Processes influencing model-data mismatch in drought-stressed, fire-disturbed eddy flux sites

Stephen Mitchell, Keith Beven, Jim Freer, and Beverly Law

Received 15 September 2009; revised 12 November 2010; accepted 14 January 2011; published 7 May 2011.

Semiarid forests are very sensitive to climatic change and among the most difficult ecosystems to accurately model. We tested the performance of the Biome-BGC model against eddy flux data taken from young (years 2004–2008), mature (years 2002–2008), and old-growth (year 2000) ponderosa pine stands at Metolius, Oregon, and subsequently examined several potential causes for model-data mismatch. We used the Generalized Likelihood Uncertainty Estimation methodology, which involved 500,000 model runs for each stand (1,500,000 total). Each simulation was run with randomly generated parameter values from a uniform distribution based on published parameter ranges, resulting in modeled estimates of net ecosystem CO2 exchange (NEE) that were compared to measured eddy flux data. Simulations for the young stand exhibited the highest level of performance, though they overestimated ecosystem C accumulation (−NEE) 99% of the time. Among the simulations for the mature and old-growth stands, 100% and 99% of the simulations underestimated ecosystem C accumulation. One obvious area of model-data mismatch is soil moisture, which was overestimated by the model in the young and old-growth stands yet underestimated in the mature stand. However, modeled estimates of soil water content and associated water deficits did not appear to be the primary cause of model-data mismatch; our analysis indicated that gross primary production can be accurately modeled even if soil moisture content is not. Instead, difficulties in adequately modeling ecosystem respiration, mainly autotrophic respiration, appeared to be the fundamental cause of model-data mismatch.


1. Introduction

A predictive science of atmosphere-biosphere interactions is essential for an understanding of how climatic change will alter earth’s ecosystems and the continued provision of ecosystem services, but this has not yet been achieved [Moorcroft, 2006]. As a result, the future of the global C cycle is a matter of ongoing debate, and a disagreement remains about whether earth’s surface will become a source or sink of atmospheric CO2 in the coming century [Cox et al., 2000; Friedlingstein et al., 2003; Moorcroft, 2006; Sitch et al., 2008]. Fortunately, long-term studies of net ecosystem CO2 exchange throughout a wide variety of biomes [Baldocchi, 2008] hold significant potential for understanding terrestrial ecosystems and their role in mitigating or exacerbating current and future atmospheric CO2 concentrations.

Net ecosystem CO2 exchange, hereafter referred to as NEE, is the net exchange of CO2 between ecosystems and the atmosphere, calculated as the difference between gross primary production and ecosystem respiration, excluding losses of respiration-derived dissolved inorganic carbon [Chapin et al., 2006]. Note that our use of NEE, based on Chapin et al. [2006], is almost identical to net ecosystem production, NEP, except with signs reversed (−NEE = NEP), where a negative NEE represents a transfer of atmospheric C to a terrestrial ecosystem. Continuous field measurements of NEE have been taken from over 400 locations using the eddy covariance method and offer a valuable baseline against which model assumptions, parameters, and performance can be ascertained [Schulz et al., 2001; Wang et al., 2001; Thornton et al., 2002; Braswell et al., 2005; Knorr and Kattge, 2005; Sacks et al., 2006]. Employing NEE as a metric against which terrestrial ecosystem model performance can be tested is useful for constraining estimates of C cycling, and not just because it gives an estimate of net gain (or loss) of C for a terrestrial ecosystem. NEE is, however, a derived variable, dependent upon uncertainty in the estimation of other metrics of the C cycle such as gross
primary production, autotrophic respiration, and heterotrophic respiration. Potential sources of model failure, therefore, are numerous and interdependent, as errors produced in one step almost inevitably lead to a propagation of errors carried into a succeeding step.

[1] One way to ascertain the extent of these uncertainties is through a model-data synthesis. Model-data synthesis, according to Raupach et al. [2005], operates under the assumption that the inherent uncertainties in any data set are just as important as the data values and should thereby be included in both parameter estimation and data assimilation. For an ecosystem model (or any environmental model), this uncertainty lies not just with the observed data but also with the parameters on which the data are conditioned, affecting both the predictive uncertainty of a model-data synthesis and the predicted best estimate. An effective analysis of these uncertainties requires an acknowledgment of the potential for equivaliney in model predictions. The concept of equivaliney implies that, within the current capacity of mechanistic modeling, there may be many model structures and different parameter sets for a particular ecosystem process, representing different hypotheses about system functioning, that may be acceptable in reproducing the observed behavior of an environmental system [Beven, 2002, 2006]. Given the limitations of observed data to fully represent the natural system, the approximate process descriptions within models, the temporal propagation of model/data errors, and lack of independent estimates of the effective parameter values by a model, it may not be possible to determine the most plausible hypothesis. However, by exploring a range of model predictions, parameter combinations that fail to predict observed responses adequately can be rejected and hence classed as unacceptable hypotheses.

[5] Semiarid forest ecosystems are a particularly interesting case for testing hypotheses about system functioning and pose a difficult challenge for the production of accurate ecosystem simulations. Fluctuations in above or belowground biomass, physiological characteristics of vegetation, or community composition that may be insignificant in more mesic, more productive ecosystems, can have substantial feedbacks that affect water balances and other ecosystem processes [Kremer and Running, 1996]. Additionally, minor variations in precipitation, potential evapotranspiration, and other environmental factors may significantly alter community phenology and productivity [Kremer and Running, 1996]. Even though they do not represent a particularly large carbon pool compared to more mesic ecosystems, they are nevertheless significant in global change studies because of their vulnerability to a changing climate, particularly drought [Fischlin et al., 2007]; extended water deficits in ponderosa pine woodlands in New Mexico have been observed to result in widespread tree mortality, producing landscape-level shifts to pinyon-Juniper woodland [Allen and Breshears, 1998]. Further drought in this region has even resulted in a widespread mortality of Pinyon pine, a species that is already highly drought-resistant [Breshears et al., 2005].

