, Katul G, Stanovick J, Mahecha M, Reichstein M, Detto M, Law B, Wohlfahrt G, Arriga N, Campos J, McCaughey J, Montagnani L, Paw U, KT, Sevanto S, Williams M (2009) Biosphere-atmosphere exchange of CO_2 in relation to climate: a cross-biome analysis across multiple time scales. Biogeosciences, 6, 2297-2312.

Taylor KE (2001) Summarizing multiple aspects of model performance in a single diagram. J. Geophys. Res., 106, 7183-7192.

Thomas CK, Law BE, Irvine J, Martin JG, Pettijohn JC, Davis KJ (2009) Seasonal hydrology explains interannual and seasonal variation in carbon and water exchange in a semi-arid mature ponderosa pine forest in Central Oregon. JGR Biogeosciences. In press.

Thornton, PE, Running SW, White MA (1997) Generating surfaces of daily meteorological variables over large regions of complex terrain. J. of Hydrology, 3-4, 214-251

Thornton PE, Running SW, Hunt ER (2005) Biome-BGC: Terrestrial Ecosystem Process Model, Version 4.1.1. Model product. Available on-line [http://daac.ornl.gov] from Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, U.S.A. doi:10.3334/ORNLDAAC/805.

Tian, HQ, Chen G, Liu M, Zhang C, Sun G, Lu C, Xu X, Ren W, Pan P, Chappelka A (2009) Model Estimates of Ecosystem Net Primary Productivity, Evapotranspiration, and Water Use Efficiency in the Southern United States during 1895-2007. Forest Ecology and Management (in press).

Trenberth KE, Jones PD, Ambenje P et al. (2007) Observations: Surface and atmospheric climate change. In: *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* (eds Solomon SD et al.). Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Waring RH, Running SW (2007) Forest ecosystems - analysis at multiple scales (third edition). Academic Press, San Diego, 440 pp.

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

Williams CA, Hanan NP, Scholes RJ, Kutsch WL (2009) Complexity in water and carbon dioxide fluxes following rain pulses in an African savanna. Oecologia, 161, 469-480.

Williamson TB, Price DT, Beverly JL, Bothwell PM, Frenkel B, Park J, Patriquin MN (2008) Assessing potential biophysical and socioeconomic impacts of climate change on forest-based communities: a methodological case study. Nat. Resour. Can., Can. For. Serv., North. For. Cent., Edmonton, AB. Inf. Rep. NOR-X-415E.

Zhan XW, Xue YK, Collatz GJ (2003) An analytical approach for estimating CO_2 and heat fluxes over the Amazonian region. Ecol. Model. 162, 97-117.

Zhou XL, Peng CH, Dang QL, Sun JF, Wu HB, Hua D (2008) Simulating carbon exchange in Canadian Boreal forests I: model structure, validation, and sensitivity analysis. Ecological Modelling, 219, 287-299.

Tables & Figures

Table 1. Models evaluated with months and sites sampled. Difference in sample sizes results from the use of standard errors of observed fluxes and site-specific drought metrics for predictive skill. Taylor diagrams included all available flux data.

		Predictive skill		Taylor diagrams		
Model	Timestep	Months	Sites	Months	Sites	Source
AgroIBIS[*]	half-hourly	165	4	192	5	Kucharik & Twine (2007)
BEPS	daily	860	10	945	10	Higuchi et al. (2004)
Biome-BGC	daily	1536	26	2001	36	Thornton et al. (2005)
Can-IBIS	half-hourly	1587	23	1978	27	Williamson et al. (2008)
CN-CLASS	half-hourly	1696	25	2082	31	Arain et al. (2006)
DLEM[%]	daily	1831	28	2246	33	Tian et al. (2009)
DNDC[*]	daily	122	4	192	5	Li et al. (2009)
Ecosys	hourly	1693	26	2450	39	Grant et al. (2005)
ED2	half-hourly	1430	22	1684	25	Medvigy et al. (2009)
EDCM	monthly	582	10	658	10	Liu et al. (2003)
EPIC[*,%]	daily	47	1	48	1	Causarano et al. (2007)
ISOLSM[%]	half-hourly	824	9	909	9	Riley et al. (2002)
LoTEC[$]	hourly	759	10	825	10	King et al. (2009)
LPJ	daily	1711	25	2126	29	Sitch et al. (2003)
MEAN[#]	native	1940	30	2776	44	this study
ORCHIDEE	half-hourly	1873	30	2332	35	Krinner et al. (2005)
SiB3[%]	half-hourly	1825	27	2258	31	Baker et al. (2008)
SiBCASA[%]	half-hourly	1952	30	2402	35	Schaefer et al. (2009)
SiBcrop[*]	half-hourly	165	4	192	5	Lokupitiya et al. (2009)
SSiB2[%]	half-hourly	1952	30	2800	44	Zhan et al. (2003)
TECO	hourly	1952	30	2414	35	Weng & Luo (2008)
TRIPLEX-Flux	daily	246	7	291	7	Zhou et al. (2008)

[*] Used only on croplands

[%] Used remotely sensed inputs of leaf area index, greenness, or fraction of absorbed fraction of photosynthetically active radiation, i.e., phenology was calculated in diagnostic mode using an average seasonal cycle by site for the full simulation period.

[$] Tuned using data assimilation

[#] Mean value across all modeled values

Table 2. Name, measurement period, and biome for all sites. Italicized site entries only used in Taylor diagrams. CA sites located in Canada; US sites in the continental United States except US-Atq and US-Brw (Alaska).

