A Model-Data Comparison of Gross Primary Productivity: Results from the North American Carbon Program Site Synthesis

Kevin Schaefer¹, Christopher Schwalm², Chris Williams³, M. Altaf Arain⁴, Alan Barr⁵, Jing M. Chen⁶, Kenneth J. Davis⁷, Dimitre Dimitrov⁸, Timothy W. Hilton⁹, David Y. Hollinger¹⁰, Elyn Humphreys¹¹, Benjamin Poulter¹², Brett M. Raczka⁷, Andrew D. Richardson¹³, Alok Sahoo¹⁴, Peter Thornton¹⁵, Rodrigo Vargas¹⁶, Hans Verbeeck¹⁷, Ryan Anderson¹⁸, Ian Baker¹⁹, T. Andrew Black²⁰, Paul Bolstad²¹, Jiquan Chen²², Peter Curtis²³, Ankur R. Desai²¹, Michael Dietze²⁴, Danilo Dragoni²⁵, Christopher Gough²⁶, Robert F. Grant²⁷, Lianhong Gu¹⁵, Atul Jain²⁸, Chris Kucharik²⁹, Beverly Law³⁰, Shuguang Liu³¹, Erandathie Lokipitiya¹⁹, Hank A. Margolis³², Roser Matamala³³, J. Harry McCaughey³⁴, Russ Monson³⁵, J. William Munger³⁶, Walter Oechel³⁷, Changhui Peng³⁸, David T. Price⁸, Dan Ricciuto¹⁵, William J. Riley³⁹, Nigel Roulet⁴⁰, Hanqin Tian⁴¹, Christina Tonitto⁴², Margaret Torn³⁹, Ensheng Weng⁴³, Xiaolu Zhou³⁸

¹National Snow and Ice Data Center, Cooperative Institute for Research in Environmental Sciences, University of Colorado at Boulder, Boulder, CO, USA
(kevin.schaefer@nsidc.org)

²School of Earth Science and Environmental Sustainability, Northern Arizona University, Flagstaff, AZ, USA

³Graduate School of Geography, Clark University, Worcester, MA, USA

⁴School of Geography and Earth Sciences and McMaster Centre for Climate Change, McMaster University, Hamilton, Ontario, Canada
5Climate Research Division, Atmospheric Sciences and Technology Directorate, Environment Canada, Saskatoon, SK, Canada

6Department of Geography, University of Toronto, Toronto, Ontario, Canada

7Department of Meteorology, the Pennsylvania State University, University Park, PA, USA

8Northern Forestry Centre, Canadian Forest Service, Natural Resources Canada, Edmonton, AB, Canada

9Department of Biology, University of New Mexico, Albuquerque, NM, USA

10Forest Service, United States Department of Agriculture, Durham, NH, USA

11Department of Geography and Environmental Studies, Carleton University, Ottawa, Ontario, Canada

12Laboratoire des Sciences du Climat et l'Environnement, Gif Sur Yvette, , France

13Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, USA

14Department of Civil and Environmental Engineering, Princeton University, Princeton, NJ, USA

15Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA

16Departamento de Biología de la Conservación, Centro de Investigación Científica y de Educación Superior de Ensenada, Ensenada, BC, México; Department of Plant and Soil Sciences; Delaware Environmental Institute, University of Delaware, Newark, DE, USA
17Laboratory of Plant Ecology, Department of Applied Ecology and Environmental Biology, Ghent University, Ghent, Belgium

18Numerical Terradynamic Simulation Group, University of Montana, Missoula, MT, USA

19Atmospheric Science Department, Colorado State University, Fort Collins, CO, USA

20Faculty of Land and Food Systems, University of British Columbia, Vancouver, BC, Canada

21Department of Atmospheric and Oceanic Sciences, University of Wisconsin-Madison, Madison, WI, USA

22Department of Environmental Sciences (DES), University of Toledo, Toledo, OH, USA

23Department Evolution, Ecology, and Organismal Biology, Ohio State University, Columbus, OH, USA

24Department of Plant Biology, University of Illinois, Urbana, IL, USA

25Department of Geography, Indiana University, Bloomington, IN, USA

26Department of Biology, Virginia Commonwealth University, Richmond, VA, USA

27Department of Renewable Resources, University of Alberta, Edmonton, AB, Canada

28Department of Atmospheric Science, University of Illinois, Urbana, IL, USA

29Department of Agronomy and Nelson Institute Center for Sustainability and the Global Environment, University of Wisconsin-Madison, Madison, WI, USA

30Department of Forest Ecosystems and Society, Oregon State University, Corvallis, OR, USA
31Earth Resources Observation and Science, United States Geological Survey, Sioux Falls, SD, USA

32Centre d’Étude de la Forêt, Laval University, Quebec City, Quebec, Canada

33Biosciences Division, Argonne National Laboratory, Argonne, IL, USA

34Department of Geography, Queen's University, Kingston, Ontario, Canada

35Laboratory for Tree Ring Research, University of Arizona, Tucson, AZ, USA

36School of Engineering and Applied Sciences and Department of Earth and Planetary Sciences, Harvard University, Cambridge, MA, USA

37Department of Biology and the Global Change Research Group, San Diego State University, San Diego, CA, USA

38Institut of Environment Sciences, Biology Science Department, University of Quebec at Montreal, Montréal, QC, Canada

39Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA

40Department of Geography, McGill University, Montreal, PQ, Canada

41International Center for Climate and Global Change Research, School of Forestry and Wildlife Sciences, Auburn University, Auburn, AL, USA

42Department of Ecology and Evolutionary Biology, Cornell University, Ithaca, NY, USA

43Department of Botany and Microbiology, University of Oklahoma, Norman, OK, USA
Abstract

Accurately simulating Gross Primary Productivity (GPP) in terrestrial ecosystem models is critical because errors in simulated GPP propagate through the model to introduce additional errors in simulated biomass and other fluxes. We evaluated simulated, daily average GPP from 26 models against estimated GPP at 39 eddy covariance flux tower sites across the United States and Canada. None of the models in this study match estimated GPP within observed uncertainty. On average, models overestimate GPP in winter, spring, and fall, underestimate GPP in summer. Models over-predicted GPP under dry conditions and for temperatures below 0 °C. Improvements in simulated soil moisture and ecosystem response to drought or humidity stress will improve simulated GPP under dry conditions. Adding a low-temperature response to shut down GPP for temperatures below 0 °C will reduce the positive bias in winter, spring, and fall and improve simulated phenology. The negative bias in summer and poor overall performance resulted from mismatches between simulated and observed Light Use Efficiency (LUE). Improving simulated GPP requires better leaf-to-canopy scaling and better values of model parameters that control the maximum potential GPP, such as $\varepsilon_{\max}$ (light use efficiency), $V_{\text{cmax}}$ (unstressed Rubisco catalytic capacity) or $J_{\max}$ (the maximum electron transport rate).

1.0 Introduction

Terrestrial Gross Primary Productivity (GPP) is the total photosynthetic uptake or carbon assimilation by plants and is a key component of terrestrial carbon balance. GPP is the main carbon input to terrestrial ecosystems, noting relatively minor inputs by
dissolved organic carbon, as well as deposition by rainwater and sedimentation [Chapin et al., 2006]. GPP depends on climate, climate variability, disturbance history, water and nutrient availability, soil type, species composition, and community structure. Understanding how these factors influence GPP remains a challenge due to complex interactions and the difficulty in quantitatively measuring GPP directly at various temporal and spatial scales. Estimates of GPP are only available at eddy covariance flux tower sites for the past decade, so we depend on models to estimate GPP over long periods of time at regional and global scales, and to project future changes in GPP in response to climate change.

Any error in simulated GPP will propagate through the model, introducing errors in simulated biomass and fluxes. If simulated GPP is too low or too high, then predicted Leaf Area Index, wood biomass, crop yield, and soil biomass may also be too low or high [Schaefer et al., 2008]. Net Ecosystem Exchange (NEE) is total ecosystem respiration ($R_{eco}$) minus GPP, with a positive NEE indicating a net release of CO$_2$ to the atmosphere. Autotrophic respiration depends on GPP and heterotrophic respiration depends on soil conditions and dead plant biomass, so errors in GPP readily propagate to errors in $R_{eco}$ and simulated diurnal and seasonal cycles of NEE. Through representation of stomatal control on GPP and transpiration, errors in GPP introduce errors in simulated latent and sensible heat flux, which in turn can introduce error in simulated atmospheric circulation. GPP is a key carbon flux that needs to be simulated as accurately as possible to ensure the most reliable values of simulated biomass and surface fluxes.