[6] Research at the Metolius sites indicates that the old-growth site is less sensitive to drought, and the mature and young pine sites experience seasonal water stress beginning in June/July that triggers a rapid decline in net carbon uptake associate with gross primary production (GPP) leveling off while total ecosystem respiration (TER) continues to increase. GPP responds to the onset of seasonal drought sooner than TER at these sites. Seasonal and diel analysis of GPP, water use efficiency and transpiration indicates that the young site experiences seasonal drought earlier and to a more severe degree, presumably because the younger stand has a shallower root system [Thomas et al., 2009; Vickers et al., 2009].

[7] The accurate simulation of such coupled climate-ecosystem changes is challenging and will depend on a number of different factors: the forcing data used to drive an ecosystem model (whether they come from observations or climate predictions), the initial conditions and disturbance history at the site, the limitations of the process representations in a model based on current understanding, the effective parameter values provided to the model, and the limitations of any observational data used to evaluate the model. These are all sources of uncertainty in simulations and their influence will depend on the ecosystem type and the spatiotemporal scale of the simulation.

[8] Our goals in this study were twofold. First, we analyzed the extent to which equifinality in parameter values can influence the reproduction of ecosystem processes in a widely used terrestrial ecosystem model. We examined this by performing multiple simulations (500,000) with multiple combinations of ecophysiological parameter values within plausible parameter ranges for a young (~20 years old), mature (~90 years old) and old-growth stand (>250 years old). Second, we analyze these results to determine the key ecosystem processes that are responsible for any model-data mismatch in these semiarid forests over multiple years of data.

2. Materials and Methods

2.1. The GLUE Methodology

[9] The Generalized Likelihood Uncertainty Estimation (GLUE) methodology [Beven and Binley, 1992] was developed from the generalized sensitivity analysis of Spear and Hornberger [1980] to assess the possibility that more than one unique parameter set for a given environmental model may provide good (behavioral) predictions. Studies of parameter responses have shown that the assumption of a single well-defined optimal parameter set rarely holds, resulting in model equifinality, where many parameter set combinations give similar predictions of observed responses [Freer et al., 1996; Franks et al., 1997; Beven and Freer, 2001; Schulz et al., 2001; Beven, 2006]. GLUE provides a means of assessing the predictive uncertainty of a model by comparing simulations to available observations using a performance measure within a Monte Carlo framework. GLUE has been used for a wide range of environmental modeling problems [see Beven and Freer, 2001; Freer et al., 2004; Beven, 2006], including the prediction of CO2 flux data [Franks et al., 1997; Schulz et al., 2001], tree mortality under drought conditions [Martinez-Vilalta et al., 2002] and forest fires [Piñol et al., 2004, 2007]. Application of the GLUE method involves a large number (typically >100,000) of model simulations, each of which is characterized by a set of randomly sampled parameters usually drawn from uniform (noninformative) prior distributions across the range of each parameter since information regarding the likely
effective parameter values is normally highly limited or unknown (Figure 1). The performance of each run is thereafter deemed behavioral or nonbehavioral based upon the comparison of simulated versus observed data. Model runs that do not meet specified acceptability criteria are rejected as nonbehavioral and are thus given zero likelihood, removing them from further analysis [Beven, 2006]. Within GLUE, a model as hypothesis is always associated with a set of parameter values. Retained parameter sets can still exhibit a wide spread of values and the associated simulated responses can be quite variable between parameter sets over the simulated period. The sensitivity of model parameters and the interactions between the parameters determines the variability in the posterior likelihoods of the parameter sets.

2.2. Evaluation of Parameter Sensitivities

In a complex model such as Biome-BGC it is inevitable that many randomly generated parameter sets will result in a simulation that is physiologically unsustainable, thus simulations that had an annual GPP of less than 10 g C m\(^{-2}\) yr\(^{-1}\) were excluded from additional analysis a priori. Once we screened parameter combinations that resulted in nonliving simulations, we calculated the resultant likelihood weights for the remaining simulations using the following equation:

\[
L(\Theta|Y) = \frac{\sum_{j=1}^{J} \left( \sum_{i=1}^{I} (C_o^i - C_m^i)^2 \right)}{\left( \sum_{i=1}^{I} (C_o^i - C_m^i)^2 \right)}
\]  

where \(C_o^i\) is observed NEE for day \(i\), and \(C_m^i\) is modeled NEE of day \(i\) for parameter set \(j\), and \(L(\Theta|Y)\) is the likelihood of simulating data \(Y\) given parameter set \(\Theta\). Note that we do not consider simulations that are merely "living" to be indicative of satisfactory model performance. Until the application of an additional performance criteria, we only consider them to be worthy of further analysis.

Our method is similar to the sensitivity analysis of Spear and Hornberger [1980] except that there is an additional step of calculating a likelihood weight for each parameter set. Equation (1) is calculated on a daily time step for the extent of the available data, which in this study is composed of carbon flux data measured by the eddy covariance technique for the years simulated (2004–2008 for the young stand, 2002–2008 for the mature stand, and 2000 for the old-growth stand).

We imposed an additional criterion for what would be considered a satisfactory model performance in relation
to inherent errors in the observed data. We did this by comparing annual estimates of NEE from the model to estimates obtained from field measurements. We used the coefficient of variation (c_i) to represent the uncertainty in annual estimates of NEE. The term c_i (calculated as σ/μ) is based on the coefficient of variation calculated for annual NEE estimates and includes estimates of errors incurred by the instrumentation used in the eddy covariance technique, gap-filling, as well as spatial and temporal variability [Oren et al., 2006; Thomas et al., 2009]. This “effective observation error” criterion is an example of the approach to model evaluation proposed by Beven [2006] but the concept of including observational uncertainties to construct model performance criteria has been applied previously [Page et al., 2003; Freer et al., 2004]. Mean annual NEE is calculated for the years during which the model-data synthesis was performed (2004–2008 for the young stand, 2002–2008 for the mature stand, and 2000 for the old-growth stand). Modeled estimates of annual NEE for those years are deemed to be satisfactory if they fall within the bounds of uncertainty (mean annual NEE ± c_i).