Site	Full Name	Period	Biome[#]
CA-Ca1	Campbell River – Mature Douglas-fir	1998-2006	ENFT
CA-Ca2	*Campbell River – Douglas-fir clearcut*	*2001-2006*	*ENFT*
CA-Ca3	*Campbell River – Douglas-fir juvenile*	*2002-2006*	*ENFT*
CA-Gro	Groundhog River Station	2004-2006	MF
CA-Let	Lethbridge Grassland	1997-2006	GRA
CA-Mer	Eastern Peatland – Mer Bleue	1999-2006	WET
CA-Oas	BERMS – Old Aspen	1997-2006	DBF
CA-Obs	BERMS – Old Black Spruce	2000-2006	ENFB
CA-Ojp	BERMS – Old Jack Pine	2000-2006	ENFB
CA-Qfo	Quebec – Mature Black Spruce	2004-2006	ENFB
CA-SJ1	*BERMS – Jack Pine, 1994 harvest*	*2002-2005*	*ENFB*
CA-SJ2	*BERMS – Jack Pine, 2002 harvest*	*2003-2006*	*ENFB*
CA-SJ3	*BERMS – Jack Pine, 1975 harvest*	*2004-2005*	*ENFB*
CA-TP3	*Turkey Point – Middle-aged*[$]	*2003-2007*	*ENFT*
CA-TP4	Turkey Point – Mature	2002-2007	ENFT
CA-WP1	Western Peatland – LaBiche River	2003-2007	WET
US-ARM	ARM – Southern Great Plains	2000-2006	CRO
US-Atq	*Atqasuk*[$]	*1999-2006*	*TUN*
US-Brw	*Barrow*[$]	*1998-2006*	*TUN*
US-Dk2	*Duke Forest – Hardwood*[$]	*2003-2005*	*DBF*
US-Dk3	Duke Forest – Loblolly Pine	1998-2005	ENFT
US-Ha1	Harvard Forest – EMS Tower	1991-2006	DBF
US-Ho1	Howland Forest – Main Tower	1996-2004	ENFT
US-IB1	*Fermi Lab – Maize/soybean rotation*	*2005-2007*	*CRO*
US-IB2	Fermi Lab – Prairie	2004-2007	GRA
US-Los	Lost Creek	2000-2006	WET
US-Me2	Metolius – Intermediate-aged Ponderosa Pine	2002-2007	ENFT
US-Me3	*Metolius – Ponderosa Pine, young #2*	*2004-2005*	*ENFT*
US-Me4	*Metolius – Ponderosa Pine, old-growth*[$]	*1996-2000*	*ENFT*
US-Me5	*Metolius – Ponderosa Pine, Young #1*	*1999-2002*	*ENFT*
US-MMS	Morgan Monroe State Forest	1999-2006	DBF
US-MOz	Missouri Ozark	2004-2007	DBF
US-Ne1	Mead – Irrigated maize	2001-2006	CRO
US-Ne2	Mead – Irrigated maize/soybean	2001-2006	CRO
US-Ne3	Mead – Rainfed maize/soybean	2001-2006	CRO
US-NR1	*Niwot Ridge*	*1998-2007*	*ENFT*
US-PFa	Park Falls / WLEF	1997-2005	MF

US-Shd	*Shidler*	*1997-2001*	*GRA*
US-SO2	Sky Oaks – Old	1998-2006	SHR
US-Syv	Sylvania Wilderness Area	2001-2006	MF
US-Ton	Tonzi Ranch	2001-2007	WSA
US-UMB	University of Michigan Biological Station (UMBS)	1998-2006	DBF
US-Var	Vaira Ranch	2001-2007	GRA
US-WCr	Willow Creek	1998-2006	DBF

[#] Biome codes: CRO = cropland, GRA = grassland, ENFB = evergreen needleleaf forest – boreal climatic zone, ENFT = evergreen needleleaf forest – temperate climatic zone, DBF = deciduous broadleaf forest, MF = mixed (deciduous/evergreen) forest, WSA = woody savanna, SHR = shrubland, TUN = tundra, WET = wetland.

[$] Sites used alternative post-processing protocol based on the La Thuile and Asilomar FLUXNET Synthesis dataset (http://www.fluxdata.org/; Moffat et al., 2007; Papale et al., 2006).

Table 3. Model-specific and site history predictants, in addition to biome, climatic season, and drought level, used to classify the presence/absence of predictive skill.

Predictant*	Value type	Comment
Model temporal resolution	Factor	Value: daily, hourly, half-hourly
Prognostic vs. diagnostic LAI	Boolean	How was the canopy, i.e., LAI (leaf area index), calculated? Does model require LAI as an input? Note: A subset of models used remotely sensed LAI to calculate canopy biomass or for phenology (diagnostic mode). This represents partial assimilation of site data into model runs insomuch as drought is reflected in the remotely sensed products used.
Number of vegetation pools	Numeric	Number of pools, both dynamic and static
Number of soil pools	Numeric	Number of pools, both dynamic and static
Number of soil layers	Numeric	Number of layers
Algorithm: Autotrophic respiration	Factor	Value: assumed fraction of instantaneous GPP, explicitly calculated, How is autotrophic respiration calculated
Algorithm: Canopy leaf biomass	Factor	Value: explicitly calculated, GPP/NPP fraction, LAI, nil How is canopy biomass calculated?
Algorithm: Ecosystem respiration	Factor	Value: autotrophic respiration + heterotrophic respiration, explicitly calculated, forced annual balance How is ecosystem respiration calculated?
Algorithm: Gross primary productivity	Factor	Value: enzyme kinetic model, light use efficiency model, nil, stomatal conductance model How is gross primary productivity calculated?
Algorithm: Heterotrophic respiration	Factor	Value: explicitly calculated, first or greater order model, zero-order model How is heterotrophic respiration calculated?
Algorithm: Net ecosystem productivity	Factor	Value: explicitly calculated, gross primary productivity - ecosystem respiration, net primary production - heterotrophic respiration How is net ecosystem productivity calculated?
Overall model complexity	Factor	Value: minimal, low, average, high, highest Total amount of first-order functional arguments for the following model-generated variables/outputs: Autotrophic respiration, canopy leaf biomass, ecosystem respiration, evapotranspiration, gross primary productivity, heterotrophic respiration, net ecosystem productivity, net primary production, soil moisture