We can classify GPP models into Enzyme Kinetic (EK), Light Use Efficiency (LUE), or empirical models. EK models represent leaf-scale enzyme-kinetics with
electron and product transport limits on simulated GPP [Farquhar et al., 1980; von Caemmerer and Farquhar, 1981; Collatz et al., 1991]. Nearly all EK models include a representation of stomatal conductance balancing GPP against water loss through leaf stomata [Collatz et al., 1991; Collatz et al., 1992]. Most stomatal conductance models are based on empirical correlations between conductance, photosynthesis, and either relative humidity [Ball et al., 1987] or Vapor Pressure Deficit (VPD) [Wang and Leuning, 1998]. LUE models estimate either GPP or Net Primary Productivity (NPP) by multiplying incident Photosynthetically Active Radiation (PAR) by a remotely sensed fraction of PAR absorbed by the vegetation ($f_{PAR}$) and an energy to biomass conversion factor (typically called Light Use Efficiency) [Monteith, 1972; Field, 1995; Prince and Goward, 1995; Landsberg and Waring 1997; Goetz et al., 1999; Running et al., 2000; Sims et al., 2006; Zhao et al., 2007; Sjöström et al., 2009; As-syakur et al., 2010].

Finally, empirical models use statistical relationships between observed environmental conditions and GPP estimated from eddy covariance flux data, which is then expanded to regional or global scales using various reanalysis weather products [Beer et al., 2010]. These GPP models vary greatly in intended application, complexity, and their representation of physical and biological processes.

The modeling community currently lacks a quantitative evaluation of multiple GPP models to gauge overall performance across different ecosystems and help prioritize long-term model development. Many modeling teams compare simulated NEE to observed NEE measured at single points using eddy covariance techniques [Baldocchi et al., 2001; Grant et al., 2010]. Others compare against NEE for large regions estimated from transport inversions that are optimally consistent with observations of atmospheric
CO2 concentration [Gurney et al., 2002; Peters et al., 2010]. However, comparisons with observed NEE do not distinguish between $R_{eco}$ and GPP and provide little information on model performance relative to GPP. Fortunately, NEE measured by eddy covariance techniques can be partitioned into $R_{eco}$ and GPP: a temperature function is tuned to nighttime $R_{eco}$, the function is used to calculate daytime $R_{eco}$, and the estimated GPP is the daytime $R_{eco}$ minus the daytime NEE [Desai et al., 2008; Lasslop et al., 2010]. There are assessments in the literature of how well terrestrial biosphere models simulate GPP, but they focused on a single or small number of models compared to GPP estimated from eddy covariance data at a small number of towers [e.g., Thornton et al., 2002; Schaefer et al., 2008; Verbeeck et al., 2008]. These studies used different techniques to estimate GPP from observed NEE [Desai et al., 2008], making it difficult to differentiate between errors in the partitioning technique and true model-data mismatches. The evaluations were run at different sites and used different input weather, making it difficult to isolate input errors from true model-data mismatches. The performance measures used in these evaluations are difficult to compare because most used qualitative performance criteria while those with quantitative performance measures used different statistical techniques and quantities. None of these model evaluations account for uncertainty in estimated GPP due to uncertainty in the eddy covariance data and partitioning techniques. The actual GPP value lies within the range defined by uncertainty, so the ideal performance target of any model is to match the observed GPP within uncertainty. These studies provide insight into the performance of individual models. However, the differences among evaluations make it very difficult to compare and synthesize the results to identify
strengths and weaknesses common to all terrestrial biosphere models and to determine what changes will provide the greatest improvements in simulated GPP.

We hypothesize that model performance depends on 1) model structure and 2) how models simulate GPP response to changing environmental conditions. Model structure refers to differences in how models represent various physical and biological processes, such as LUE vs. EK models. To test our hypotheses, we compared simulated GPP from 24 models against estimated GPP at 39 eddy covariance flux tower sites in the North American Carbon Program (NACP) Site-level Interim synthesis. Our analysis includes observation uncertainty to determine if the models hit the desired performance target: matching observed GPP within uncertainty. The number and variety of models and sites in the NACP Site Synthesis are sufficient to identify the strengths and weaknesses common to all GPP models and, most importantly, how to improve the models.

2.0 Methods

2.1 Estimated GPP

Our analysis used daily average GPP estimated at 39 eddy covariance flux tower sites (Table 1). Observed NEE at all towers were processed and partitioned into GPP and $R_{eco}$ using standard techniques [Barr et al., 2004]. The NACP Site Synthesis included 47 tower sites, but we did not include those sites where the GPP partitioning was not done or the algorithm failed to converge. The chosen sites represented eight major biome types across North America except tundra, with 24 sites from the AmeriFlux network and 15 sites from the Fluxnet Canada Research Network/Canadian Carbon Program.
partitioning was not done for tundra sites due to large data gaps in winter and the lack of nighttime data in summer to train the $R_{eco}$ model. NACP Site Synthesis used the International Global Biosphere Program (IGBP) biome classifications [Loveland et al., 2001]. Some models were designed for specific biome types, such as forest or agriculture sites only, so not every model simulated all sites, resulting in a total of 627 simulations and an average of 16 simulations per site.

Observed, hourly NEE at each site was gap-filled and decomposed into hourly $R_{eco}$ and GPP using a standard procedure [Barr et al., 2004] and converted into 24-hour, daily averages. Before processing, observed NEE was screened to remove outliers [Papale et al., 2006] and exclude data during periods of low turbulence based on a friction velocity threshold [Barr et al., 2011]. $R_{eco}$ was set equal to observed, nighttime NEE and fitted to an empirical model based on air and soil temperature using a moving window approach. The function was then used to calculate the daytime $R_{eco}$, which was subtracted from daytime NEE to get GPP. Finally, gaps in GPP were filled using an empirical model that was tuned to the estimated GPP values. GPP was set to zero when the soil was frozen and the air temperature was below 0 °C, assuming air temperature represents the temperature of the entire system. The partitioning procedure occasionally produced negative GPP values at dawn and dusk at most sites and occasionally for entire days at the dry, grassland sites. The negative GPP values typically greatly exceeded estimates of random uncertainty and probably resulted from errors in the $R_{eco}$ temperature response or the fact that the partitioning algorithm did not account for influences of soil moisture on GPP. We set all negative GPP values to zero and transferred the flux to $R_{eco}$ to maintain the observed NEE, and recalculated the daily average GPP. One third of the
models used a daily time step, so we calculated the daily average as the average rate of GPP over a 24-hour period using both estimated and gap-filled values.

Although treated here as “observed” GPP, they are not strictly observed, but rather estimated from observed tower-based NEE. The GPP includes all the strengths, weaknesses, and assumptions of the original NEE observations. The energy budget at flux towers does not balance because the eddy covariance technique captures small-scale turbulent fluxes less than 1 km, but can underestimate latent and sensible heat fluxes due to large-scale eddies on the order of 10 km [Foken, 2008]. Assuming similarity with the lack of surface energy balance closure, the measured NEE might be between 15-20% less than the actual values [Foken, 2008], indicating a potential underestimate of GPP. Eddy covariance techniques can underestimate nighttime $R_{eco}$, depending on the threshold used to filter out Reco under stable conditions, also resulting in an underestimate of GPP. The empirical formulation for $R_{eco}$ used to estimate GPP from NEE assumed air temperature represents the temperature of the entire system, underestimating GPP when the canopy temperature is greater than zero and the air temperature is less than zero. Lastly, the partitioning algorithm did not account for how soil moisture and other factors control $R_{eco}$, resulting in errors in the estimated GPP (as evidenced by occurrences of negative GPP at some grassland sites).

Total GPP uncertainty included gap filling algorithm uncertainty, partitioning uncertainty, random uncertainty, and threshold friction velocity ($u^*$) uncertainty, summed in quadrature. Summing in quadrature assumes that these sources of error are uncorrelated. We did not correct for potential biases due to the lack of energy closure or underestimates of $R_{eco}$ at night. Random and $u^*$ filtering uncertainty was estimated using
a Monte Carlo technique [Barr et al., 2011; Richardson and Hollinger, 2007]. Barr et al. [in prep.] generated a synthetic flux time series using the gap-filling algorithm, randomly introduced artificial gaps, added noise, and then refilled the gaps. Repeating the process 1000 times for each site-year produced probability distribution functions with the 2.5 and 97.5 percentiles representing uncertainty. Assuming the GPP uncertainty due to the gap-filling algorithm was the same as that for NEE, the GPP gap filling uncertainty was the fraction of filled GPP values for each day times the standard deviation of multiple algorithms [Moffat et al., 2007]. Partitioning uncertainty was based on the standard deviation of multiple partitioning algorithms [Desai et al., 2008]. Total uncertainty generally increased with the magnitude of GPP and varied from a minimum of ~1 µmol m$^{-2}$ s$^{-1}$ in winter to 2-4 µmol m$^{-2}$ s$^{-1}$ in summer. Random uncertainty dominated over other sources of uncertainty, ranging from ~90% of total uncertainty in summer months to ~50% of total uncertainty in winter months.