2.3. Parameter Estimates

[Biome-BGC requires 37 ecophysiological parameter values for the simulation of evergreen needleleaf forests. Of these 37, 13 were allowed to vary (Table 1) assuming independent uniform prior distributions across feasible ranges of the parameters in the absence of any strong information about effective parameter values and their covariation. White et al. [2000] performed a sensitivity analysis of model parameters, showing that LAI, FLNR, and C:N_leaf were among the most sensitive parameters, and these were some of the parameters we included. Additional selection of parameters that were allowed to vary was based on the range in variation of parameters in the literature. For instance, parameters with a wide range of variability, such as FRC:LC, were chosen for this reason, and literature values that exhibited little to no variability were likewise excluded. Our study site has a canopy comprised primarily of *Pinus ponderosa*, a species with input parameters that are generally provided in the compilation by White et al. [2000], many of which were also available from site measurements. Each parameter range was subsequently expanded to allow for the possibility of yet-unpublished values that may be present in the field. However, it must also be stressed that field observations may not be commensurate with the effective parameter values required in a particular model structure to give a good fit to the observed carbon flux data.]

### 2.4. Study Sites

[The Metolius Research National Area is located approximately 64 km north of Bend, Oregon. Net ecosystem CO2 exchange data for the young stand at Metolius were collected from 2004 to 2008. Data from the mature stand at Metolius were collected 2002–2008, and data for the old-growth stand were collected in 2000. The young stand is recovering from a clear-cut performed in 1987 while the mature stand is recovering from a harvest in ~1905. All sites are dominated primarily by Ponderosa pine (*Pinus ponderosa*) and have a complex disturbance history. Ponderosa pine forests in central Oregon typically are subject to frequent, low-severity wildfires that occur naturally approximately every 16 years [Bork, 1985]. High-severity fires occur with considerably less frequency. In the old-growth stand at Metolius, a stand replacing fire appears to have occurred in 1750, killing 99% of all overstory trees in the stand. Additional moderate severity burns occurred in 1850 and 1950. More recently, anomalously warm, dry years (1985–1994, 2000–2005) contributed to regional drought stress [Thomas et al., 2009]. Low-intensity underburns were implemented in small areas of the old-growth forest (e.g., 1 ha) at the Research Natural Area since the 1950s. The age distribution in the tower footprint at the time of eddy covariance measurements was ~50% mixed age (45 and 250 years), 25% young (45 years) and 25% pure stands of old trees (250+) [Law et al., 1999]. Soils are well drained. Descriptions of site-specific data are referred to as they were used.

#### Table 1. Biome-BGC Parameters Randomly Sampled Between Limits

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Variable</th>
<th>Estimated Literature Values</th>
<th>Measured Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>FM</td>
<td>annual leaf and fine root turnover fraction (1/yr)</td>
<td>0.1–0.9</td>
<td>0.25</td>
</tr>
<tr>
<td>FRC:LC</td>
<td>new fine root C: new leaf C (ratio)</td>
<td>0.1–6.0</td>
<td>3.0</td>
</tr>
<tr>
<td>SC:LC</td>
<td>new stem C: new leaf C (ratio)</td>
<td>0.2–2.0</td>
<td>0.96</td>
</tr>
<tr>
<td>CRC:NSC</td>
<td>new croot C: new stem C (ratio)</td>
<td>0.2–0.5</td>
<td>N/A</td>
</tr>
<tr>
<td>C:N_leaf</td>
<td>C:N of leaves (kgC/kgN)</td>
<td>20–90</td>
<td>53.53</td>
</tr>
<tr>
<td>C:N_litter</td>
<td>C:N of leaf litter, after retranslocation (kgC/kgN)</td>
<td>90–150</td>
<td>88.6</td>
</tr>
<tr>
<td>C:N_root</td>
<td>C:N of fine roots (kgC/kgN)</td>
<td>20–90</td>
<td>79</td>
</tr>
<tr>
<td>C:N_dead wood</td>
<td>C:N of dead wood (kgC/kgN)</td>
<td>200–1800</td>
<td>330</td>
</tr>
<tr>
<td>LAI</td>
<td>canopy average specific leaf area, projected area basis (m²/kgC)</td>
<td>0.5–4.0</td>
<td>2.25</td>
</tr>
<tr>
<td>FLNR</td>
<td>fraction of leaf N in Rubisco (unitless)</td>
<td>0.01–0.15</td>
<td>N/A</td>
</tr>
<tr>
<td>g_max</td>
<td>maximum stomatal conductance, projected area basis (m/s)</td>
<td>0.002–0.012</td>
<td>0.008</td>
</tr>
<tr>
<td>Ψ_s</td>
<td>leaf water potential: start of conductance reduction (MPa)</td>
<td>−0.85 to −0.2</td>
<td>−1.5</td>
</tr>
<tr>
<td>Ψ_c</td>
<td>leaf water potential: complete conductance reduction (MPa)</td>
<td>−2.3 to −0.9</td>
<td>N/A</td>
</tr>
</tbody>
</table>

*For description of data collection procedures, see Law et al. [2003].

