Rate my data: quantifying the value of ecological data for the development of models of the terrestrial carbon cycle

TREVOR F. KEENAN,1,4 ERIC A. DAVIDSON,2 J. WILLIAM MUNGER,3 AND ANDREW D. RICHARDSON1

1Department of Organismic and Evolutionary Biology, Harvard University, 22 Divinity Avenue, Cambridge, Massachusetts 02138 USA
2Woods Hole Research Center, 149 Woods Hole Road, Falmouth, Massachusetts 02540 USA
3School of Engineering and Applied Sciences and Department of Earth and Planetary Sciences, Harvard University, Cambridge, Massachusetts 02138 USA

Abstract. Primarily driven by concern about rising levels of atmospheric CO2, ecologists and earth system scientists are collecting vast amounts of data related to the carbon cycle. These measurements are generally time consuming and expensive to make, and, unfortunately, we live in an era where research funding is increasingly hard to come by. Thus, important questions are: “Which data streams provide the most valuable information?” and “How much data do we need?” These questions are relevant not only for model developers, who need observational data to improve, constrain, and test their models, but also for experimentalists and those designing ecological observation networks.

Here we address these questions using a model-data fusion approach. We constrain a process-oriented, forest ecosystem C cycle model with 17 different data streams from the Harvard Forest (Massachusetts, USA). We iteratively rank each data source according to its contribution to reducing model uncertainty. Results show the importance of some measurements commonly unavailable to carbon-cycle modelers, such as estimates of turnover times from different carbon pools. Surprisingly, many data sources are relatively redundant in the presence of others and do not lead to a significant improvement in model performance. A few select data sources lead to the largest reduction in parameter-based model uncertainty. Projections of future carbon cycling were poorly constrained when only hourly net-ecosystem-exchange measurements were used to inform the model. They were well constrained, however, with only 5 of the 17 data streams, even though many individual parameters are not constrained. The approach taken here should stimulate further cooperation between modelers and measurement teams and may be useful in the context of setting research priorities and allocating research funds.

Key words: biosphere-atmosphere interaction; carbon fluxes; carbon sequestration; climate change research; data assimilation; Harvard Forest, Massachusetts, USA; process-based models.

INTRODUCTION

In recent years our ability to collect vast amounts of data related to the structure and function of the biosphere, at both high temporal and spatial frequency, has greatly increased (Luo et al. 2008). New large-scale monitoring through networks such as NEON (national ecological observatory network), ICOS (integrated carbon observation system), FLUXNET (a network of regional networks integrating worldwide measurements of CO2, water, and energy flux between terrestrial systems and the atmosphere), and LTER (long-term ecological research sites) along with the extended satellite record, and data collation efforts such as TRY (a plant trait database; Kattge et al. 2011), are amassing tremendous amounts of data. However, the ultimate value of the accumulating diverse data sources will depend on the extent to which the data can be used to improve our understanding of, and ability to model, the earth system.

One of the main motivations for the increase in data availability is the need to improve our understanding of terrestrial carbon cycling (IPCC 2007). Much of the anthropogenically emitted CO2 cycles through terrestrial ecosystems. Current estimates of CO2 removed from the atmosphere by global photosynthesis stand at around 120 Pg C (Beer et al. 2010). A slightly smaller amount is respired back into the atmosphere, giving an estimated net global carbon sink in terrestrial ecosystems of 1-2 Pg C (Le Quéré et al. 2009, Pan et al. 2011). The main biological processes of photosynthesis and respiration that drive this cycle have long been identified. Large uncertainty remains, however, as to the mechanisms controlling the response of these processes to drivers at different spatial and temporal scales. This uncertainty is reflected in the broad range of model projections of the future of global terrestrial carbon storage (Friedlingstein et al. 2006, Heimann and Reichstein 2008), making the
implementation of effective policy difficult at best (IPCC 2007).