		Note: Factor values correspond to quintiles.
Grazed Fertilized Fire Harvest Herbicide Insects and pathogens Irrigation Natural regeneration Pesticide Planted Residue management Thinning Underburn	Boolean	Has listed management activity or disturbance or event occurred on site?
Stand age class	Factor	Value: young, intermediate, nil, mature, multi-cohort, Values based on stand age in forested sites; stands without a clear dominant stratum are treated as multi-cohort; non-forest types have nil.

* Modeling architecture predictants were taken from the Metadata for Forward (Ecosystem) Model Intercomparison survey data collated by the NACP Site Synthesis (http://daac.ornl.gov/SURVEY8/survey_results.shtml). Of the 21 distinct modeling platforms used (excluding the mean model ensemble) 15 models contributed metadata. The predictive skill classification exercise excluded runs from: AgroIBIS, EDCM, LoTEC, SiBCASA, SSiB2, and TRIPLEX-Flux.

Table 4. Predictive skill and sample size by drought level and climatic season. Drought level was based on monthly values of 3-month SPI, Dry required a Standard Precipitation Index (SPI) value of < -0.8; Wet > +0.8. Otherwise Normal conditions existed. Sample size is the total number of paired data-model site-months.

	Predictive skill				Sample size			
	Dry	Normal	Wet	Overall	Dry	Normal	Wet	Overall
Winter	0.42	0.49	0.47	0.47	1479	3996	1096	6571
Spring	0.27	0.35	0.22	0.31	1534	3990	1209	6733
Summer	0.45	0.54	0.42	0.49	1359	3910	1442	6711
Fall	0.28	0.30	0.33	0.30	1491	3936	1305	6732
Overall	0.35	0.42	0.36	0.39	5863	15832	5052	26747

Table 5. Predictive skill by climatic season, drought level, and biome. Drought level was based on monthly values of 3-month SPI, Dry required a Standard Precipitation Index (SPI) value of < -0.8; Wet > +0.8. Otherwise Normal conditions existed.

	Climatic season				Drought level			
Biome[#]	Winter	Spring	Summer	Fall	Dry	Normal	Wet	Overall
CRO	0.50	0.43	0.52	0.06	0.45	0.34	0.22	0.37
DBF	0.57	0.15	0.52	0.21	0.23	0.39	0.36	0.36
ENFB	0.32	0.49	0.58	0.23	0.30	0.37	0.52	0.41
ENFT	0.49	0.50	0.57	0.79	0.52	0.63	0.51	0.59
GRA	0.45	0.34	0.32	0.06	0.23	0.37	0.13	0.30
MF	0.61	0.07	0.68	0.16	0.35	0.42	0.30	0.38
SHR	0.02	0.16	0	0	0.04	0.04	0.05	0.04
WET	0.40	0.12	0.42	0.11	0.23	0.29	0.20	0.26
WSA	0.07	0.49	0.27	0.05	0.14	0.25	0.18	0.22
Overall	0.47	0.31	0.49	0.30	0.35	0.42	0.36	0.39

[#] Biome codes: CRO = cropland, GRA = grassland, ENFB = evergreen needleleaf forest – boreal climatic zone, ENFT = evergreen needleleaf forest – temperate climatic zone, DBF = deciduous broadleaf forest, MF = mixed (deciduous/evergreen) forest, WSA = woody savanna, SHR = shrubland, WET = wetland.

Figure 1.