#### Table 2. Site Characteristics for the Metolius AmeriFlux Sites Simulated

<table>
<thead>
<tr>
<th></th>
<th>Young Stand</th>
<th>Mature Stand</th>
<th>Old Stand</th>
</tr>
</thead>
<tbody>
<tr>
<td>Latitude</td>
<td>44.32</td>
<td>44.45</td>
<td>44.49</td>
</tr>
<tr>
<td>Longitude</td>
<td>−121.61</td>
<td>−121.56</td>
<td>−121.62</td>
</tr>
<tr>
<td>Elevation</td>
<td>1008 m</td>
<td>1253 m</td>
<td>915 m</td>
</tr>
<tr>
<td>Stand age (90th percentile)</td>
<td>23</td>
<td>90</td>
<td>250</td>
</tr>
<tr>
<td>Overstory LAI</td>
<td>1.1 m² m⁻²</td>
<td>2.8 m² m⁻²</td>
<td>2.1 m² m⁻²</td>
</tr>
<tr>
<td>Dominant species</td>
<td>ponderosa pine</td>
<td>ponderosa pine</td>
<td>ponderosa pine</td>
</tr>
<tr>
<td>Soil texture</td>
<td>sandy loam</td>
<td>sandy loam</td>
<td>sandy loam</td>
</tr>
</tbody>
</table>

(90th percentile)
during the time in which the system was simulated. Additional site characteristics are summarized in Table 2.

2.5. Modeling Protocol

All simulations used version 4.1.2 of the Biome-BGC model [Thornton et al., 2002], a widely used terrestrial ecosystem model. Biome-BGC simulates water, carbon, and nitrogen dynamics in plants, litter, and soil, using a daily time step for all processes [Running and Coughlin, 1988; White et al., 2000]. Biome-BGC allows for the option of a spin-up simulation to serve as a basis for an initial estimate of soil C content. Spin-up time is determined by the amount of time it takes to allow soil C to reach equilibrium [Thornton and Rosenbloom, 2005]. We incorporated the same randomly generated parameter values in the spin-up simulations for each of our GLUE analysis simulations. Among the young and mature stands, we incorporated a representation of the stand’s disturbance history into the regular (non-spin-up) simulations. Our methodology for this was similar to, but not an exact replicate of, the methodology developed by Law et al. [2001].

The incorporation of disturbance history for the old-growth stand was more complex. Thornton et al. [2002] report a disturbance history involving a stand replacing (~99% mortality) fire in 1750, a medium-severity fire (~25% mortality) in 1850, and another medium-severity fire (~25% mortality) in 1950. We incorporated this disturbance history by running the spin-up simulation and then beginning the normal simulation in the year 1750. The simulation was then run for 100 years until 1850, at which point a disturbance resulted in a stem mortality of between 20 and 30% (to account for the uncertainty in the disturbance severity estimate of 25%). The mortality percentage generated for the disturbance in 1850 was then multiplied by the estimates of stem C to estimate the contribution of newly dead stem C and leaf C to the coarse woody debris and litter pools, respectively. Next, a new initialization file was generated that incorporated these estimates of newly created coarse woody debris C and litter C, which were then added to the pools left over from the simulation before the fire. The amount of stem C and leaf C remaining after the disturbance was calculated by multiplying the amount of leaf C remaining after the simulation to 1850 (1 - disturbance mortality fraction) and was also incorporated into the new initialization file. Prefire estimates of soil C, fine root C, and soil N, although likely to have marginally changed in the disturbance, were not modified from their values at the end of the simulation ending in 1850. The same procedure was performed from 1851 until 1950, when another fire with ~25% mortality occurred. Finally, starting in 1951, a new simulation was run with no major disturbances apart from the annual fire mortality fraction applied to every year. The final simulation was run until the end of the year 2000, the year in which the eddy covariance data were collected.

2.6. Data Collection

Flux data collection protocols are described by Anthoni et al. [2002] and Thomas et al. [2009]. In brief, the eddy covariance method estimates NEE flux from the covariance of high-frequency fluctuations in vertical wind velocity and CO₂ concentrations. NEE is calculated as the sum of this flux term and a canopy CO₂ storage term, the latter of which is calculated from the change in CO₂ concentration in the canopy air space as a function of height [Law et al., 1999; Anthoni et al., 2002]:

\[
\text{NEE} = \overline{w c'} + \int_0^z \frac{dc}{dt} dz
\]

where \(\overline{w c'}\) is the time-averaged eddy flux for CO₂ [covariance between the turbulent fluctuations for vertical wind speed (\(w\)) and scalar concentration (\(c'\))] and \(dc/dt\) is a vertical storage term that is a function of canopy height (z), which approximates change in CO₂ storage in the canopy air space. The scalar concentrations were measured at 20 Hz and NEE are averaged in 30 min intervals which form the data set of daily estimates of NEE for each respective stand.

There are several potential sources of error in NEE estimates that form the basis of our uncertainty estimate to determine model acceptability. First, the fluxes that are computed over half-hour intervals to minimize sampling errors are associated with nonstationarity in the effective footprint over time; micrometeorological sampling errors [Baldočhi, 2003]; and statistical sampling errors from gap-filling methodologies [Falge et al., 2001]. We used the uncertainty estimates calculated by Thomas et al. [2009] for the Metolius sites, where the uncertainty (coefficient of variation) for annual NEE is estimated to be 16% of the mean. All sites have been evaluated for accuracy and consistency of eddy covariance-based GPP and ecosystem respiration with biometric measurements of NPP, autotrophic respiration, and heterotrophic respiration, including closure of the CO₂ balance to account for errors in each of the component fluxes, and the sites passed both tests [Luyssaert et al., 2009; Vickers et al., 2009].