2.2 Modeled GPP

Our analysis used simulated GPP from 24 different models (Table 2) plus two model averages. We used the model characteristics in Table 2 as covariates to determine if the different representations of physical and biological processes produce statistically significant differences in model performance relative to GPP. All the EK models included the effects of stomatal conductance and all but two used the Ball-Berry stomatal conductance model [Ball et al., 1987]. The leaf-to-canopy column indicates whether the strategy for scaling from a single leaf to the entire canopy accounts for the effects of diffuse light on shaded leaves (2-leaf) or not (big-leaf). To determine if resolving the
diurnal cycle improved performance, we included the ensemble average of all models and the average of all models that resolved the diurnal cycle. Including the two ensemble averages at each site, we have a total of 627 simulations or 4242 site-years of model output with an average of 23 simulations per model.

We included estimates of MODIS GPP from Collection 5.0 and Collection 5.1, [Heinsch et al., 2003; Running et al., 2004]. We extracted a 3 by 3 pixel window of 8-day maximum composite GPP values at 1 km² spatial resolution, with the center pixel containing the tower site [MODIS GPP, 2010]. We filtered out low quality pixels using the simple binary quality control flag to remove the effects of potential cloud contamination and averaged the rest of the pixels to represent the GPP at each site. We linearly interpolated between 8-day composite values to obtain daily GPP.

The MODIS GPP from Collection 5.0 and Collection 5.1 were not based on observed meteorology, so we also calculated GPP using gap-filled observed weather and the MODIS algorithm [Heinsch et al., 2003; Running et al., 2004]:

\[
GPP = 0.45 \varepsilon_{\text{max}} F_{SW} f_{\text{PAR}} S_{\text{VPD}} S_{T}
\]

where \( \varepsilon_{\text{max}} \) is the maximum light use efficiency, \( F_{SW} \) is the incident shortwave radiation flux, \( f_{\text{PAR}} \) is the absorbed fraction of PAR, \( S_{\text{VPD}} \) is the VPD scaling factor, and \( S_{T} \) is the air temperature scaling factor. \( S_{\text{VPD}} \) represents the GPP response to drought and humidity stress and \( S_{T} \) represents the GPP response to temperature, with both varying between zero and one. We used the MODIS Biome-Property-Look-Up-Table [Zhao and Running, 2010] and daily \( f_{\text{PAR}} \) values interpolated from the monthly mean GIMMSg NDVI dataset [Tucker et al., 2005]. Although the complexity and sophistication varies widely, all the
models have a GPP formulation similar to MODIS: a peak potential rate times the amount of absorbed light, multiplied by a series of scaling factors representing how GPP responds to changing environmental conditions. The scaling factors represent the ratio of actual to a reference or optimal GPP and vary between zero and one.

All models used gap-filled observed weather from each tower site [Ricciuto et al., 2009; http://www.nacarbon.org/nacp/] with input parameters and biophysical characteristics derived from local observations, such as soil texture. Missing air temperature, atmospheric humidity, shortwave radiation, and precipitation data were filled using DAYMET [Thornton et al., 1997] 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 values using the phasing from observed mean diurnal cycles calculated from a 15-day moving window. When station data were unavailable, a 10-day running mean diurnal cycle was used [Ricciuto et al., 2009; http://nacp.ornl.gov/docs/Site_Synthesis_Protocol_v7.pdf]. The models were run for as many years as required, repeating the gap-filled weather, until they reached steady state initial conditions where $R_{ec}$ balances GPP and the average NEE over the entire simulation is near zero. The steady state assumption influences $R_{ec}$, but has little or no effect on simulated GPP. All models used their standard values for various biophysical parameters except LoTEC, which used optimized parameter values obtained through data assimilation [Ricciuto et al., 2011].
2.3 Model Performance

We quantified model performance using a statistical analysis of model-data residuals using daily average GPP. We first calculated residuals:

\[ r_i = (GPP_{si} - GPP_{oi}), \]

where \( r_i \) is the residual for the \( i \)th model-data pair, \( GPP_{si} \) is simulated daily average GPP, and \( GPP_{oi} \) is estimated daily average GPP. Bias is the residual mean and the Root Mean Square Error (RMSE) is the residual standard deviation. \( \chi^2 \) is the mean of residuals normalized by uncertainty:

\[ \chi^2 = \frac{1}{n} \sum_{i=1}^{n} \frac{(r_i)^2}{\varepsilon_i^2}, \]

where \( n \) is the number of residuals and \( \varepsilon_i \) is the uncertainty for the \( i \)th daily average estimated GPP. We filtered out ~0.1% of daily estimated GPP values with \( \varepsilon_i \leq 0.3 \) µmol m\(^{-2}\) s\(^{-1}\) which produced extreme outlier \( \chi^2 \) values that skewed our results. Such unrealistically small \( \varepsilon_i \) values occasionally occurred when GPP was near zero. We did not filter out daily average GPP values based on the number of filled values per day because these values have higher uncertainty and a proportionally lower influence on \( \chi^2 \).

The ideal target for any model is \( \chi^2 < 1.0 \), which means, on average, the residuals are less than uncertainty or the model matches observations within measurement uncertainty. Variations of \( \chi^2 \) within this target range have no meaning relative to model performance. A model with \( \chi^2 \) of 0.8, for example is not “better” than a model with \( \chi^2 \) of 0.9, since both models show no statistically significant differences with observations. Consequently, we identified performance categories based on ranges of \( \chi^2 \) values.
1.0 indicated good model performance. $X^2$ between one and two indicated marginal model performance, where the model-data mismatch is on the order of two times the observation uncertainty. $X^2 > 2.0$ indicated poor model performance, where the model-data mismatch is several times the observation uncertainty. A $X^2$ of nine, for example, indicated that the model-data mismatch is, on average, three times the uncertainty.

To test our hypothesis that model structure influence performance, we aggregated the performance measures by model, model characteristic, site, and month-of-year. To identify any statistically significant differences in performance based on how the models represented various physical and biological processes, we aggregated performance measures by the model structural characteristics listed in Table 2. To evaluate seasonal variation in performance parameters, we aggregated by month-of-year, where January is the average of all Januaries, February the average of all Februaries, etc.

To test our hypothesis that model performance depends on how they represent the GPP response to changing environmental conditions, we compared observed and simulated environmental response curves. We sorted the daily average GPP values into bins based on daily average values of input driver variables and calculated the mean, standard deviation, and uncertainty of the daily average GPP for each bin. We focused on downwelling shortwave radiation flux, air temperature, and relative humidity. The relative humidity response function reflects the reduction in GPP due to stomatal closure under drier atmospheric conditions. VPD, the difference between saturated and actual water vapor pressure, also reflects stomatal closure, but varies with temperature such that the range and magnitude varies among sites. Relative humidity always varies between zero and one and greatly simplifies our analysis by allowing easy comparison among
sites. More importantly, 16 of the 20 models in this analysis that account for stomatal
conductance used the Ball-Berry stomatal conductance model, which is based on relative
humidity. Each model’s mathematical formulation and associated parameters values
determined the shape of the simulated response curves. We compared simulated and
observed shape characteristics, such as slope, to isolate those model formulations or
parameters that determine model performance. The response function for each driver
variable differs slightly from site to site and is only weakly correlated to response
functions for other driver variables. For example, the optimal temperature for GPP is
different for each site, but does not depend on humidity or light. Consequently, we made
no attempt to remove covariance between driver variables.

3.0 Results

3.1 Performance Summary

None of the models in this study achieved a good overall performance of $\chi^2$ less
than one for all sites (Figure 1). LoTEC, which was optimized against flux data and
DLEM achieved marginal performance, while SSIB2 and TECO had large $\chi^2$ values due
to large biases. Generally speaking, higher RMSE resulted in higher $\chi^2$, but the
relationship was weak because $\chi^2$ accounts for uncertainty while RMSE does not. There
was no relationship between model bias and either RMSE or $\chi^2$. The RMSE fell within a
narrow performance range, with a mean and standard deviation of 2.8±1.0 $\mu$mol m$^{-2}$ s$^{-1}$.
On average, the models underestimated GPP (negative bias) with a mean bias of -0.3
$\mu$mol m$^{-2}$ s$^{-1}$, but the spread between models was very large compared to the mean, with a
bias standard deviation of ±0.6 $\mu$mol m$^{-2}$ s$^{-1}$.
Figure 2 shows that models, on average, underestimated GPP in summer (negative bias) and overestimated GPP in winter, spring, and fall (positive bias). These seasonal biases of opposite sign tended to cancel, resulting in the lower overall biases seen in Figure 1. Figure 2 shows the average monthly bias, but every month showed both positive and negative biases for individual models with standard deviations ranging from two to ten times the mean bias. The estimated GPP is smaller in spring and fall, with correspondingly smaller uncertainties, which magnified the relatively small model biases to produce slight peaks in $\chi^2$ in spring and fall. The models performed worst in the summer and the best in winter, indicating the models properly shut down GPP during winter, but the real challenge is to capture GPP dynamics during the growing season.