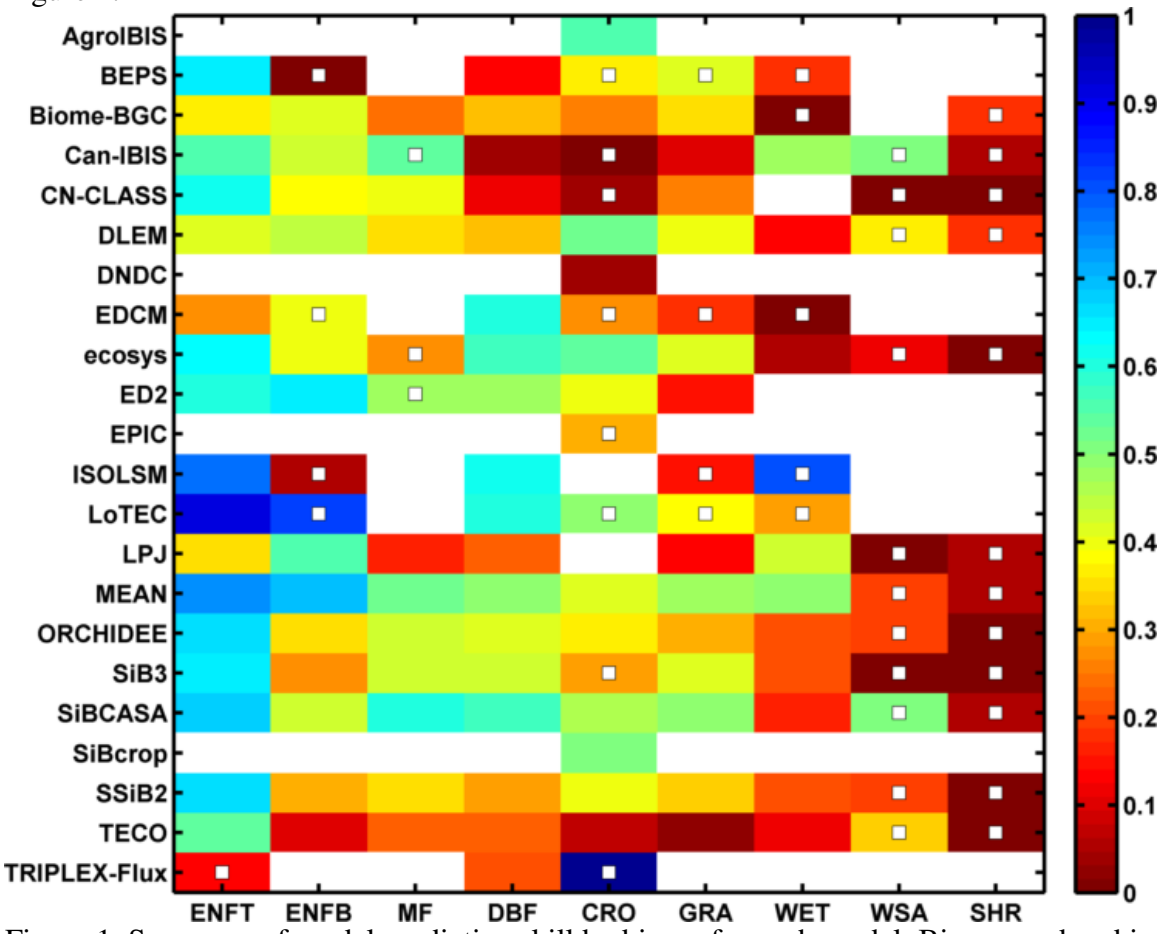

Figure 1. Summary of model predictive skill by biome for each model. Biomes ordered in descending order based on across-model average predictive skill. White blocks: no observations; gray squares: undersampled ($n < 100$ months).

Figure 2.

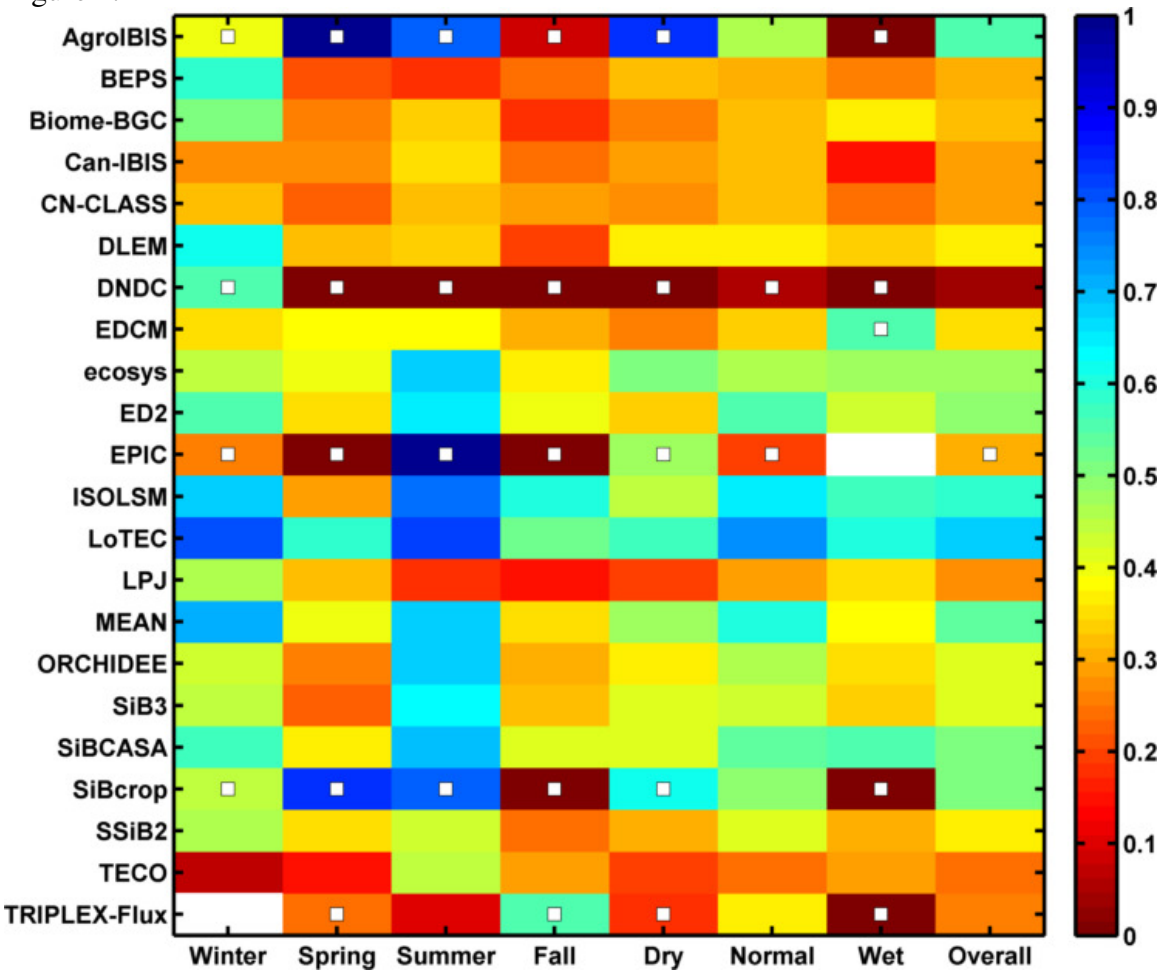

Figure 2. Summary of model predictive skill by climatic season, drought level, and overall. White blocks: no observations; gray squares: undersampled ($n < 100$ months).