2.7. Meteorological Data

The driving meteorological data for Biome-BGC is composed of the following inputs given on a daily time step: maximum temperature (\(T_{\text{max}}\)), minimum temperature (\(T_{\text{min}}\)), average temperature (\(T_{\text{avg}}\)), average vapor pressure deficit (VPD) (MPa), average incoming shortwave radiation (\(S_{\text{rad}}\)) (W m⁻²), total precipitation (mm), and day length (s). Meteorological instrumentation did not exist at the Metolius site prior to its establishment as an AmeriFlux site, requiring the generation of such data for the years before the eddy covariance instrumentation was installed at the site. This need was met using the DAYMET climate model, a model which gathers data for a user-specified location by extrapolating meteorological readings from surrounding climate stations and adjusting for any changes in elevation [Thornton and Running, 1999; Thornton et al., 2000]. DAYMET generated daily climate data from 1980 through 2000. For the young stand, meteorological data was used starting in 1987. For the mature and old-growth stands, the meteorological data was recycled so that it could provide a record of appropriate length for each stand. Meteorological data taken from the AmeriFlux instrumentation then replaced the data generated by the DAYMET model for the time span of our analysis. In addition to incorporating meteorological data, Biome-BGC allows for the user to specify yearly CO₂ concentrations at the site, based on annual CO₂ concentrations recorded since 1901, and we utilized this feature to account
for changes in atmospheric CO₂ at this point. CO₂ concentrations prior to 1901 were based on the inferred CO₂ concentrations from Intergovernmental Panel on Climate Change [2007].

We suggest that the assumptions made in setting up Biome-BGC for this study are rather typical of this type of ecosystem simulation. Unique to this study is the exploration of uncertainty and potential equifinality of different models as hypotheses about functioning of the *Pinus ponderosa* system (see also Mitchell et al. [2009] for application to an additional *Pinus ponderosa* ecosystem). The inputs described above were the basis for 500,000 simulations performed with different randomly chosen parameter sets from the ranges specified in Table 1. A complete list of the parameter values used in the simulations can be found in Table S1 (available as auxiliary material).

We recognize that there will be an interaction between errors in the inputs and any parameter sets that are identified as behavioral within the GLUE methodology (see discussions in the works of Beven [2006]) but, as in very many environmental modeling studies, there is little information available with which to assess the potential input errors.

3. Results

3.1. Parameter Sensitivities

Of the simulations that were run for the young, mature, and old-growth stands, 80.77%, 79.26%, and 73.00% were nonliving (GPP < 10 g C m⁻² yr⁻¹) and were thus excluded from further analysis. The direction in which the errors were propagated differed for the simulations (Figure 2). In the young stand, 99.73% of the simulations overestimated the magnitude of measured NEE. However, among the living simulations for the mature and old-growth
stands, 100% and 99.84% underestimated the magnitude of measured NEE.

[22] Beven [2006] has suggested an approach to model rejection based on setting prior limits of acceptability. Here this approach has been implemented by defining such limits on the basis of the effect of error in the field measurements as in condition above. Such an approach allows a refinement of the set of retained models to treat as behavioral only those providing predictions that are satisfactory in the sense of being consistent with the estimated errors in the measurements and that might therefore be considered as providing reliable simulations. Among the living simulations in the young stand, 14.04% of the estimates of mean annual GPP were within 16% of the measured estimate of 745 g C m\(^{-2}\) yr\(^{-1}\). For the living simulations in the mature stand, 2.89% of the estimates of mean annual GPP were within 16% of the measured estimate of 1583 g C m\(^{-2}\) yr\(^{-1}\) (Figure 3). For the living simulations in the old-growth stand, 19.79% of the estimates of GPP (year 2000) were within 16% of the measured estimate of 938 g C m\(^{-2}\) yr\(^{-1}\). Perhaps surprisingly, estimates of total ecosystem respiration (TER) were more accurate than estimates of GPP. Among the living simulations in the young stand, 16.64% of the estimates of mean annual TER were within 16% of the measured estimate of 1049 g C m\(^{-2}\) yr\(^{-1}\). Among the living simulations in the mature stand, 15.90% of the estimates of mean annual TER were within 16% of the measured estimate of 1115 g C m\(^{-2}\) yr\(^{-1}\). Among the living simulations for the old-growth stand, 25.00% of the estimates of mean annual TER were within 16% of the measured estimate of 430 g C m\(^{-2}\) yr\(^{-1}\) (Figure 3).

[23] Reproduction of estimates of GPP and TER were considerably more accurate than estimates of NEE. This is due, in part, to the fact that estimates of NEE are dependent on the estimates of both GPP and TER. An accurate estimate of NEE does not imply accurate estimates of both GPP and TER; it only means that the difference between the two is equal to the difference between the two estimates measured in the field. Among the living simulations for the young stand, a mere 1.04% of the simulations resulted in satisfactory (=16% of mean annual NEE) estimates for the stand (Figure 3) while 0.63% of living old-growth simulations had reproductions of mean annual NEE that were satisfactory. None (0.00%) of the simulations for the mature stand were within 16% of the mean annual NEE (Figure 3).

[24] Estimates of NEE at the Metolius sites show significant interannual variability, and part of the model-data mismatch may result from an inability to respond to interannual climatic fluctuations. Thomas et al. [2009] demonstrated that soil water content is the main determinant of ecosystem carbon flux in the Metolius sites, and that the interannual variability of the onset of the growing season is large (45 days). From 2002 to 2008, annual NEE at the mature stand ranged from −250 g C m\(^{-2}\) yr\(^{-1}\) in 2003 to −603 g C m\(^{-2}\) yr\(^{-1}\) in 2008, with an average annual NEE of −471 g C m\(^{-2}\) yr\(^{-1}\) for the overall time period. Also worth noting are the findings of Vickers et al. [2009], which demonstrate that the young stand may be even more sensitive to drought than the mature stand. From 2004 to 2008, NEE at the young stand varied from 45 g C m\(^{-2}\) yr\(^{-1}\) in 2005 to 189 g C m\(^{-2}\) yr\(^{-1}\) in 2008, with an average of 119 g C m\(^{-2}\) yr\(^{-1}\) for that time period. (Note that nonnegative NEE values imply a transfer of terrestrial ecosystem C to the atmosphere.)