New approaches that can combine models with multiple data sources—"model-data fusion"—are emerging as a means to better understand the dominant processes controlling terrestrial carbon cycling. Such techniques can be employed both to directly inform carbon cycle models and as a tool to synthesize the growing amounts of data. The basic philosophy is that using data in a statistically rigorous manner to give the best model possible (conditional on model structure) can both highlight model deficiencies and integrate different data sources. Recent efforts have used a diverse range of data types with process-oriented models (e.g., Braswell et al. 2005, Williams et al. 2005, Moore et al. 2008, Richardson et al. 2010, Weng and Luo 2011, Keenan et al. 2012). A strength of the approach is that it can be used to assess the model against all observations simultaneously. Using multiple constraints goes beyond simple testing of a model against a single measurement type—the approach uses data to both test and inform model behavior for all aspects of the system for which observations are available. The result is a data-informed process-oriented model, which allows the researcher to quantify the degree of uncertainty in model projections.

Carbon-cycle modelers typically rely on experimental and observational data that have been collected by others. One of the most common questions asked of modelers by experimentalists and (more recently) data acquisition network designers is "What data are most useful?" In response to such questions, however, modelers generally do not have a better answer than what is essentially an educated guess. Indeed, from a modeling perspective using more data does not always lead to a better-constrained model (Richardson et al. 2010). In an environment of increasingly organized data-acquisition networks (Keller et al. 2008) and efforts that seek to merge models with data (Wang et al. 2007, Keenan et al. 2011a), it becomes imperative to develop ways of quantifying the usefulness of different data sources. By identifying the next measurement that should be made, which maximizes the information gained from all measurements together, the efficiency and cost effectiveness of measurement campaigns can be improved, along with model projections.

Here, we develop a framework to address the question, "How useful is a particular measurement for reducing uncertainty in a process-oriented model of terrestrial carbon cycling?" We use multiple data sources from long-term records at the Harvard Forest, in the northeastern United States, in combination with a model-data fusion framework. We rank the different data streams according to the incremental information that each data stream provides. We do this by iteratively testing the reduction in model uncertainty achieved by informing the model with each data source. At each step in the process we assess the impact of a particular measurement type on both short-term (diurnal, seasonal, annual) model projections, and long-term (decadal) model responses to climate change.

Materials and Methods

Site

Hourly model simulations were run for 12 complete years (1992–2003) at the Harvard Forest Environmental Measurement Site (HFEMS; information available online), located in the northeastern United States (42.53° N, 72.17° W; elevation 340 m) (Wofsy et al. 1993, Goulden et al. 1996, Barford et al. 2001, Urbanski et al. 2007). Measurements and simulations pertain to the area within the HFEMS tower footprint, which is largely comprised of deciduous trees. The area is dominated by the deciduous species red oak (Quercus rubra, 52% basal area), and red maple (Acer rubrum, 22% basal area), with a small conifer component that includes eastern hemlock (Tsuga Canadensis, 17% basal area), and occasional white and red pine (Pinus strobus, Pinus resinosa).

Data

All data used were gathered between 1992 and 2003. We used hourly meteorological and eddy-covariance (Wofsy et al. 1993, Urbanski et al. 2007) measurements of net ecosystem exchange (NEE; available online). Gap-filled meteorological variables used include hourly incident photosynthetically active radiation (PAR), air temperature above the canopy, soil temperature at a depth of 5 cm, vapor pressure deficit, and atmospheric CO₂ concentrations.

Quality-controlled hourly eddy-covariance observations (without gap-filling) of NEE were used to constrain parameters of an ecosystem model. Gap-filled data, or model-based partitioning of NEE to respiration and photosynthesis components, were not used. For ancillary data constraints we used 15 different data sources, which included measurements of leaf-area index, soil organic-carbon content, carbon in roots, carbon in wood, wood carbon annual increment, observer-based estimates of bud-burst and leaf senescence, leaf litter, woody litter, soil carbon turnover times, and three different measurement sets of soil respiration that capture spatial and methodological variability (Table 1). These data are freely available from the Harvard Forest Data Archive (available online) or the references in Table 1.