The models generally performed the best at forest sites and the worst at crop, grassland, and savanna sites (Figure 3). The models did not show good overall performance at any site, but did show marginal performance at seven sites. The models performed best for deciduous broadleaf, mixed forest, and evergreen needleleaf biome types and all of the ten sites with the best overall performance were forest. The spread in performance within biome types was broad: two of the ten sites with the worst performance were evergreen needleleaf forest sites. However, seven of the ten sites with the worst model performance were crop, grassland or savanna sites.

The models showed a large spread in both the magnitude and timing of the simulated GPP seasonal cycle, as indicated by the large spread in Figure 4. The models captured the basic observed seasonal pattern in GPP with near zero values in winter and a peak value in mid-summer, so the standard deviation of the seasonal cycle is an alternative measure of seasonal amplitude. Consequently, the radial distance in Figure 4
is effectively the ratio of simulated to estimated seasonal amplitude of GPP, an alternative measure of bias. Ratios less than one indicate the model underestimates the GPP seasonal amplitude (negative bias). The models had standard deviation ratios ranging from 0.5 to 1.4, which means the simulated GPP ranged from 50% to 140% of estimated GPP. The models showed correlations between 0.6 and 0.9, indicating they varied widely in how well they captured the timing of the estimated GPP seasonal cycle, regardless of how well they captured the magnitude. For example, LOTEC, ISOLSM, and LPJ all had standard deviation ratios near one, consistent with the low biases seen in Figure 1. However, these three models showed progressively smaller correlations and correspondingly larger $\chi^2$, indicating that simulated phenology and associated phasing of the GPP seasonal cycle played an important role in determining overall model performance.

The two model means showed the highest correlation with observations. This indicates that model-data mismatches associated with the timing of the GPP seasonal cycle can partly cancel out when averaging the results from multiple models. Essentially, the ensemble mean gave better results than any single model alone. Although errors in timing cancel, an ensemble mean does not eliminate overall bias. The standard deviation ratios for the two model means in Figure 4 reflected the overall, average negative bias of all the models in our analysis. Although there was positive bias in winter, spring, and fall, the negative bias in summer dominated and, on average, the models as a whole underestimate GPP by 20-30%.

These results complement and extend previous analyses from the NACP Site Synthesis. Richardson et al. [2011] found that overestimation of GPP in spring and fall
resulted in models predicting the start of spring uptake about two weeks earlier than observed and the end of uptake in fall about two weeks later than observed. Schwalm et al., [2010] found models simulate NEE better at forest sites than grassland sites. The positive biases in spring and fall can help explain the decreased model performance relative to NEE in spring and fall [Schwalm et al., 2010]. Underestimating GPP in summer can explain the peaks in the spectral signature of NEE residuals at the annual time scale [Dietze et al., 2011]. Our results and those from prior studies indicate that seasonal biases in simulated GPP can help explain problems in simulating seasonal changes in NEE.

Overall, there was a very large spread in model performance. On average, a single model showed good performance at 12% of the sites, marginal performance at 26%, and poor performance for the rest. Nearly every model had one “outlier” site where it performed considerably worse than the other sites, with $\chi^2$ values often exceeding 20. Conversely, nearly all models showed good or marginal performance at least one site. The spread in performance across sites was equally broad, with three outlier sites where none of the models performed well. The spread among models at a single site was also wide: each site, on average, had two good simulations, four marginal simulations, and two outlier simulations with $\chi^2 > 20$.

3.3 Model Structure

Model performance did not depend on model structure, as defined by the model characteristics in Table 2. We did not find any statistically significant relationships between performance and how models represent various physical and biological
processes (Table 3). In all cases, the difference in mean values between model groups was much smaller than the standard deviation within groups such that none of the differences were statistically significant. For example, LUE models performed better than EK models, but when excluding SSIB2 and TECO, which had large biases, EK models performed better than LUE models. Essentially, EK and LUE models performed equally well in simulating observed GPP. The performance difference for a daily vs. hourly time step was consistent with the difference between EK and LUE models, since nearly all EK models use an hourly time step. The same was true for the other model structural characteristics: models that include a nitrogen cycle, a soil model, shaded leaves, or prognostic phenology performed equally well as models that do not.

We found no significant relationships, but this does not mean that model structure does not influence performance. Sprintzin et al. [2012], for example, clearly demonstrate that accounting for diffuse light and changing from a big-leaf to a 2-leaf formulation improved BEPS performance. However, SiB is a big-leaf model and performed just as well as BEPS. The lack of significant relationships means that model performance is dominated by some other aspect of model design not represented by the model characteristics in Table 3, such as how models simulate GPP responses to changing environmental conditions.

3.2 Light Response

The poor overall performance and the negative bias in summer resulted from mismatches between simulated and observed LUE. The light response curve is GPP as a function of downwelling shortwave radiation and its slope is the LUE. Figure 5 shows a
light response curve based on daily average GPP for US-Me2, which we chose because it
had a large number of simulations and was typical of all sites. The uncertainty in Figure
5 was dominated by gap-filling and partitioning uncertainty because the bin averaging
tended to greatly reduce the random uncertainty. US-Me2 is an evergreen needleleaf site
with simulations from 19 models with a marginal overall performance ($X^2 = 1.9$). Five
models had good performance ($X^2 \leq 1.0$), two showed marginal performance ($1.0 < X^2 \leq
2.0$), and the rest showed poor performance. Four out of the five models with good
performance all had LUEs that matched observed values within uncertainty. We saw no
clear pattern in bias, with each model under-predicting GPP at some sites and over-
predicting at others, but the spread in simulated GPP among models at US-Me2 was
typical of all sites.

Bias decreased as the ratio of simulated to observed LUE approached one (Figure
6). We calculated the observed and simulated LUE as the regression of GPP vs.
shortwave radiation flux with a Y-intercept forced to be zero. Both the observed and
simulated light response curves were noisy, so we forced the Y-intercept to be zero to
guarantee that GPP was zero for zero incident shortwave light. The LUE ratio was
simulated LUE divided by observed LUE, with a ratio less than one indicating the model
underestimated GPP (negative bias). The LUE ratios formed a diagonal line with near
zero bias when the LUE ratio was one. Plots of LUE ratios for subgroups defined by the
model structural characteristics in Table 3, individual models, and individual sites all
showed the same pattern as in Figure 6: when the LUE ratio was one, the bias was zero.

To improve performance in simulated GPP, model developers should focus first
on those parameters that determine the simulated LUE. The LUE is determined by the
leaf-to-canopy scaling and a small number of parameters that define the maximum potential GPP. For the MODIS algorithm described above, for example, the LUE is determined by $\varepsilon_{\text{max}}$, so a better value of $\varepsilon_{\text{max}}$ will improve performance. For other models, $V_{c_{\text{max}}}$ (the unstressed Rubisco carboxylation rate), $\alpha$ (quantum yield), or $J_{\text{max}}$ (the maximum electron transport rate) determine LUE. These maximum potential GPP parameters are either held constant for all sites, like in ECOSYS [Grant et al., 2009] or vary with biome or Plant Functional Type (PFT), like SiB3 [Baker et al., 2008]. Models account for changing environmental conditions by multiplying the maximum potential GPP by temperature, moisture, and humidity scaling factors that represent the ratio of actual to peak GPP.

How a model scales from a single leaf to an entire canopy also influences the simulated LUE. The maximum potential GPP parameters typically represent peak or optimal values for a single leaf at the canopy top. The leaf-to-canopy scaling factor represents the ratio of GPP for a single leaf to GPP for the entire canopy. A model assumes the distribution of leaf nitrogen and light levels within the canopy, and integrates from canopy top to bottom to calculate a leaf-to-canopy scaling factor. SiB and BEPS, for example, both assume the distribution of light is governed by Beer’s law [Sellers et al., 1996; Sprintsin et al., 2012]. Unfortunately, the leaf-to-canopy scaling and the maximum potential GPP parameters are coupled and can compensate, indicating that the model has to get both right to get the correct GPP.