Figure 3.

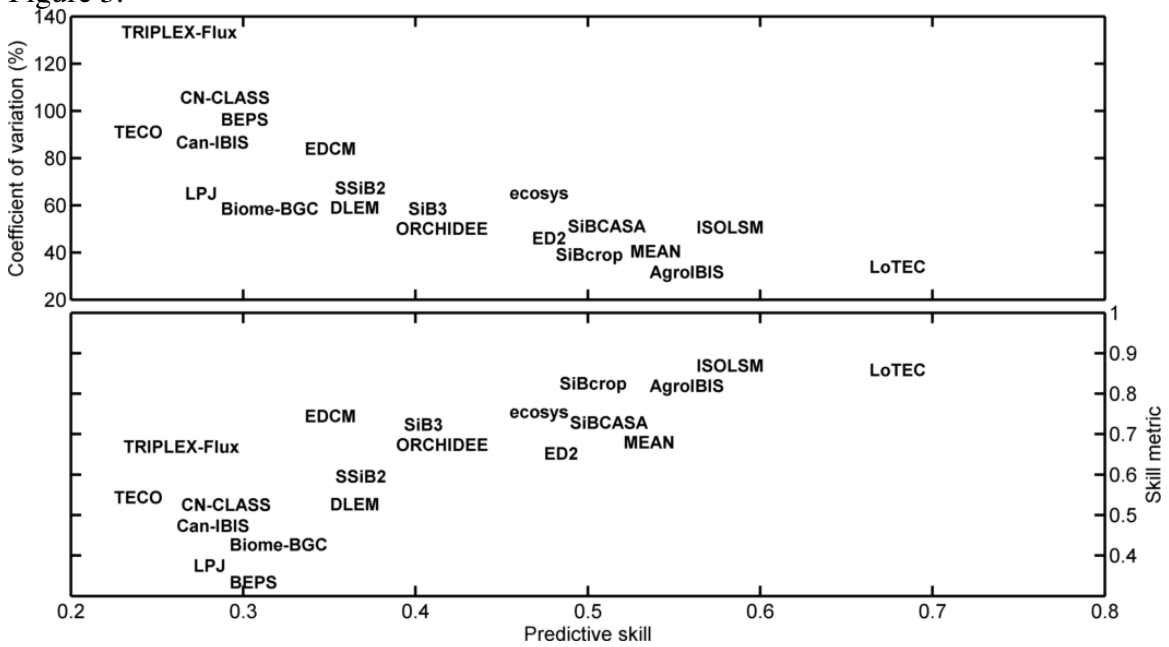

Figure 3. Model performance metrics. Upper panel: coefficient of variation for predictive skill (CV_π) vs. predictive skill (π). Lower panel: overall model skill (S) from Taylor diagrams vs. predictive skill. DNDC ($S = 0.34$, $\pi = 0.04$, $CV_\pi = 174\%$; outlier) and EPIC ($S = 0.95$, $\pi = 0.30$, CV_π = undefined; only one site simulated) not shown. Model names jittered to improve readability.

Figure 4.