Measurement-based estimates of uncertainty were used for each data stream in the optimization. Flux uncertainty estimates were taken from Richardson et al. (2006), where uncertainties were shown to follow a double-exponential distribution, with the standard deviation of the distribution specified as a linear
function of the flux. Soil respiration uncertainty estimates were taken from Savage et al. (2009) and Phillips et al. (2010), where measurement uncertainty increased linearly with the magnitude of the flux. Estimates of uncertainties for the remaining data streams were based on either spatial variation or standard deviations from repeat sampling. Full details of uncertainty estimates are given in Keenan et al. (2012).

The FöBAAR Model

We used a forest carbon cycle model—FöBAAR (Forest Biomass, Assimilation, Allocation and Respiration; Keenan et al. 2012)—that strikes a balance between parsimony and detailed process representation. Working on an hourly timescale, FöBAAR calculates photosynthesis from two canopy layers, and respiration from eight carbon pools (leaf, wood, roots, soil organic matter [microbial, slow, and passive pools], leaf litter and [during phenological events] mobile stored carbon). Meteorological drivers considered are: canopy air temperature (Ta), 5 cm soil temperature (Ts), photosynthetic active radiation (PAR), vapor pressure deficit (VPD), and atmospheric CO2. Model parameters are given in Table 1.

The canopy in FöBAAR is described in two compartments representing sunlit and shaded leaves (Sinclair et al. 1976, Wang and Leuning 1998). Canopy light penetration depends on the position of the sun, and the area of leaf exposed to the sun based on leaf angle and the canopy’s ellipsoidal leaf distribution (Campbell 1986), assuming a spherical leaf angle distribution. Assimilation rates are calculated via the Farquhar approach (Farquhar et al. 1980, De Pury and Farquhar 1997). Stomatal conductance is calculated using the Ball–Berry model (Ball et al. 1987), coupled to photosynthetic rates through the analytical solution of the Farquhar, Ball Berry coupling (Baldocchi 1994).

Maintenance respiration is calculated as a fraction of assimilated carbon. The remaining assimilate is allocated to different carbon pools (foliar, wood and root) on a daily time step. Root respiration is calculated hourly and coupled to photosynthesis through the direct allocation to roots. Dynamics of soil organic matter is modeled using a three-pool approach (microbial, slow, and passive pools) (Knorr and Kattge 2005). Decomposition in each pool is calculated hourly, with a pool-specific temperature dependency. Litter decomposition is also calculated hourly, but on an air-temperature basis. Litter and root carbon are transferred to the microbial pool, then to the slow and finally to the passive pool. For further details on model structure see Keenan et al. (2012).

Model–data fusion

An adaptive multiple-constraints Markov-chain Monte Carlo approach was used to optimize the process-oriented model and explore model uncertainty. The algorithm uses the Metropolis-Hastings (M-H) approach (Metropolis and Ulam 1949, Metropolis et al. 1953, Hastings 1970) combined with simulated annealing (Press et al. 2007). Prior distributions for each parameter (Table 1) were assumed to be uniform (non-informative, in a Bayesian context).

The optimization process uses a two-step approach. (1) In the first stage the parameter space is explored for 100,000 iterations using the optimization algorithm. At
each iteration the current step size is used as the
standard deviation of random draws from a normal
distribution with mean 0, by which parameters are
varied around the previous accepted parameter set. This
stage identifies the optimum parameter set by minimizing
the cost function (see next paragraph). (2) In the
second stage, the parameter space is again explored
using a Markov chain starting from the optimum
identified in step 1. Acceptance of a parameter set is
based on whether the cost function for each data stream
passes a \( \chi^2 \) test (at 90% confidence) for acceptance/
rejection, after variance normalization (e.g., Franks et
al. 1999, Richardson et al. 2010).