Our results indicate that better LUE parameter values and leaf-to-canopy scaling will improve overall performance in simulated GPP, although we could not delve into
individual models to identify the correct parameters and the best values. The number, nomenclature, definition, and units of the various parameters that define LUE differ widely among models. Model developers can use data assimilation of, for example, eddy covariance data to estimate parameter values [Hanson et al., 2004]. The LOTEC model illustrates the potential to improve performance by estimating the maximum potential GPP parameters with data assimilation. However, due to differences among models, parameter values estimated for one model may not work in another. The TRY database of plant characteristics [Kattge et al., 2011] includes observations of $V_{cmax}$, $J_{max}$, and $\alpha$ compiled from many studies scattered that could potentially minimize these inter-model differences. The leaf-to-canopy scaling depends on the assumed variation of light levels and parameter values within the canopy, which may require additional field observations. For example, measuring leaf-level nitrogen content, which determines $V_{cmax}$, is relatively easy, but what models need is canopy-level nitrogen content, which, unfortunately, is much more difficult to measure. Changes to how models treat the distribution of light within the canopy could improve the leaf-to-canopy scaling, such as better canopy radiative transfer models coupled to the GPP models or separating sunlit and shaded leaves, but developers have to demonstrate that such changes improve GPP performance. The observed LUE shows strong variability within PFT classes and the biases may result from the fact that the models assume constant parameter values for each PFT. Although spatially explicit maps of LUE parameter values currently do not exist, remote sensing of canopy nitrogen shows promise [Ollinger et al., 2008]. However, nitrogen in Rubisco is the variable of interest and relating total nitrogen to a canopy $V_{cmax}$ in a way that will work for all models will require new theoretical development.
3.4  

Humidity Response

We found that difficulties in simulating GPP under dry conditions can explain why models performed worse at grassland and savanna sites than forest sites. Figure 7 shows normalized humidity response curves for the US-Var grassland and US-Ha1 deciduous broadleaf forest sites. We normalized the humidity response curves to emphasize the shape of the curves, which were typical of all sites: low GPP under dry conditions, an optimal GPP at 70-80% relative humidity, and a decrease for higher humidity associated with colder temperatures. Models calculate lower GPP under stressed conditions using scaling factors that represent the ratio of stressed to optimal GPP. The scaling factors determine the shape of the response curves in Figure 7, but the simulated LUE determines the GPP magnitude. Decreased GPP at low relative humidity can be caused either by humidity stress reducing stomatal conductance, high temperature stress, or drier soils with reduced water availability (drought stress). Half of the models over-predicted GPP at both sites under low humidity conditions (relative humidity less than 60%). Such dry conditions occurred only 23% of the time at US-Ha1, but occurred 46% of the time at US-Var, which has dry summers with near zero growth. Even though half the models did not capture GPP under dry conditions at both sites, the effect on performance was much stronger at US-Var because the dry periods occurred twice as often than at US-Ha1. This explains the poor performance at the evergreen needleleaf forest sites CA-SJ1 and CA-SJ2, where the dry periods occurred nearly as often as US-Var. Essentially, the more often the dry periods occurred, the worse the performance.

The NEE partitioning algorithm can partly explain the model-data mismatch at drier sites. The algorithm was not designed for drier sites where soil moisture has greater
influence on $R_{eco}$ than temperature. Consequently the algorithm either did not converge or produced negative GPP, which we changed to zero as described above. However, filtering out these zero GPP values did not change model performance at these sites. The partitioning algorithm could be improved to account for moisture, but the models also need to improve simulated GPP under dry conditions.

Determining exactly how to improve simulated GPP under dry conditions was not possible in our analysis because the effects of drought and humidity stress are intertwined. Periods of low rainfall simultaneously reduce both soil moisture and atmospheric humidity, making the associated effects of drought and humidity stress on GPP difficult to separate. Models that account for drought stress typically calculate a GPP scaling factor using either input precipitation or plant water availability from simulated soil moisture [Schaefer et al., 2008; Potter et al., 1993]. Models that account for humidity stress either calculate a GPP scaling factor based on humidity or directly reduce stomatal conductance [Heinsch et al., 2003; Baker et al., 2008]. Thus, over-predicting GPP under dry conditions could result from problems with the simulated soil moisture, the calculation of plant water availability, or the representation of humidity stress.

To complicate matters, a model’s representation of humidity stress can compensate for poor representation of drought stress, and vice versa. For example, the MODIS algorithm above does not account for drought stress at all, but the humidity response was strengthened to compensate [Heinsch et al., 2003], such that MODIS reproduced the shape (but not magnitude) of the observed humidity response curves. Determining whether models should improve simulated soil moisture, drought stress, or
humidity stress requires a simultaneous analysis of simulated and observed soil moisture, latent heat flux, and GPP, which is beyond the scope of our analysis.

3.6 Temperature response

Figure 8 shows a typical temperature response function for the evergreen needleleaf forest site, US-Ho1, which had simulations from 23 models and a marginal overall performance ($X^2 = 1.8$). US-Ho1 was the site closest to the “average” temperature response function for all sites. The observed optimal temperature for GPP was 19 °C and average across all sites was 20±5 °C. Differences between simulated and observed GPP near the peak or optimal temperature reflected differences in simulated and observed LUE in summer, as described above. For US-Ho1, GPP shut down for daily average temperatures below -6.5 °C and the average low temperature cutoff across all sites was -6±3 °C. Cutoff temperatures below zero reflected conditions in spring and fall where daytime temperatures were above freezing to allow photosynthesis while the nighttime temperatures were below freezing, resulting in a negative daily average air temperature. The average winter season at US-Ho1, defined as the time with daily average temperature below 0 °C, was 92 days and the average for all sites was 75±47 days.

Temperature was the dominant control of the seasonal variation of GPP at most sites. The simulated start of GPP in spring and the stop of GPP in fall is a representation of phenology. Models primarily use temperature to control phenology, but how this is done varies widely. Some models using growing degree days to predict bud burst in spring while others simply shut down GPP below a lower temperature limit. How
models simulate GPP at temperatures near 0 °C determined the start and stop of the simulated growing season.

The models tended to over-predict GPP at low temperatures: half of all simulations predicted more than double the observed GPP at temperatures below 0 °C. The observations indicated that no more than 0.8±0.6% of total annual GPP occurred in winter, but the models simulated anywhere from 0% to 15% of the annual GPP in winter. Part of this may have resulted from the partitioning algorithm itself, which set GPP to zero when the soil and the air temperature are below zero. However, observations indicate this is realistic since the recovery of photosynthesis after freezing temperatures can be delayed for weeks with repeated exposure to frost and cold and frozen soils limit root uptake of water and stomatal conductance [Strand and Oquist, 1985; Waring and Winner, 1996]. Over-predicting GPP under cold conditions explained the positive biases in winter, spring, and fall, which in turn resulted in uptake starting earlier than observed in spring and later than observed in fall.

Better low temperature inhibition functions will improve simulated GPP in winter, spring, and fall and improve simulated phenology. Most models use an exponential or “$Q_{10}$” response function to represent the effects of low temperature on GPP:

$$S_T = \left( Q_{10} \right)^{\frac{T-T_{\text{ref}}}{10}},$$

where $S_T$ is a temperature scaling factor applied to GPP, $T$ is temperature (°C), and $T_{\text{ref}}$ is a reference temperature (°C). All models based on the Farquhar et al. [1980] EK model, for example, use this type of formulation. Most models have a second exponential
function with separate $Q_{10}$ and $T_{ref}$ values to reflect reduced GPP for high temperatures. Combined, the low and high temperature functions produce an optimal temperature for simulated GPP. The combined temperature scaling factor represents the ratio of actual to optimal GPP and varies between zero and one. The simulated LUE determined GPP magnitude under optimal conditions in mid-summer, but the temperature scaling determined the seasonal cycle in simulated GPP and simulated phenology. The exact values of $Q_{10}$ and $T_{ref}$ vary widely between models and we made no attempt to determine which values are correct.

The positive bias in winter, spring, and fall resulted from the fact that the $Q_{10}$ function alone will never reach zero no matter how cold the temperature, so those models using this type of formulation without a frost inhibition function over predicted GPP at low temperatures. The frost inhibition function is an additional scaling factor that shuts down GPP below a specified threshold temperature [Kucharik and Twine, 2007; Li et al., 2010]. In addition, some models also include a GPP recovery period after the frost event [Baker et al., 2008; Schaefer et al., 2008]. Data on photosynthesis at low temperatures are relatively scarce, so developing a low temperature inhibition function to incorporate the effects of nutrient and water availability in partially frozen soils may require more observations. Improving the modeled low temperature inhibition function will improve simulated GPP in spring and fall, and thus simulated phenology.

5.0 Conclusions

None of the models in this study match estimated GPP within the range of uncertainty of observed fluxes. On average, the models achieved good performance for
only 12% of the simulations. Two models achieved overall marginal performance, matching estimated GPP within roughly two times the uncertainty. Our first hypothesis proved false: we found no statistically significant differences in performance due to model structure, mainly due to the large spread in performance among models and across sites. The models in our study reproduced the observed seasonal pattern with little or no GPP in winter and peak GPP in summer, but did not capture the observed GPP magnitude. We found, on average, that models overestimated GPP in spring and fall and underestimated GPP in summer. Our second hypothesis proved true: model performance depended on how models represented the GPP response to changing environmental conditions. We identified three areas of model improvement: simulated LUE, low temperature response function, and GPP response under dry conditions.

