A model-data comparison of gross primary productivity: Results from the North American Carbon Program site synthesis


Received 20 January 2012; revised 16 May 2012; accepted 26 May 2012; published 19 July 2012.

[1] 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, and underestimate GPP in summer. Models overpredicted 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 $\epsilon_{\text{max}}$ (LUE), $V_{\text{cmax}}$ (unstressed Rubisco catalytic capacity) or $J_{\text{max}}$ (the maximum electron transport rate).

1. Introduction

[2] 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.

[3] 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.

[4] 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; 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 versus Ek models. To test our hypotheses, we compared simulated GPP from 26 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.

\section{Methods}

\subsection{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. GPP 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., 2000]. 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 (A. G. Barr et al., Use of change-point detection for u*—threshold evaluation for the North American Carbon Program interim synthesis, manuscript in preparation, 2012). 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 and 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 R_{eco} 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. Last, 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).

[10] 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.
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>BEPS</td>
<td>10</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>
</tr>
<tr>
<td>Biome-BGC</td>
<td>33</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>
</tr>
<tr>
<td>Can-IBIS</td>
<td>24</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>
</tr>
<tr>
<td>CN-CLASS</td>
<td>28</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>
</tr>
<tr>
<td>DLEM</td>
<td>30</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>
</tr>
<tr>
<td>DNDC</td>
<td>5</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>
</tr>
<tr>
<td>Ecosys</td>
<td>35</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>
</tr>
<tr>
<td>ED2</td>
<td>24</td>
<td>Hourly</td>
<td>9</td>
<td>Prognostic</td>
<td>Yes</td>
<td>EK</td>
<td>2-Leaf</td>
<td>Modvig et al. [2009]</td>
</tr>
<tr>
<td>EDM</td>
<td>9</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>
</tr>
<tr>
<td>ESDM</td>
<td>13</td>
<td>Hourly</td>
<td>10</td>
<td>Prognostic</td>
<td>Yes</td>
<td>LUE</td>
<td>2-Leaf</td>
<td>Yang et al. [2009]</td>
</tr>
<tr>
<td>ISOLSM</td>
<td>9</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>
</tr>
<tr>
<td>Lsp-TEC</td>
<td>10</td>
<td>Hourly</td>
<td>14</td>
<td>Prognostic</td>
<td>No</td>
<td>EK</td>
<td>Big-Leaf</td>
<td>Hanson et al. [2008]</td>
</tr>
<tr>
<td>LPj</td>
<td>26</td>
<td>Daily</td>
<td>2</td>
<td>Prognostic</td>
<td>No</td>
<td>EK</td>
<td>Big-Leaf</td>
<td>Stich et al. [2003]</td>
</tr>
<tr>
<td>MODIS 5.0</td>
<td>38</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>
</tr>
<tr>
<td>MODIS 5.1</td>
<td>37</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>
</tr>
<tr>
<td>MODIS alg</td>
<td>39</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>
</tr>
<tr>
<td>ORCHIDEE</td>
<td>32</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>
</tr>
<tr>
<td>SiB3</td>
<td>28</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>
</tr>
<tr>
<td>SiBCASA</td>
<td>32</td>
<td>Hourly</td>
<td>25</td>
<td>Semi-prognostic</td>
<td>No</td>
<td>EK</td>
<td>Big-Leaf</td>
<td>Schaef er et al. [2009]</td>
</tr>
<tr>
<td>SiBroop</td>
<td>5</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>
</tr>
<tr>
<td>SISB2</td>
<td>39</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>
</tr>
<tr>
<td>TECO</td>
<td>32</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>
</tr>
<tr>
<td>TRIPLEX</td>
<td>6</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>
</tr>
</tbody>
</table>

*a*Zero soil layers indicate the model does not have a prognostic submodel for soil temperature and moisture.

*b*Observed phenology means the model uses remote sensing data to determine leaf area index (LAI) and gross primary productivity (GPP). Semi-prognostic means that remote sensing data is used to specify either LAI or GPP, but not both.

*GPP model types: EK (enzyme kinetic) and LUE (light use efficiency).*

[Richardson and Hollinger, 2007; A. G. Barr et al., manuscript in preparation, 2012]. A. G. Barr et al. 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 ~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

[12] 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.

[12] 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 [Distributed Active Archive Center for Biogeochemical Dynamics, 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.

[13] 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]:

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

where \(\varepsilon_{\text{max}}\) is the maximum light use efficiency, \(F_{\text{SW}}\) 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. \(S_{\text{PPD}}\) represents the GPP response to drought and humidity stress and \(S_{T}\) represents the GPP response to temperature, with both varying between 0 and 1.
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. [14] 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_{eco} balances GPP and the average NEE over the entire simulation is near zero. The steady state assumption influences R_{eco} 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

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

\[ r_i = ( \text{GPP}_{si} - \text{GPP}_{oi} ), \]

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

\[ X^2 = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{r_i}{e_i} \right)^2, \]

where \( n \) is the number of residuals and \( e_i \) is the uncertainty for the \( i^{th} \) daily average estimated GPP. We filtered out \( \sim 0.1\% \) of daily estimated GPP values with \( e_i \leq 0.3 \text{ \mu mol m}^{-2} \text{s}^{-1} \) which produced extreme outlier \( X^2 \) values that skewed our results. Such unrealistically small \( e_i \) 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 \( X^2 \).