The cost function quantifies the extent of model–data
mismatch using all available data (eddy-covariance,
biometric, and so forth). Individual data stream cost
functions, \( j_i \), are calculated as the total uncertainty-
weighted squared data–model mismatch, averaged by
the number of observations for each data stream (\( N_i \)):

\[
j_i = \frac{\left( \sum_{t=1}^{N_i} \frac{(y_i(t) - p_i(t))^2}{\delta_i(t)} \right)^2}{N_i}
\]  

(1)

where \( y_i(t) \) is a data constraint at time \( t \) for data stream \( i \)
and \( p_i(t) \) is the corresponding model predicted value.
\( \delta_i(t) \) is the measurement-specific model predicted value. For the
aggregate multi-objective cost function we use the
average of the individual cost functions, which can be
written as

\[
J = \frac{\sum_{i=1}^{M} j_i}{M}
\]  

(2)

where \( M \) is the number of data streams used.

Each individual cost function is averaged by the
number of observations for the relative data stream. The
average of the cost functions from all data streams is
taken as the total cost function. In this manner each
data stream is given equal importance in the optimization
(Franks et al. 1999, Barrett et al. 2005).

**Experimental set-up**

We used a simple three-step iterative algorithm for the
model experiment. The basic premise is to successively
add data streams as model constraints, according to
which data stream gives the best incremental reduction
in model uncertainty:

- **Step 1** For \( i = 1 \), perform model–data fusion with each
measurement type in Table 1 individually.
- **Step 2** Identify the single best measurement type, i.e.,
that which gives the minimum posterior distribution
of model–data mismatch (see below).
- **Step 3** For \( i = 2 \ldots M \), repeat steps 1 and 2 again to
identify the next best measurement type (in
addition to the data streams already selected). Do
this until no more data streams are available.

We calculate the reduction in model uncertainty
through the posterior distribution of model–data
mismatch (the difference between modeled and observed
variables, Eq. 2). At each iteration of Step 2 above, we
calculate the model uncertainty using the entropy of the
posterior distribution of model–data mismatch for each
data combination. Entropy is a measure of the
uncertainty associated with a random variable (Shannon
1948, Jaynes 1957, Kolmogorov 1968) and can be used
to quantify the information gained by the use of a
particular data source (e.g., Weng and Luo 2011). At
each Step 2 in the above algorithm, we identify the best
data combination as that which gives the lowest entropy
(and thus lowest model uncertainty) in the posterior
distribution of model–data mismatch. Running the
above algorithm took about three days on an 18-core
computational cluster.

**Climate projections to 2100**

We used the extracted posterior parameter distributions
to project carbon cycling and stocks to 2100 for
each step in the above-outlined experiment. This served
as an additional means by which to quantify the
incremental benefit of each additional data stream. For
the future climate scenario, we used downscaled data
(Hayhoe et al. 2007) from the regionalized projection of
the GFDL-CM global coupled climate–land model
(Delworth et al. 2006) driven with socio-economic
change scenario A1fi (IPCC 2007). Mean annual
temperature at Harvard forest, using this projection, is
predicted to increase from 7.1\(^{\circ}\) to 11.9\(^{\circ}\)C, with an
associated increase in atmospheric CO\(_2\) to 969 ppm by
2100.

**RESULTS**

**What measurements are most important?**

At each stage in the optimization process, we
identified the next-best measurement type by quantifying
how much each data stream reduced the uncertainty in
model projections (via Eq. 2). The most useful
measurements were those that quantified how carbon
flowed through the ecosystem at different time scales
(Fig. 1). In particular, the combination of measurements
on fast (net ecosystem exchange, soil respiration) and
slow (soil carbon turnover rates, monthly/annual
cumulative fluxes, litter from wood/leaves) carbon flows
in the ecosystem lead to the largest improvement in
model performance. Many measurements did not
inform the model in the presence of others. For example,
the use of data on the size of the soil carbon pool did not
lead to a large reduction in model uncertainty when soil
respiration data were available along with turnover rates
from the different soil carbon pools (Fig. 1). Estimates
of bud-burst dates did not lead to a large reduction in
model uncertainty, as they could be inferred by the
model from the eddy-covariance CO\(_2\) flux data. Observ-
vations of leaf senescence dates, on the other hand, were
highly ranked. Autumn shifts in carbon cycling are
driven by gradual biotic changes in canopy status, and co-occur with gradual abiotic changes in mean climate forcings. The senescence data, being biotic in nature, therefore improved the ability of the model to distinguish between autumn dynamics driven by biotic and abiotic changes. In addition to bud-burst data, litter turnover, and the proportion of autotrophic respiration in soil respiration measurements were ranked low, implying that the information contained in these measurements is also available from the higher ranked data (Fig. 1). The low ranking of nighttime net ecosystem exchange (NEE) is a good example of a situation where the information provided by a measurement is already present in another, as both annual and monthly NEE sums are constructed using nighttime NEE data.