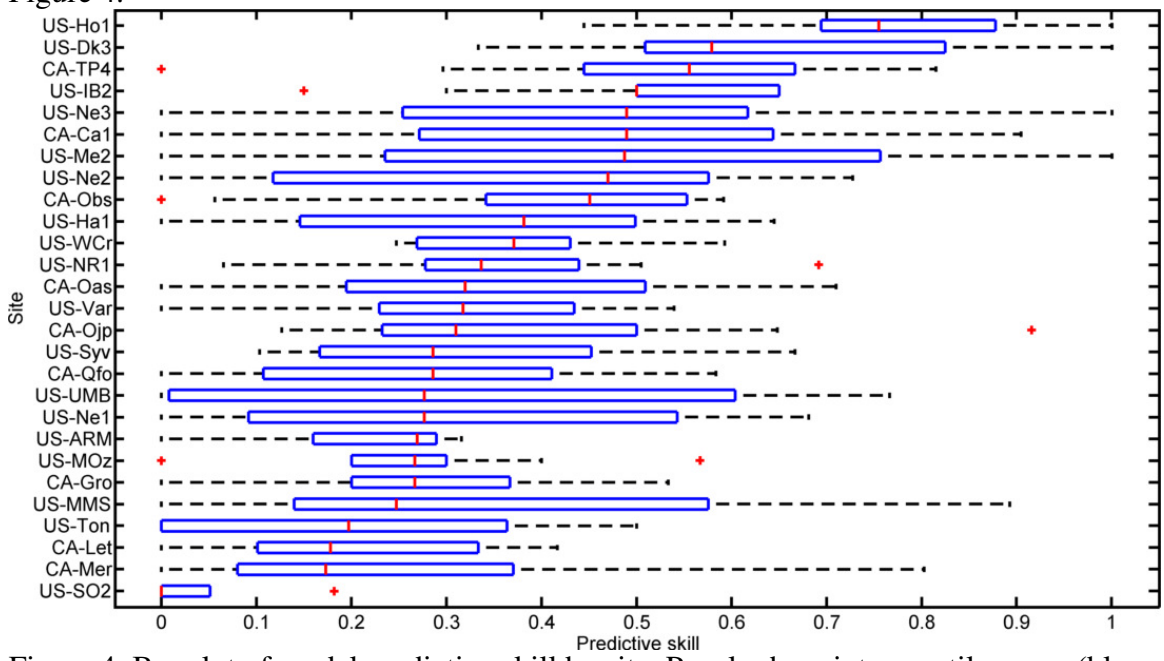

Figure 4. Boxplot of model predictive skill by site. Panels show interquartile range (blue box), median (solid red line), range (whiskers), and outliers (red cross; values more than 1.5 x interquartile range from the median). Only sites (n = 27) simulated with at least 10 unique models using steady state spinup shown. Sites sorted by median predictive skill.

Figure 5.

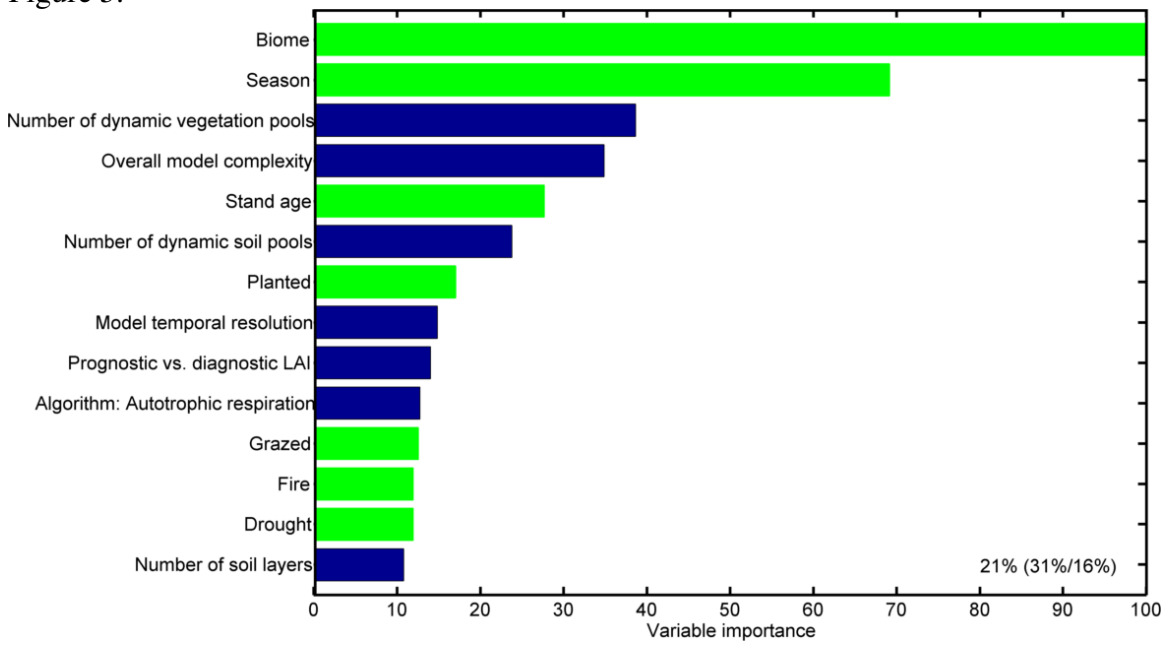

Figure 5. Variable importance scores for model-specific (blue) and site-specific (green) predictants. Scores were generated from a regression tree with the presence/absence of predictive skill by data group (n = 2258) as the response. *Planted*, *Grazed*, and *Fire* were boolean variables related to the occurrence of the event on site. *Stand age* was young, intermediate, or mature for forested stands; nil otherwise. For model-specific predictants *Algorithm: Autotrophic respiration* was either assumed fraction of instantaneous gross primary productivity, explicitly calculated, or nil (not calculated by the model). Numbers in lower right corner indicate overall misclassification rate and, parenthetically, within class misclassification rates for data groups with then without predictive skill. Only the 14 of 29 predictants with score > 10 shown; see Table 3 for complete listing of evaluated model structural and site attributes.

Figure 6.