[16] The ideal target for any model is \( X^2 < 1.0 \), which means, on average, the residuals are less than uncertainty or the model matches observations within measurement uncertainty. Variations of \( X^2 \) within this target range have no meaning relative to model performance. A model with an \( X^2 \) value of 0.8, for example, is not “better” than a model with an \( X^2 \) value of 0.9, since both models show no statistically significant differences with observations. Consequently, we identified performance categories based on ranges of \( X^2 \) values. An \( X^2 \) value of \( \leq 1.0 \) indicated good model performance. An \( X^2 \) value between 1.0 and 2.0 indicated marginal model performance, where the model-data mismatch is on the order of two times the observation uncertainty. An \( X^2 \) value of \( > 2.0 \) indicated poor model performance, where the model-data mismatch is several times the observation uncertainty. An \( X^2 \) value of 9, for example, indicates that the model-data mismatch is, on average, three times the uncertainty.

[17] 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.

[18] 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. Results

3.1. Performance Summary

[19] None of the models in this study achieved a good overall performance of \( X^2 \) 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 \( X^2 \) values due to large biases. Generally speaking, higher RMSE resulted in higher \( X^2 \), but the
relationship was weak because $\chi^2$ accounts for uncertainty and bias 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 1.0 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. [2012] 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 \( \chi^2 > 20 \).

### 3.2. 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 versus 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. Sprintsin 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.

3.3. 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

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Value</th>
<th>$X^2$</th>
<th>Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPP Model Type</td>
<td>EK</td>
<td>$4.2 \pm 3.9$</td>
<td>$2.4 \pm 1.2$</td>
</tr>
<tr>
<td>GPP Model Type</td>
<td>LUE</td>
<td>$3.2 \pm 2.3$</td>
<td>$2.8 \pm 1.7$</td>
</tr>
<tr>
<td>Leaf-to-Canopy</td>
<td>Big-Leaf</td>
<td>$3.7 \pm 3.2$</td>
<td>$2.6 \pm 1.5$</td>
</tr>
<tr>
<td>Leaf-to-Canopy</td>
<td>2-leaf</td>
<td>$4.2 \pm 3.8$</td>
<td>$2.4 \pm 1.2$</td>
</tr>
<tr>
<td>Phenology</td>
<td>Observed</td>
<td>$3.1 \pm 2.5$</td>
<td>$2.3 \pm 1.3$</td>
</tr>
<tr>
<td>Phenology</td>
<td>Prognostic</td>
<td>$4.8 \pm 4.1$</td>
<td>$2.7 \pm 1.4$</td>
</tr>
<tr>
<td>Soil Model</td>
<td>No</td>
<td>$3.3 \pm 2.6$</td>
<td>$2.6 \pm 1.7$</td>
</tr>
<tr>
<td>Soil Model</td>
<td>Yes</td>
<td>$4.2 \pm 3.8$</td>
<td>$2.4 \pm 1.2$</td>
</tr>
<tr>
<td>Nitrogen Model</td>
<td>No</td>
<td>$4.1 \pm 3.6$</td>
<td>$2.5 \pm 1.4$</td>
</tr>
<tr>
<td>Nitrogen Model</td>
<td>Yes</td>
<td>$3.7 \pm 3.3$</td>
<td>$2.5 \pm 1.4$</td>
</tr>
<tr>
<td>Time Step</td>
<td>Daily</td>
<td>$3.5 \pm 3.0$</td>
<td>$2.7 \pm 1.6$</td>
</tr>
<tr>
<td>Time Step</td>
<td>Hourly</td>
<td>$4.3 \pm 3.8$</td>
<td>$2.4 \pm 1.2$</td>
</tr>
</tbody>
</table>
GPP for the entire canopy. A model assumes the distribution scaling factor represents the ratio of GPP for a single leaf to values for a single leaf at the canopy top. The leaf-to-canopy potential GPP parameters typically represent peak or optimal canopy also influences the simulated LUE. The maximum value of ε
\text{max}
, so a better value of ε
\text{max}
 will improve performance. For other models, V
\text{cmax}
 (the unstressed Rubisco carboxylation rate), α (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
\text{cmax}
, J
\text{max}
, and α compiled from many studies 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
\text{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
\text{cmax}
 in a way that will work for all models will require new theoretical development.
G03010

SCHAEFER ET AL.: A MODEL-DATA COMPARISON OF GPP

Figure 7. Normalized humidity response curves for (a) US-Var and (b) US-Ha1 showing simulated and observed GPP as a function of relative humidity. The gray bar indicates uncertainty in the observed GPP response. All response curves are normalized such that the maximum value is 1.

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 overpredicted 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 [Schafer 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, overpredicting 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.

3.5. 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 $20 \pm 5°C$. 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 $-6 \pm 3°C$. 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...
improve simulated phenology. Most models use an exponential or 
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 deter-
mined 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.

[41] The positive bias in winter, spring, and fall resulted 
from the fact that the $Q_{10}$ function alone will never reach zero 
nmatter 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.

4. Conclusions

[42] 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 mar-
ginal performance, matching estimated GPP within roughly 
two times the uncertainty. Our first hypothesis proved false: 
we found no statistically significant differences in perform-
ance 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 
ot 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.

[43] The poor overall model performance resulted pri-
marily 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), 
$V_{\text{cm}}$ (unstressed Rubisco catalytic capacity) or $J_{\text{max}}$ (the 
maximum electron transport rate). The temperature, humid-
ity, and drought scaling factors determined temporal vari-
ability 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 overpredicted 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 Q10 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.

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; funding was also provided by the VIIth International Congress on Photosynthesis, in Domains of Photosynthesis, in Domains of Photosynthesis.

References


Distributed Active Archive Center for Biogeochemical Dynamics (2010), AmeriFlux-Canada synthesis, AF l u x n e t - C a n a d a s y n t h e s i s , 148, 2022, doi:10.1890/06-0922.1.