3.2. Uncertainties in Disturbance History

[25] To ascertain the role of disturbance history on the validity of NEE estimates for the old-growth site, we took the parameter set that best emulated NEE estimates and ran 10,000 simulations, each with different fire severities. Unlike the previous GLUE exercise, which simulated disturbances with severities of 20–30% during the years of 1850 and 1950 (see section 2.5 above), we simulated the effects of variation in disturbance history simulating disturbances with severities of 0–100% in 1850 and 1950. We also allowed annual fire-related tree mortality to vary from 0.1 to 2.0% to incorporate the variation in fire-related mortality that results from frequent (MFRI = 16 years), low-severity fires among Pinus ponderosa trees, which are well adapted to low-intensity surface fires. In a study on recent fires in Ponderosa pine near the site, mean basal area mortality ranged from 14% in low-severity stands to 50% in moderate severity and 100% in high-severity stands [Meigs et al., 2009]. Results indicate that annual fire-related tree mortality was a highly sensitive parameter, with lower mortality estimates resulting in greatly improved estimates of NEE (Figure 4). Similarly, low estimates of fire severity in both 1850 and 1950 resulted in the best reproduction of NEE data (Figure 4), but the sensitivities of these parameters were much lower than that of annual fire-related tree mortality.

4. Discussion

[26] Semiarid ecosystems are a particularly challenging case study in ecosystem modeling, and a failure to reproduce data for one semiarid ecosystem does not indict the model’s performance for all ecosystem types. However, assuming our methodology for parameterization is representative of typical model applications, and that we characterized disturbance history sufficiently, the model performed inadequately in this sensitivity test because there does not appear to be a single instance in which both GPP and TER were accurately modeled (Figure 5) in the mature and old-growth stands, even if certain combinations of them could result in accurate estimates of NEE for the old-growth stand.

4.1. Soil Hydrology

[27] We initially suspected that difficulties in modeling the soil hydrology in the Metolius ecosystem accounted for a large proportion of model-data mismatch, as the sensitivity of this ecosystem to soil hydrology has been documented [Irvine et al., 2004; Thomas et al., 2009]. We investigated this possibility by examining the estimates of mean annual soil water content (θ_s) and mean annual GPP. In the young and old-growth stands, modeled estimates of mean annual soil water content were universally higher than the measured values. Conversely, the model reproduced a lower amount of soil water content at the mature stand, though the difference was not nearly as dramatic as it was for the young and old-growth stands.

[28] Much of the discrepancy may be due to difficulties in reproducing taproot dynamics of semiarid forests. Taproots can extend deep into the soil in semiarid ecosystems as an adaptation for chronic water limitations, and it is striking
Figure 3. Proportions of carbon flux values for live simulations reproduced by the model. Nonliving simulations are not shown. Black lines represent range of values that could be deemed satisfactory (±cv).
that, in the old-growth stand, the model produced a higher soil water content than that observed in the field. We suspect that this may be due to the inability to account for, and respond to, the availability of deep soil water, which is associated with precipitation at longer timescales than the daily precipitation that is used in the simulations; Biome-BGC simulates water availability with a simplified representation of the soil water column and associated water uptake.

Fortunately, even though the model did not accurately simulate mean soil water content, this does not mean that all of the simulated estimates of mean annual GPP were inaccurate. On the contrary, there were many estimates of annual GPP that were comparable to field data. Similarly, successful reproductions of the ratio of annual GPP to annual evapotranspiration (ET) were found in an earlier GLUE analysis of the Biome-BGC model for young and mature aged ponderosa pine stands, even when no simulation accurately reproduced the relationship between soil water content and ET [Mitchell et al., 2009].

4.2. Soil Respiration

Approximately 80% of global GPP is respired back to the atmosphere, and soil respiration, which incorporates elements of both \( R_a \) and \( R_h \), accounts for more than two-thirds of this flux [Law et al., 1999; Janssens et al., 2001; Xu et al., 2001]. Difficulties in modeling soil respiration are well-documented. Braswell et al. [2005] applied nonlinear inversion to the eddy covariance flux measurements from Harvard Forest using a simplified model of photosynthesis and evapotranspiration and concluded that multiyear eddy flux measurements allow for a tight constraint on photosynthesis, but poor constraints on parameters relating to soil decomposition, which varies at considerably longer timescales than photosynthesis and evapotranspiration. Similarly, Verbeeck et al. [2006] found that the parameter responsible for the greatest amount of uncertainty in the FORUG model was related to soil respiration, and Williams et al. [2005] concluded that long-term measurements of carbon pool sizes are needed to estimate parameters relating to soil decomposition.

We investigated the extent to which modeled estimates of soil respiration contributed to model-data mismatch by performing a subspecies of GLUE analyses for 10,000 runs. While the simulations for the young stand overlapped with the field data, the vast majority of the simulations for the mature stand and all of the simulations for the old-growth stand underestimated soil respiration estimates from the field (Figure 6). We note that only periodic soil respiration measurements were made at these sites, and scaling with temperature response curves over the year can lead to larger uncertainty in annual estimates than use of continuous measurements. Likewise, Law et al. [1999] showed that on calm nights, soil respiration could exceed eddy covariance estimates of total respiration, suggesting periods of nocturnal cold air drainage. However, even though soil respiration was underestimated in most of the 10,000 runs for both the mature and old-growth stands, this underestimation actually improved model performance. Even though modeled estimates of soil respiration were
Figure 5. Mean annual GPP plotted against total ecosystem respiration (TER), mean annual GPP plotted against autotrophic respiration ($R_a$), and mean annual NPP plotted against heterotrophic respiration ($R_h$) for all three stands. Modeled values are shown in blue, while the range in uncertainty (± cv) regarding the measured values is shown in black. Note that field-derived estimates of $R_a$ in the old stand are nearly 0 and are likely to be significantly underestimated, meaning that eddy flux estimates of GPP and/or TER for the old stand are probably inaccurate.
underestimated, it was an inaccuracy that actually increased the accuracy of the estimates of NEE, since TER was almost always overestimated. Thus, even though our results lend support to prior findings [Braswell et al., 2005; Williams et al., 2005; Verbeeck et al., 2006] that suggest that the frameworks for modeling soil respiration are inadequate, we cannot attribute the entirety of mismatch in our study to soil respiration.