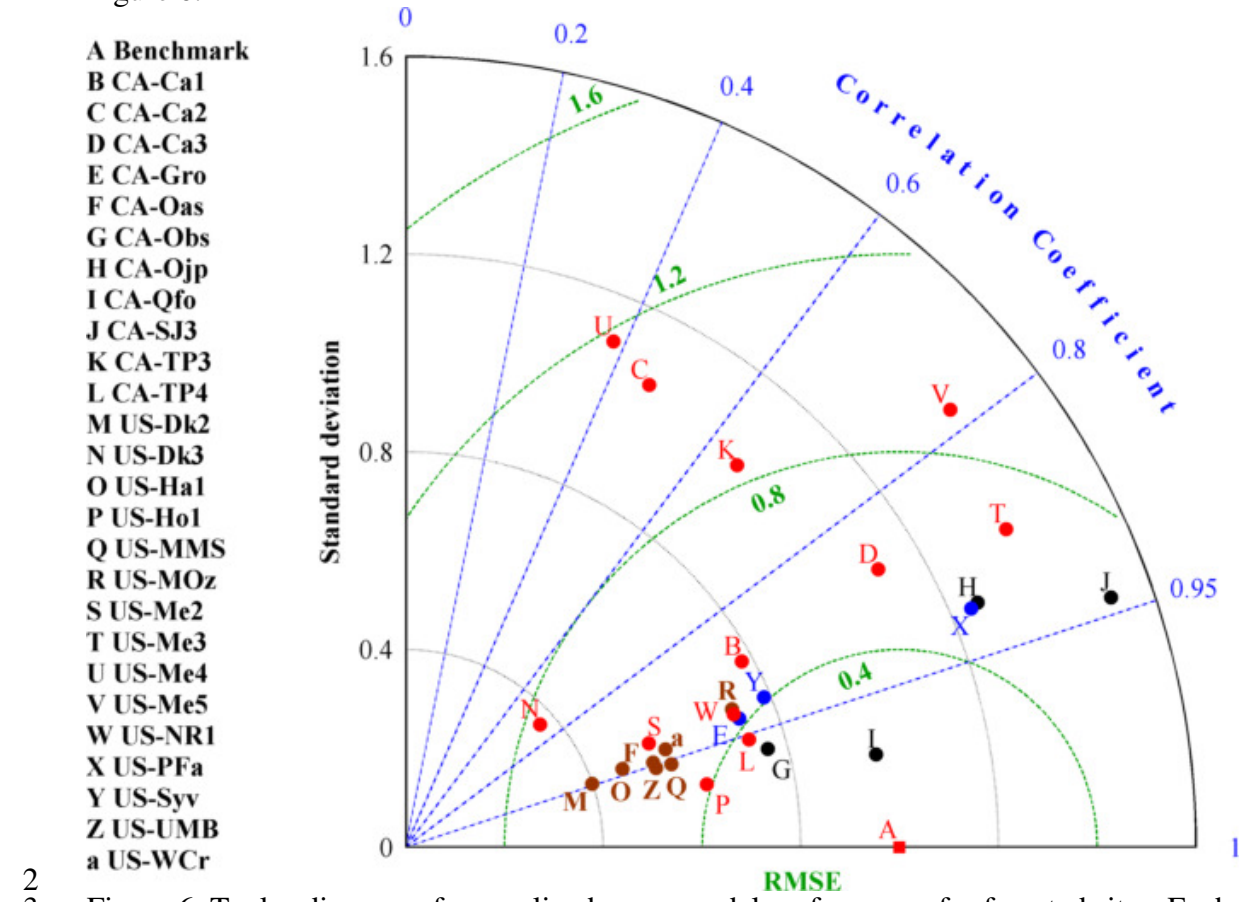

Figure 6. Taylor diagram of normalized mean model performance for forested sites. Each circle (n = 26 sites) is the site-specific mean model ensemble (MEAN). Benchmark (red square) corresponds to observed normalized monthly NEP; units of σ and RMSE are multiples of observed σ. Color coding of site letter and circles indicates biome: evergreen needleleaf forest – temperate climatic zone (red), deciduous broadleaf forest (brown), mixed (deciduous/evergreen) forest (blue), evergreen needleleaf forest – boreal climatic zone (black). Outlying sites (evergreen needleleaf forest – boreal climatic zone) not shown: CA-SJ1 (ρ = 0.81, σ = 3.9, RMSE = 3.1) and CA-SJ2 (ρ = –0.67, σ = 4.3, RMSE = 5.1).

Figure 7.