The extent to which measurements can identify model parameters

When using all data, 26 of the 40 parameters included could be effectively identified (parameters \( a \) to \( y \), Table 2 and Fig. 2). Here, we consider a parameter identifiable if
the size of the posterior parameter distribution was half that of the prior distribution. In general, posterior parameter distributions were gradually reduced as more data streams were added to the system. Using all data together reduced the posterior parameter distributions by 60\% over all parameters (Fig. 3), when compared to the priors. The majority of the reduction in the range of posterior parameter distributions, however, was achieved with the use of relatively few data streams. For example, 14 parameters were well constrained with the use of only six different data sources (Fig. 2). The top 10 parameters that were most informed by the data related to the respiration rates of the different soil carbon pools, phenology, and litterfall. Fourteen parameters were not constrained, even when using all data together (parameters \( z \) to \( k_2 \)). These were predominantly related to canopy processes (e.g., leaf mass per area, dark respiration, photosynthetic potential, and the fraction of photosynthesis used for maintenance respiration), and rates of transfer between soil organic-matter carbon pools.

**Equifinality and parameter interactions**

When analyzing parameter posterior distributions in terms of parameter correlations, using additional data constraints increased the number of correlated parameters for the six data constraints that gave the largest reduction in model uncertainty (Fig. 4a). Using more data streams, in addition to these six, did not significantly change parameter correlations. Eight of the 40 parameters optimized were strongly correlated (\( r^2 / C^2 | 2 > 0.3 \)) when using all data to constrain the model. For example, the extracted values for photosynthetic potential (\( V_{cmax} \), Table 2; \( l_2 \), Fig. 4b) were highly correlated with the proportion of photosynthate lost as maintenance respiration (parameter 8, Table 2 and Fig. 4b).

---

**Table 2.** Parameters and pools used in our study are from the FöBAAR (Forest Biomass, Assimilation, Allocation, and Respiration) model (Keenan et al. 2012b).