4.3. Gross Primary Production and Total Ecosystem Respiration

While some of the simulations for the young stand were successful in reproducing the relationship between GPP and TER, none of the modeled estimates for the mature and old-growth stands were successful in reproducing the mean annual GPP: mean annual TER ratio (Figure 5). Similar results were found in an earlier GLUE analysis of the Biome-BGC model for two younger stands by Mitchell et al. [2009]. To determine what component of ecosystem respiration (autotrophic or heterotrophic) is causing model-data mismatch, we utilized field estimates of NPP to determine autotrophic respiration. Once autotrophic respiration \( R_a \) was known, the heterotrophic portion of ecosystem respiration \( R_h \) was calculated \( (R_h = TER - R_a) \).

Our results demonstrate that autotrophic respiration is not being simulated reliably in both the mature and old-

![Figure 6. Mean annual GPP plotted against mean annual soil respiration for all three stands. Even though soil respiration was underestimated in most of the 10,000 simulations for the mature stand and all of the simulations for the old-growth stand, this underestimation actually improved model performance since total ecosystem respiration (TER) was almost always overestimated for these stands.](image)

![Figure 7. Modeled estimates \( R_a/GPP \) compared to field estimates. Note that \( R_a/GPP \) is increasingly overestimated in the mature and old-growth stands, indicating that \( R_a \) is a significant source of model-data mismatch. Note that field-derived estimates of GPP in the old stand may be significantly underestimated (see Figure 5), meaning that field estimates of TER for the old stand are probably inaccurate.](image)
growth stands. Simulations for both stands showed that the model almost always overestimated the portion of GPP that was lost via $R_a$ for the mature and old-growth stands (Figure 7). The trend of simulated fluxes observed in Figure 7 runs contrary to field observations where $R_a/GPP$ is initially low [DeLucia et al., 2007]. The role of $R_a$ in model-data mismatch is further illustrated in Figure 8, which shows that when estimates of $R_a$ and $R_h$ are adjusted according to estimates of field measurements of $R_a/GPP$ and $R_h/R_a$. (bottom right) Histogram showing the distribution of the corresponding values of $GPP - R_a - R_h$ obtained from the adjusted values of $R_a$ and $R_h$ from the scatterplot in Figure 8 (top right). Note that when the estimates of $R_a$ and $R_h$ are adjusted according to the ratio observed from field measurements, the new estimates of $GPP - R_a - R_h$ are substantially improved.

Figure 8. (top left) Scatterplot showing the modeled values of $R_a$ and $R_h$ plotted against GPP for the living simulations for the mature stand, with (bottom left) the histogram representing the distribution of the corresponding modeled values of $GPP - R_a - R_h$. (top right) Scatterplot showing the adjusted values of $R_a$, calculated by multiplying the modeled value of $R_a$ by the ratio of measured $R_a/GPP$ to modeled $R_a/GPP$. Adjusted $R_h$ was calculated by multiplying the adjusted modeled value of $R_a$ by the field measurements of $R_h/R_a$. (bottom right) Histogram showing the distribution of the corresponding values of $GPP - R_a - R_h$ from the scatterplot in Figure 8 (top right). Note that when the estimates of $R_a$ and $R_h$ are adjusted according to the ratio observed from field measurements, the new estimates of $GPP - R_a - R_h$ are substantially improved.

Simulations for both stands showed that the model almost always overestimated the portion of GPP that was lost via $R_a$ for the mature and old-growth stands (Figure 7). The trend of simulated fluxes observed in Figure 7 runs contrary to field observations where $R_a/GPP$ is initially low [DeLucia et al., 2007]. The role of $R_a$ in model-data mismatch is further illustrated in Figure 8, which shows that when estimates of $R_a$ and $R_h$ are adjusted according to estimates of field measurements of $R_a/GPP$ and $R_h/R_a$, estimates of $GPP - TER$ (i.e., $GPP - R_a - R_h$) are substantially improved. Evidence for a failure to adequately simulate $R_a$ was also found by Wang et al. [2009], who calibrated Biome-BGC to produce highly accurate estimates of GPP and evapotranspiration (ET), but noted that NPP ($GPP - R_a$) was nevertheless underestimated when compared to field measurements. An overestimation of $R_a$ also explains why our simulations for the mature and old-growth stands were run best with parameter values that were often on the extreme ends of value ranges, as they resulted in an increase in GPP that compensated for an overestimation of $R_a$. Consequently, parameters that have a more direct control over potential GPP, such as those that control leaf production and leaf nitrogen concentration, are clearly among the most sensitive model parameters in Biome-BGC [White et al., 2000]. The framework for modeling $R_a$ was largely derived from that outlined by Ryan [1991].

[34] Of the living simulations for the mature and old-growth stands, 100% and 99.84% underestimated the magnitude of NEE. An underestimation of NEE in forest ecosystems is not unprecedented for the Biome-BGC model. Thornton et al. [2002] reported that, in the year 2000, Biome-BGC estimated an annual NEE of $-25$ g C m$^{-2}$ y$^{-1}$.
for the old-growth stand at Metolius, a significant underestimation of the −508 g C m$^{-2}$ yr$^{-1}$ that was recorded. An overestimation of both autotrophic and heterotrophic respiration cascades through the years simulated, as each inaccurate year of simulation inevitably carries over into the following year. The results of these cascading errors can result in a significant deviation from field-based estimates of total ecosystem C storage. Thornton et al. [2002] provided modeled estimates of total C storage (not to be confused with annual NEE) for the old-growth site at Metolius at 140 Mg C ha$^{-1}$, while the field-based estimates of Law et al. [2001, 2003] are 210 and 239 Mg C ha$^{-1}$ (different years of measurement), ~50% higher than the modeled estimates. A similar tendency for estimates of the magnitude of annual NEE to be underestimated appears to be present in modeled estimates of other ecosystems as well. Simulations performed by Thornton et al. [2002] for the Wind River Experimental Forest in the western Cascade Range of Washington estimated stand biomass to be 443 Mg C ha$^{-1}$, while field measurements reported by Smithwick et al. [2002] estimated stand C storage to be 614 Mg C ha$^{-1}$, 39% higher than the modeled estimates.