The poor overall model performance resulted primarily from inadequate representation of observed LUE. Simulated LUE is controlled by the leaf-to-canopy scaling strategy and a small set of model parameters that define the maximum potential GPP, such as $\varepsilon_{\text{max}}$ (light use efficiency), $V_{\text{cmax}}$ (unstressed Rubisco catalytic capacity) or $J_{\text{max}}$ (the maximum electron transport rate). The temperature, humidity, and drought scaling factors determined temporal variability in simulated GPP, but the LUE parameters determined the magnitude of simulated GPP. To improve simulated GPP, model developers should focus first on improving the leaf-to-canopy scaling and the values of those model parameters that control the LUE.

Many models over-predicted GPP under dry conditions, explaining why, on average, models performed worse at grassland and savanna sites than at forest sites. The importance of this to model performance increases at sites where drier conditions occur
more frequently. Since dry conditions occur more frequently at grassland and savanna sites than at forest sites, models tended to perform worse at grassland and savanna sites compared to forest sites. Improving how models simulate soil moisture, drought stress, or humidity stress can improve simulated GPP under dry conditions.

Many models over-predicted GPP under cold conditions, partly explaining the positive bias in simulated GPP in winter, spring, and fall. The estimated GPP completely shut down for daily average temperatures less than -6 °C, but the $Q_{10}$ formulation used by many models did not shut down GPP under cold or frozen conditions. The simulated GPP started too early in spring and persisted too late in fall, resulting in a positive bias and phasing errors in phenology. Using an ensemble mean can cancel out errors in phenology, but does not cancel out bias. Improving or imposing a low temperature inhibition function in the GPP model will resolve the problem.

6.0 Acknowledgments

We thank the North American Carbon Program Site-Level Interim Synthesis team 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. We thank Dennis Baldocchi, Lawrence Flanagan, Ni Golaz, Gabriel Katul, Kim Novick, Paul Stoy, and Shashi B. Verma for providing valuable data and advice during the development of this paper. This research was partly funded by NOAA Award NA07OAR4310115 and U.S. National Science Foundation grant ATM-0910766, and by the U.S. Department of Energy’s Office of Science for AmeriFlux Science Team research.
to develop measurement and data submission protocols and conduct quality assurance of measurements for AmeriFlux investigators (Grant No. DE-FG02-04ER63911).

7.0 References


Grant, R. F., Barr, A. G., Black, T. A., Margolis, H. A., McCaughey, J. H., Trofymow, J. A. (2010), Net ecosystem productivity of temperate and boreal forests after
clearcutting - a Fluxnet-Canada measurement and modelling synthesis, *Tellus*, 62B: 475–496.


MODIS Gross Primary Photosynthesis (GPP) MOD17A2 and MYD17A2 (2010), ASCII Subsets, 7x7-km, 8-day, January 1, 2001 Forward. Available online at [http://www.modis.ornl.gov/modis/index.cfm] from the Oak Ridge National Laboratory (ORNL) Distributed Active Archive Center (DAAC), Oak Ridge, Tennessee, USA.


source to springtime carbon sink in a high-elevation, subalpine forest, *Oecologia* 146: 130-147.


Peters, W., et al. (2010), Seven years of recent European net terrestrial carbon dioxide exchange constrained by atmospheric observations, *Global Change Biology*, 16(4), 1317-1337.


Running, S. W., P. E. Thornton, R. Nemani and J. M. Glassy (2000), Global terrestrial gross and net primary productivity from the Earth observing system. In O. E.


Waring, R.H., and W.E. Winner (1996), Constraints on terrestrial primary production along the Pacific Coast of North and South America. In “High Latitude Rain


Figure captions

**Figure 1**: Overall model performance per model for all sites as defined by $\chi^2$ (a), RMSE (b), and bias (c). The models are arranged in order of increasing $\chi^2$ and color coded for Enzyme Kinetic (EK), Light Use Efficiency (LUE) and model means. $\chi^2$ less than one indicates good performance, $\chi^2$ between one and two indicates marginal performance, and $\chi^2$ greater than two indicates poor performance.

**Figure 2**: The monthly average bias in simulated GPP (a) and monthly $\chi^2$ (b) based on all 627 simulations from all models. A $\chi^2$ below 2.0 indicates marginal performance.

**Figure 3**: Overall performance per site for all models as defined by $\chi^2$ (a), RMSE (b), and bias (c). The sites are arranged in order of increasing $\chi^2$ and color coded by biome type.

**Figure 4**: Taylor plot by model for all sites. A Taylor plot is a polar plot where the cosine of the angle from the x-axis is the correlation coefficient between simulated and observed GPP. The correlation coefficient measures how well the simulated GPP captures the phasing and timing of the observed GPP. The radial direction is the ratio of simulated to observed standard deviation. The green lines represent Root Mean Squared Error normalized by standard deviation. An ideal model would have a standard deviation ratio of one and a correlation of one (point A).
Figure 5: The light response curve for US-Me2 showing simulated and observed GPP as a function of downwelling shortwave radiation. The grey bar indicates uncertainty in the observed GPP response to shortwave radiation.

Figure 6: GPP bias as a function of the ratio of simulated to observed LUE for summer (June-July-August) for all 627 simulations.

Figure 7: Normalized humidity response curves for US-Var (a) and US-Ha1 (b) showing simulated and observed GPP as a function of relative humidity. The grey bar indicates uncertainty in the observed GPP response. All response curves are normalized such that the maximum value is one.

Figure 8: The temperature response curve for US-Ho1 showing simulated and observed GPP as a function of daily mean air temperature. US-Ho1 is the site closest to the average response for all sites. The grey bar indicates uncertainty in the observed GPP response to air temperature.
### Table 1: Summary of site characteristics

<table>
<thead>
<tr>
<th>Site</th>
<th>IGBP Class</th>
<th>Description</th>
<th>longitude (deg)</th>
<th>latitude (deg)</th>
<th>Start (yr)</th>
<th>Stop (yr)</th>
<th>Number Models</th>
<th>Reference(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CA-Ca1</td>
<td>ENF</td>
<td>Campbell River - Mature Douglas-fir</td>
<td>-125.3</td>
<td>49.9</td>
<td>1998</td>
<td>2006</td>
<td>23</td>
<td>Krishnan et al. [2009]; Humphreys et al. [2006]</td>
</tr>
<tr>
<td>CA-Ca2</td>
<td>ENF</td>
<td>Campbell River - Douglas-fir, clearcut</td>
<td>-125.3</td>
<td>49.9</td>
<td>2001</td>
<td>2006</td>
<td>8</td>
<td>Krishnan et al. [2009]; Humphreys et al. [2006]</td>
</tr>
<tr>
<td>CA-Ca3</td>
<td>ENF</td>
<td>Campbell River - Douglas-fir, juvenile</td>
<td>-124.9</td>
<td>49.5</td>
<td>2002</td>
<td>2006</td>
<td>8</td>
<td>Krishnan et al. [2009]; Humphreys et al. [2006]</td>
</tr>
<tr>
<td>CA-Gro</td>
<td>MF</td>
<td>Groundhog River</td>
<td>-82.2</td>
<td>48.2</td>
<td>2004</td>
<td>2006</td>
<td>17</td>
<td>McCaughey et al. [2006]</td>
</tr>
<tr>
<td>CA-Let</td>
<td>GRA</td>
<td>Lethbridge Grassland</td>
<td>-112.9</td>
<td>49.7</td>
<td>1997</td>
<td>2007</td>
<td>22</td>
<td>Flanagan and Adkinson [2011]</td>
</tr>
<tr>
<td>CA-Mer</td>
<td>WET</td>
<td>Eastern Peatland - Mer Bleue</td>
<td>-75.5</td>
<td>45.4</td>
<td>1999</td>
<td>2006</td>
<td>18</td>
<td>Roulet et al. [2007]</td>
</tr>
<tr>
<td>CA-Oas</td>
<td>DBF</td>
<td>BERMS - Old Aspen</td>
<td>-106.2</td>
<td>53.6</td>
<td>1997</td>
<td>2006</td>
<td>23</td>
<td>Krishnan et al. [2006]; Barr et al. [2006]</td>
</tr>
<tr>
<td>CA-Obs</td>
<td>ENF</td>
<td>BERMS - Old Black Spruce</td>
<td>-105.1</td>
<td>54.0</td>
<td>2000</td>
<td>2006</td>
<td>22</td>
<td>Krishnan et al. [2008]; Kjun et al. [2006]</td>
</tr>
<tr>
<td>CA-Ojp</td>
<td>ENF</td>
<td>BERMS - Old Jack Pine</td>
<td>-104.7</td>
<td>53.9</td>
<td>2000</td>
<td>2006</td>
<td>18</td>
<td>Kjun et al. [2006]</td>
</tr>
<tr>
<td>CA-Qfo</td>
<td>ENF</td>
<td>Quebec Mature Black Spruce</td>
<td>-74.3</td>
<td>49.7</td>
<td>2004</td>
<td>2006</td>
<td>17</td>
<td>Bergeron et al. [2008]; Bergeron et</td>
</tr>
<tr>
<td>Location</td>
<td>Type</td>
<td>Site</td>
<td>Description</td>
<td>t</td>
<td>n</td>
<td>1990s Start</td>
<td>1990s End</td>
<td>2000s Start</td>
</tr>
<tr>
<td>----------</td>
<td>------</td>
<td>------</td>
<td>-------------</td>
<td>---</td>
<td>---</td>
<td>-------------</td>
<td>-----------</td>
<td>-------------</td>
</tr>
<tr>
<td>CA-SJ1</td>
<td>ENF</td>
<td>BERMS - Jack Pine, 1994 harvest</td>
<td>-104.7</td>
<td>53.9</td>
<td>2002</td>
<td>2005</td>
<td>7</td>
<td>Zha et al. [2009]</td>
</tr>
<tr>
<td>CA-SJ2</td>
<td>ENF</td>
<td>BERMS - Jack Pine, 2002 Harvest</td>
<td>-104.6</td>
<td>53.9</td>
<td>2003</td>
<td>2006</td>
<td>8</td>
<td>Zha et al. [2009]</td>
</tr>
<tr>
<td>CA-SJ3</td>
<td>ENF</td>
<td>BERMS - Jack Pine, 1995 Harvest</td>
<td>-104.6</td>
<td>53.9</td>
<td>2005</td>
<td>2006</td>
<td>7</td>
<td>Zha et al. [2009]</td>
</tr>
<tr>
<td>CA-TP4</td>
<td>ENF</td>
<td>Turkey Point - Mature</td>
<td>-80.4</td>
<td>42.7</td>
<td>2002</td>
<td>2007</td>
<td>17</td>
<td>Arain and Restrepo [2005]; Peichl and Arain [2006]</td>
</tr>
</tbody>
</table>