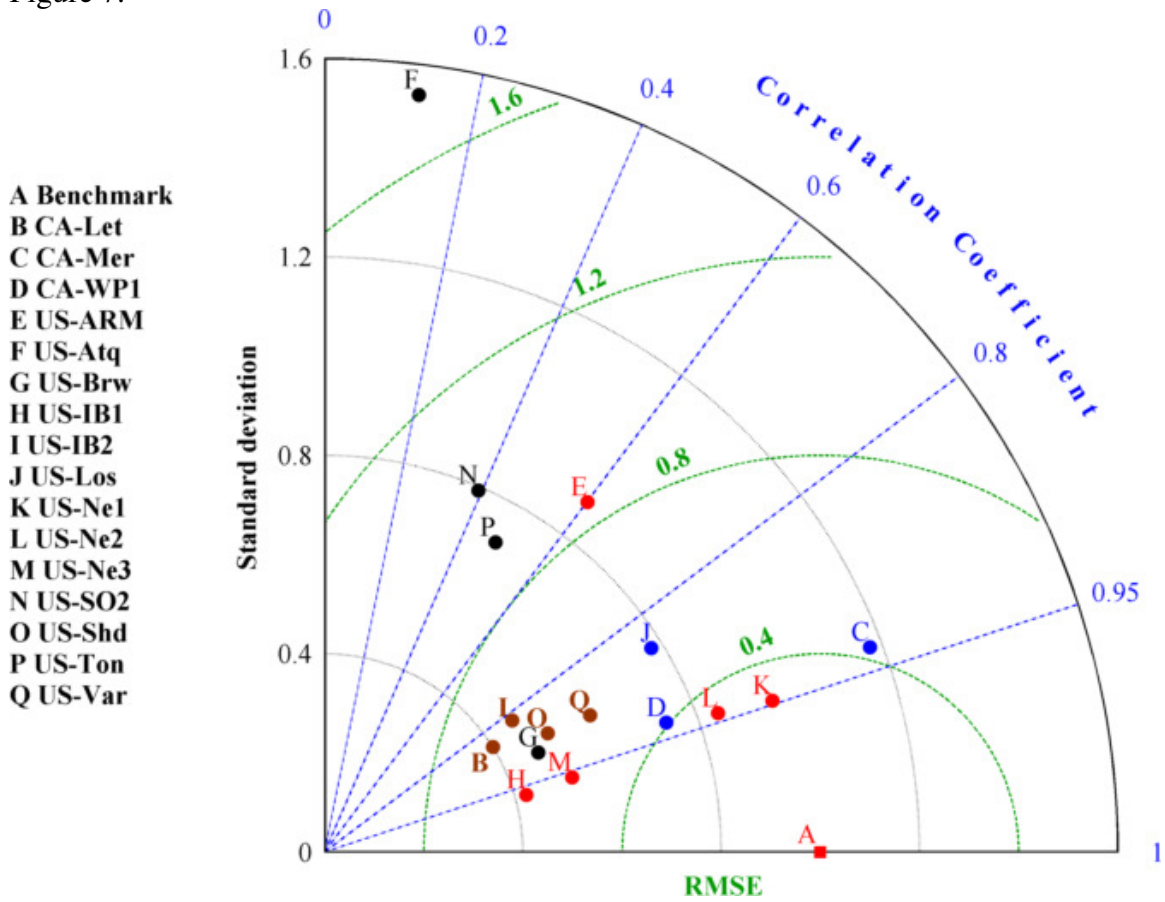

Figure 7. Taylor diagram of normalized mean model performance for non-forested sites. Each circle (n = 16 sites) is the site-specific mean model ensemble (MEAN). Benchmark (red square) corresponds to observed normalized monthly NEP; units of σ and RMSE are multiples of observed σ. Color coding of site letter and circles indicates biome: croplands (red), grasslands (brown), wetlands (blue), all other biomes (black).

Figure 8.

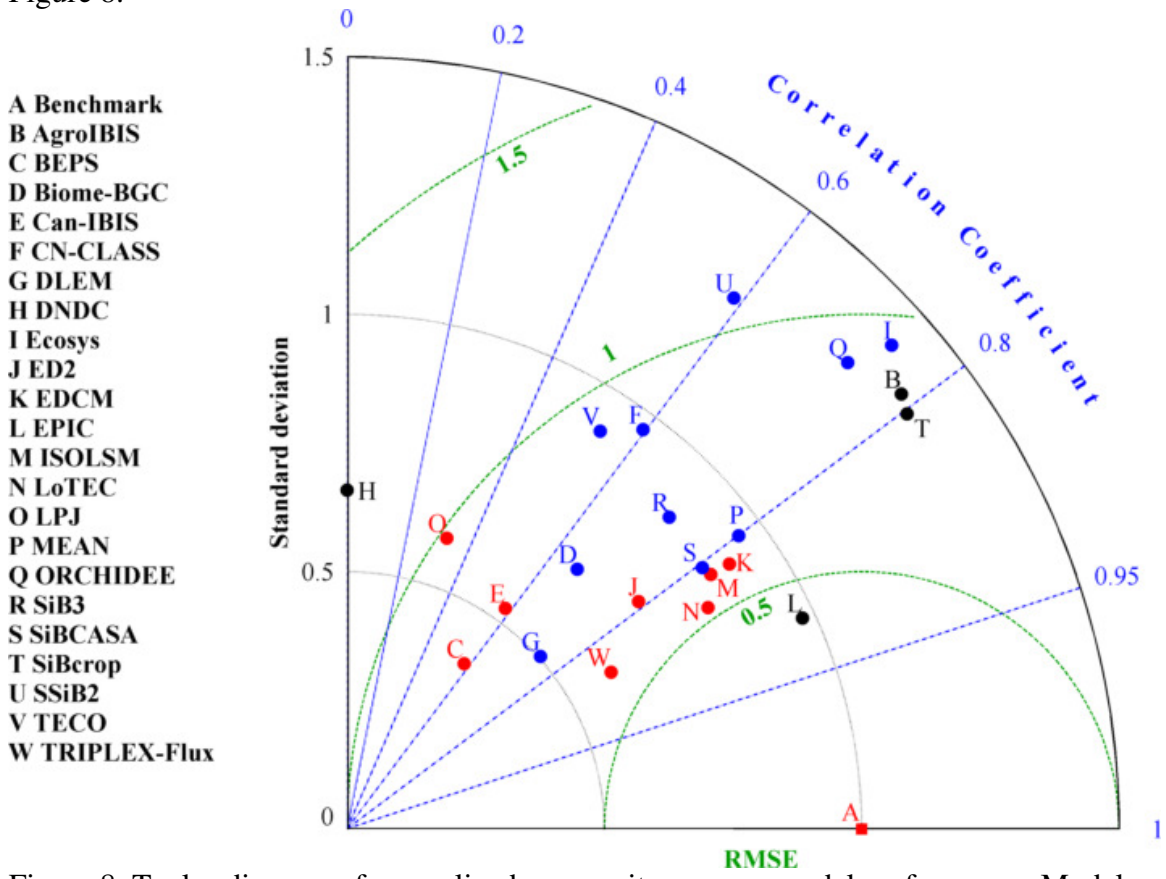

Figure 8. Taylor diagram of normalized across-site average model performance. Model σ and RMSE were normalized by observed σ. Each circle (n = 22 models) corresponds to the mean across all sites. Benchmark (red square) corresponds to observed normalized monthly NEP; units of σ and RMSE are multiples of observed σ. Color coding of model letter and circles indicates generality of model performance: specialist models used only in croplands ($n \leq 5$ sites; black), generalist models used across range of biomes and sites ($n \geq 30$ sites, blue), all other models (red). The correlation for DNDC ($\rho = -0.13$) is displayed as zero for readability.

[End of document]