[35] Finally, the frequency and severity of previous disturbances, such as high-frequency low-severity wildfires, as well as moderate-frequency wildfires of moderate-to-high severity, exert significant influence on estimates of NEE, particularly at stands with long and complex disturbance histories, like the old-growth stand. Our work on the sensitivity of disturbance history to NEE estimates reiterated the findings by Thornton et al. [2002], which emphasized the importance of incorporating site-specific patterns of disturbance into simulations. After running 10,000 simulations with the highest likelihood model parameter set, we found that, for the disturbances simulated in 1850 and 1950, lower-severity disturbances resulted in simulations with higher likelihood (Figure 4). This was true even for disturbance severity values that were significantly lower than the actual severities of the disturbances. Similarly, the annual rate of fire disturbance had the higher likelihoods at the lower end of the parameter value range (Figure 4). The reason for this confluence is likely similar to the reasons for other shaping in parameter values, as lower mortality from disturbance results in more photosynthetic machinery, thus higher GPP to offset the model’s inability to compensate for the difficulties in simulating ecosystem respiration.

[36] However, there is an important difference in the application of field-based measurements of year-specific mortality from average annual mortality. Ascertainment of the extent of significant disturbances, such as those that occurred in 1750, 1850, and 1950, can be done with considerably more certainty than estimating the fine-scale mortality processes that occur with a high-frequency, low-severity fire regime endemic to the east Cascades; reconstructing a stand-level disturbance with moderate to high severity can be easily done through dendrochronology, while estimating the annual fraction of fire-related mortality from fires of lower severity is more difficult and may yet prove to be essential to estimates of NEE in terrestrial ecosystem models. Since the fine-scale annual rate of fire-related mortality is a sensitive model parameter that is difficult to estimate, incorporating such disturbances, while necessary, is likely to be a major challenge to finding behavioral estimates of NEE. We suspect the same is true for other fine-scale mortality processes, such as pest outbreaks (i.e., pine beetle infestations). Simulating an aggregated estimate of the annual disturbance-related mortality, rather than simulating the full magnitude of actual disturbances in the years in which they occur, is likely too coarse a representation of the disturbance processes. Furthermore, the frequency and extent of both of these mortality agents is likely to have changed over the time period in which they were measured, and may change in the future as a result of continued climatic change, making a reconstruction difficult and a future projection of them even more challenging.

5. Conclusions

[37] We have incorporated the uncertainty that arises from disturbance history, multidimensional parameter variability, and eddy flux measurement uncertainty by simulating 500,000 combinations of 13 parameter values for three age classes of semiarid forest and tested to see if estimates of NEE from those simulations can fall within the bounds of measurement uncertainty inherent in estimates of NEE based on eddy flux measurements. Studies that provide an evaluation of the sources of uncertainty to this extent are rare, and our simulations of net ecosystem CO$_2$ exchange in these semiarid forests with Biome-BGC were challenged by some inconsistencies with observations. We note that the model can provide accurate reproductions of GPP, meaning that what is arguably the most critical ecosystem process, photosynthesis, is reasonably well understood. Nevertheless, it is clear that substantial uncertainties remain in modeling ecosystem respiration, as our results indicate that respiration frameworks, particularly those for autotrophic respiration, require modification. However, even if both autotrophic and heterotrophic respiration can be accurately modeled, coping with the equifinality that results from multidimensional parameter variability may still present obstacles to the accurate reproduction of ecosystem processes. We also note that uncertainties in disturbance history can undermine the accuracy of even the most robustly calibrated models. For these reasons, we contend that an incorporation of measurement uncertainty, multidimensional parameter variability, and uncertainties in ecosystem disturbance history may prove valuable in efforts to detect specific problems in model structure and we encourage such a procedure in future model assessments.

[38] Acknowledgments. The Metolius AmeriFlux sites are supported by the U.S. Department of Energy Terrestrial Carbon Program (grant DE-FG02-03ER63653). Discussions regarding Biome-BGC with Ron Neilson at the United States Forest Service Pacific Northwest Research Station were also very helpful, as was his generosity in letting us use his Linux cluster. Biome-BGC version 4.1.2 was provided by Peter Thornton at the National Center for Atmospheric Research (NCAR), and by the Numerical Terradynamic Simulation Group (NTSG) at the University of Montana. NCAR is sponsored by the National Science Foundation. Support for this research was provided by an NSF IGERT graduate fellowship to S. R. Mitchell (NSF award 0333257) in the Ecosystem Informatics IGERT program at Oregon State University, a NASA New Investigator Program grant to K. E. B. O’Connell (NN604GR436) at Oregon State University, and the H. J. Andrews LTER (DEB-0218088). Development of the GLUE methodology has been supported by UK NERC grant NER/L/S/2001/00659 and NE/E002242/1.


K. Beven, Lancaster Environment Centre, Lancaster University, Lancaster LA1 4YQ, UK.
J. Freer, School of Geographical Sciences, University of Bristol, Bristol BS8 1SS, UK.
B. Law, Department of Forest Ecosystems and Society, College of Forestry, Oregon State University, Corvallis, OR 97331, USA.
S. Mitchell, Nicholas School of the Environment, Duke University, Durham, NC 27708, USA. (stephen.mitchell@duke.edu)