CA-WP1   | WET  | Western Peatland - LaBiche River | -112.5 | 55.0 | 2003 | 2007 | 14 | Flanagan and Syed [2011] |

US-ARM   | CRO  | ARM - Southern Great Plains | -97.5 | 36.6 | 2000 | 2007 | 17 | Fischer et al. [2007] |


US-Ha1   | DBF  | Harvard Forest - EMS Tower | -72.2 | 42.5 | 1991 | 2006 | 23 | Urbanski et al. [2007] |

US-Ho1   | ENF  | Howland Forest - Main Tower | -68.7 | 45.2 | 1996 | 2004 | 23 | Richardson et al. [2009] |

US-IB1   | CRO  | Fermi Lab - agriculture | -88.2 | 41.9 | 2005 | 2007 | 16 | Post et al. [2004] |

US-IB2   | GRA  | Fermi - Prairie | -88.2 | 41.8 | 2004 | 2007 | 17 | Post et al. [2004] |

US-Los   | WET  | Lost Creek | -90.0 | 46.1 | 2000 | 2006 | 12 | Sulman et al. [2009] |

US-Me2   | ENF  | Metolius - Intermediate-aged Ponderosa Pine | -121.6 | 44.5 | 2002 | 2007 | 20 | Thomas et al. [2009] |

US-Me3   | ENF  | Metolius - Ponderosa Pine, young (2) | -121.6 | 44.3 | 2004 | 2005 | 8 | Vickers et al. [2009] |

US-Me5   | ENF  | Metolius - Ponderosa Pine, young (1) | -121.6 | 44.4 | 1999 | 2002 | 8 | Law et al. [2001] |

<table>
<thead>
<tr>
<th>Region</th>
<th>Biome</th>
<th>Site</th>
<th>Climate</th>
<th>Soil</th>
<th>Layers</th>
<th>Phenology</th>
<th>Nitrogen Cycle</th>
<th>GPP Model</th>
<th>Leaf-to-Canopy</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>US-MOz</td>
<td>DBF</td>
<td>Missouri Ozark</td>
<td>-92.2</td>
<td>38.7</td>
<td>2004 - 2007</td>
<td>17</td>
<td>Gu et al. [2006]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>US-Ne1</td>
<td>CRO</td>
<td>Mead - Irrigated maize</td>
<td>-96.5</td>
<td>41.2</td>
<td>2001 - 2006</td>
<td>17</td>
<td>Verma et al. [2005]; Suyker and Verma [2010]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>US-Ne2</td>
<td>CRO</td>
<td>Mead - Irrigated maize/soybean</td>
<td>-96.5</td>
<td>41.2</td>
<td>2001 - 2006</td>
<td>17</td>
<td>Verma et al. [2005]; Suyker and Verma [2010]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>US-Ne3</td>
<td>CRO</td>
<td>Mead - Rainfed</td>
<td>-96.4</td>
<td>41.2</td>
<td>2001 - 2006</td>
<td>23</td>
<td>Verma et al. [2005]; Suyker and Verma [2010]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>US-NR1</td>
<td>ENF</td>
<td>Niwot Ridge</td>
<td>-105.5</td>
<td>40.0</td>
<td>1998 - 2007</td>
<td>17</td>
<td>Monson et al. [2002]; Monson et al. [2005]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>US-PF2</td>
<td>MF</td>
<td>Park Falls / WLEF</td>
<td>-90.3</td>
<td>45.9</td>
<td>1995 - 2005</td>
<td>14</td>
<td>Davis et al. [2003]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>US-Sh2</td>
<td>GRA</td>
<td>Shidler</td>
<td>-96.7</td>
<td>36.9</td>
<td>1997 - 2000</td>
<td>15</td>
<td>Suyker et al. [2003]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>US-Sy2</td>
<td>MF</td>
<td>Sylvania Wilderness Area</td>
<td>-89.3</td>
<td>46.2</td>
<td>2001 - 2006</td>
<td>15</td>
<td>Desai et al. [2005]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>US-Ton</td>
<td>WSA</td>
<td>Tonzi Ranch</td>
<td>-121.0</td>
<td>38.4</td>
<td>2001 - 2007</td>
<td>15</td>
<td>Ma et al. [2007]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>US-Umb</td>
<td>DBF</td>
<td>UMBS</td>
<td>-84.7</td>
<td>45.6</td>
<td>1999 - 2006</td>
<td>22</td>
<td>Gough et al. [2008]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>US-Var</td>
<td>GRA</td>
<td>Vaira Ranch</td>
<td>-121.0</td>
<td>38.4</td>
<td>2000 - 2007</td>
<td>17</td>
<td>Ma et al. [2007]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>US-WCr</td>
<td>DBF</td>
<td>Willow Creek</td>
<td>-90.1</td>
<td>45.8</td>
<td>1998 - 2006</td>
<td>16</td>
<td>Cook et al. [2004]</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1 IGBP biome class definitions: CRO (Croplands), CSH (Closed Shrublands), DBF (Deciduous Broadleaf Forest), ENF (Evergreen Needleleaf Forest), GRA (Grasslands), MF (Mixed Forests), WET (Permanent Wetlands), WSA (Woody Savannas).

Table 2: Summary of model characteristics

<table>
<thead>
<tr>
<th>Model</th>
<th>Number Sites</th>
<th>Time Step</th>
<th>Soil Layers</th>
<th>Phenology</th>
<th>Nitrogen Cycle</th>
<th>GPP Model</th>
<th>Leaf-to-Canopy</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>AgroIBIS</td>
<td>5</td>
<td>Hourly</td>
<td>11</td>
<td>Prognostic</td>
<td>Yes</td>
<td>EK</td>
<td>Big-Leaf</td>
<td>Kucharik and Twine [2007]</td>
</tr>
<tr>
<td>Model</td>
<td>Frequency</td>
<td>Soil Layers</td>
<td>Prognostic</td>
<td>LUE</td>
<td>Sub-model</td>
<td>Reference</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------</td>
<td>-----------</td>
<td>-------------</td>
<td>------------</td>
<td>-----</td>
<td>-----------</td>
<td>----------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BEPS</td>
<td>Daily</td>
<td>3</td>
<td>Semi-prognostic</td>
<td>Yes</td>
<td>EK</td>
<td>2-Leaf</td>
<td>Liu et al. [1999]</td>
<td></td>
</tr>
<tr>
<td>Biome-BGC</td>
<td>Daily</td>
<td>1</td>
<td>Prognostic</td>
<td>Yes</td>
<td>EK</td>
<td>2-Leaf</td>
<td>Thornton et al. [2005]</td>
<td></td>
</tr>
<tr>
<td>Can-IBIS</td>
<td>Hourly</td>
<td>7</td>
<td>Prognostic</td>
<td>Yes</td>
<td>EK</td>
<td>2-Leaf</td>
<td>Liu et al. [2005]</td>
<td></td>
</tr>
<tr>
<td>CN-CLASS</td>
<td>Hourly</td>
<td>3</td>
<td>Prognostic</td>
<td>Yes</td>
<td>EK</td>
<td>2-Leaf</td>
<td>Arain et al. [2006]</td>
<td></td>
</tr>
<tr>
<td>DLEM</td>
<td>Daily</td>
<td>2</td>
<td>Semi-prognostic</td>
<td>Yes</td>
<td>EK</td>
<td>2-Leaf</td>
<td>Tian et al. [2010]</td>
<td></td>
</tr>
<tr>
<td>DNDC</td>
<td>Daily</td>
<td>10</td>
<td>Prognostic</td>
<td>Yes</td>
<td>LUE</td>
<td>Big-Leaf</td>
<td>Li et al. [2010]</td>
<td></td>
</tr>
<tr>
<td>Ecosys</td>
<td>Hourly</td>
<td>15</td>
<td>Prognostic</td>
<td>Yes</td>
<td>EK</td>
<td>2-Leaf</td>
<td>Grant et al. [2009]</td>
<td></td>
</tr>
<tr>
<td>ED2</td>
<td>Hourly</td>
<td>9</td>
<td>Prognostic</td>
<td>Yes</td>
<td>EK</td>
<td>2-Leaf</td>
<td>Medvigy et al. [2009]</td>
<td></td>
</tr>
<tr>
<td>EDCM</td>
<td>Daily</td>
<td>10</td>
<td>Prognostic</td>
<td>Yes</td>
<td>LUE</td>
<td>Big-Leaf</td>
<td>Liu et al. [2003]</td>
<td></td>
</tr>
<tr>
<td>ISAM</td>
<td>Hourly</td>
<td>10</td>
<td>Prognostic</td>
<td>Yes</td>
<td>LUE</td>
<td>Big-Leaf</td>
<td>Yang et al. [2009]</td>
<td></td>
</tr>
<tr>
<td>ISOLSM</td>
<td>Hourly</td>
<td>20</td>
<td>Observed</td>
<td>No</td>
<td>EK</td>
<td>2-Leaf</td>
<td>Riley et al. [2002]</td>
<td></td>
</tr>
<tr>
<td>LoTEC</td>
<td>Hourly</td>
<td>14</td>
<td>Prognostic</td>
<td>No</td>
<td>EK</td>
<td>Big-Leaf</td>
<td>Hanson et al. [2004]</td>
<td></td>
</tr>
<tr>
<td>LPJ</td>
<td>Daily</td>
<td>2</td>
<td>Prognostic</td>
<td>No</td>
<td>EK</td>
<td>Big-Leaf</td>
<td>Sitch et al. [2003]</td>
<td></td>
</tr>
<tr>
<td>MODIS_5.0</td>
<td>Daily</td>
<td>0</td>
<td>Observed</td>
<td>No</td>
<td>LUE</td>
<td>Big-Leaf</td>
<td>Heinsch et al. [2003]</td>
<td></td>
</tr>
<tr>
<td>MODIS_5.1</td>
<td>Daily</td>
<td>0</td>
<td>Observed</td>
<td>No</td>
<td>LUE</td>
<td>Big-Leaf</td>
<td>Heinsch et al. [2003]</td>
<td></td>
</tr>
<tr>
<td>MODIS_alg</td>
<td>Daily</td>
<td>0</td>
<td>Observed</td>
<td>No</td>
<td>LUE</td>
<td>Big-Leaf</td>
<td>Heinsch et al. [2003]</td>
<td></td>
</tr>
<tr>
<td>ORCHIDEE</td>
<td>Hourly</td>
<td>2</td>
<td>Prognostic</td>
<td>No</td>
<td>EK</td>
<td>Big-Leaf</td>
<td>Krinner et al. [2005]</td>
<td></td>
</tr>
<tr>
<td>SiB3</td>
<td>Hourly</td>
<td>10</td>
<td>Observed</td>
<td>No</td>
<td>EK</td>
<td>Big-Leaf</td>
<td>Baker et al. [2008]</td>
<td></td>
</tr>
<tr>
<td>SiBCASA</td>
<td>Hourly</td>
<td>25</td>
<td>Semi-prognostic</td>
<td>No</td>
<td>EK</td>
<td>Big-Leaf</td>
<td>Schaefer et al. [2009]</td>
<td></td>
</tr>
<tr>
<td>SiBcrop</td>
<td>Hourly</td>
<td>10</td>
<td>Prognostic</td>
<td>Yes</td>
<td>EK</td>
<td>Big-Leaf</td>
<td>Lokupitiya et al. [2009]</td>
<td></td>
</tr>
<tr>
<td>SSIB2</td>
<td>Hourly</td>
<td>3</td>
<td>Observed</td>
<td>No</td>
<td>EK</td>
<td>Big-Leaf</td>
<td>Zhan et al. [2003]</td>
<td></td>
</tr>
<tr>
<td>TECO</td>
<td>Hourly</td>
<td>10</td>
<td>Prognostic</td>
<td>No</td>
<td>EK</td>
<td>2-Leaf</td>
<td>Weng and Luo [2008]</td>
<td></td>
</tr>
<tr>
<td>TRIPLEX</td>
<td>Daily</td>
<td>0</td>
<td>Observed</td>
<td>Yes</td>
<td>LUE</td>
<td>Big-Leaf</td>
<td>Peng et al. [2002]</td>
<td></td>
</tr>
</tbody>
</table>

1 Zero soil layers indicate the model does not have a prognostic sub-model for soil temperature and moisture.
2 Observed phenology means the model uses remote sensing data to determine Leaf Area Index and GPP. Semi-prognostic means that remote sensing data is used to specify either LAI or GPP, but not both.

3 GPP Model types: EK (Enzyme Kinetic) and LUE (Light Use Efficiency)

**Table 3:** Differences in Performance based on Model Structural Characteristics (value ± standard deviation) for all 627 simulations.

<table>
<thead>
<tr>
<th>Char</th>
<th>Value</th>
<th>chisqr (-)</th>
<th>RMSE ($\mu$mol m$^{-2}$ s$^{-1}$)</th>
<th>bias ($\mu$mol m$^{-2}$ s$^{-1}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPP Model Type</td>
<td>EK</td>
<td>4.2±3.9</td>
<td>2.4±1.2</td>
<td>0.1±1.2</td>
</tr>
<tr>
<td>GPP Model Type</td>
<td>LUE</td>
<td>3.2±2.3</td>
<td>2.8±1.7</td>
<td>-0.9±1.3</td>
</tr>
<tr>
<td>Leaf-to-Canopy</td>
<td>Big-leaf</td>
<td>3.6±3.2</td>
<td>2.6±1.5</td>
<td>-0.2±1.4</td>
</tr>
<tr>
<td>Leaf-to-Canopy</td>
<td>2-leaf</td>
<td>4.3±3.8</td>
<td>2.4±1.2</td>
<td>-0.1±1.3</td>
</tr>
<tr>
<td>Phenology</td>
<td>Observed</td>
<td>3.1±2.5</td>
<td>2.3±1.3</td>
<td>-0.3±1.2</td>
</tr>
<tr>
<td>Phenology</td>
<td>Prognostic</td>
<td>4.8±4.1</td>
<td>2.7±1.4</td>
<td>-0.1±1.4</td>
</tr>
<tr>
<td>Soil Model</td>
<td>No</td>
<td>3.3±2.6</td>
<td>2.6±1.7</td>
<td>-0.6±1.4</td>
</tr>
<tr>
<td>Soil Model</td>
<td>Yes</td>
<td>4.2±3.8</td>
<td>2.4±1.2</td>
<td>0.0±1.3</td>
</tr>
<tr>
<td>Nitrogen Model</td>
<td>No</td>
<td>4.1±3.6</td>
<td>2.5±1.4</td>
<td>-0.1±1.3</td>
</tr>
<tr>
<td>Nitrogen Model</td>
<td>Yes</td>
<td>3.7±3.3</td>
<td>2.5±1.4</td>
<td>-0.3±1.3</td>
</tr>
<tr>
<td>Time Step</td>
<td>Daily</td>
<td>3.5±3.0</td>
<td>2.7±1.6</td>
<td>-0.7±1.3</td>
</tr>
<tr>
<td>Time Step</td>
<td>Hourly</td>
<td>4.3±3.8</td>
<td>2.4±1.2</td>
<td>0.2±1.2</td>
</tr>
</tbody>
</table>