<table>
<thead>
<tr>
<th>Parameter identification</th>
<th>Parameter name</th>
<th>Definition</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>SOM(_{Fp})</td>
<td>passive SOMC respiration rate (log)</td>
<td>-10</td>
<td>-1</td>
</tr>
<tr>
<td>b</td>
<td>SOM(_{Fp})SDT</td>
<td>fast-cycling SOMC temperature dependence</td>
<td>0.01</td>
<td>0.1</td>
</tr>
<tr>
<td>c</td>
<td>SOM(_{Fp})D</td>
<td>fast-cycling SOMC respiration rate (log)</td>
<td>-6</td>
<td>-1</td>
</tr>
<tr>
<td>d</td>
<td>ArTS</td>
<td>leaf senescence onset mean air temperature (°C)</td>
<td>0</td>
<td>15</td>
</tr>
<tr>
<td>e</td>
<td>Lf(_{s})</td>
<td>litterfall from foliage (log)</td>
<td>-6</td>
<td>-1</td>
</tr>
<tr>
<td>f</td>
<td>SOM(_{Fp})SD</td>
<td>slow- cycling SOMC respiration rate (log)</td>
<td>-6</td>
<td>-1</td>
</tr>
<tr>
<td>g</td>
<td>Lf(_{w})</td>
<td>litterfall from wood (log)</td>
<td>-6</td>
<td>-1</td>
</tr>
<tr>
<td>h</td>
<td>FC</td>
<td>fraction of Cf not transferred to mobile carbon</td>
<td>0.4</td>
<td>0.7</td>
</tr>
<tr>
<td>i</td>
<td>GDD(_{0})</td>
<td>day of year for growing degree day initiation</td>
<td>50</td>
<td>150</td>
</tr>
<tr>
<td>j</td>
<td>Lit2SOM</td>
<td>litter to fast SOMC transfer rate (log)</td>
<td>-6</td>
<td>-1</td>
</tr>
<tr>
<td>k</td>
<td>Lf(_{f})</td>
<td>litterfall from roots (Log)</td>
<td>-6</td>
<td>-1</td>
</tr>
<tr>
<td>l</td>
<td>Af</td>
<td>fraction of GPP allocated to foliage</td>
<td>0.5</td>
<td>1</td>
</tr>
<tr>
<td>m</td>
<td>Lit3T</td>
<td>litter respiration temperature dependence</td>
<td>0.01</td>
<td>0.1</td>
</tr>
<tr>
<td>n</td>
<td>Litc</td>
<td>carbon in litter</td>
<td>10</td>
<td>1000</td>
</tr>
<tr>
<td>o</td>
<td>Rc</td>
<td>carbon in roots</td>
<td>20</td>
<td>500</td>
</tr>
<tr>
<td>p</td>
<td>Ltt</td>
<td>litter respiration rate (log)</td>
<td>-6</td>
<td>-1</td>
</tr>
<tr>
<td>q</td>
<td>R(_{root})</td>
<td>root respiration rate (log)</td>
<td>-6</td>
<td>-1</td>
</tr>
<tr>
<td>r</td>
<td>Mobc-Tr</td>
<td>fraction of mobile transfers respired</td>
<td>0</td>
<td>0.01</td>
</tr>
<tr>
<td>s</td>
<td>Rsoil(_{1})</td>
<td>soil respiration scaling co-efficient (data set 1)</td>
<td>0.5</td>
<td>1.5</td>
</tr>
<tr>
<td>t</td>
<td>WC</td>
<td>carbon in wood</td>
<td>8000</td>
<td>14000</td>
</tr>
<tr>
<td>u</td>
<td>Ar</td>
<td>fraction of NPP allocated to roots</td>
<td>0.5</td>
<td>1</td>
</tr>
<tr>
<td>v</td>
<td>LitSOMT</td>
<td>litter to fast SOMC temperature dependence</td>
<td>0.03</td>
<td>0.5</td>
</tr>
<tr>
<td>w</td>
<td>Rsoil(_{2})</td>
<td>soil respiration scaling co-efficient (data set 2)</td>
<td>0.5</td>
<td>1.5</td>
</tr>
<tr>
<td>x</td>
<td>SOM(_{Fp})P</td>
<td>carbon in passive cycling SOM layer</td>
<td>2000</td>
<td>12000</td>
</tr>
<tr>
<td>y</td>
<td>SOM(_{Fp})S</td>
<td>carbon in slow cycling SOM layer</td>
<td>2000</td>
<td>12000</td>
</tr>
<tr>
<td>z</td>
<td>Mobc-R</td>
<td>mobile stored carbon respiration rate (log)</td>
<td>-6</td>
<td>-1</td>
</tr>
<tr>
<td>1</td>
<td>GDD(_{s})</td>
<td>growing degree days for spring onset</td>
<td>150</td>
<td>300</td>
</tr>
<tr>
<td>2</td>
<td>SOM(<em>{Fp})SOM(</em>{p})S</td>
<td>fast SOMC to slow SOMC rate</td>
<td>0.03</td>
<td>0.5</td>
</tr>
<tr>
<td>3</td>
<td>SOM(_{Fp})S</td>
<td>carbon in slow-cycling SOM layer</td>
<td>2000</td>
<td>12000</td>
</tr>
<tr>
<td>4</td>
<td>SOM(<em>{Fp})SOM(</em>{p})P</td>
<td>transfer rate from slow to passive SOM</td>
<td>0.001</td>
<td>0.4</td>
</tr>
<tr>
<td>5</td>
<td>SOM(<em>{Fp})SOM(</em>{p})PT</td>
<td>fast SOMC to slow SOMC temp. dependence</td>
<td>0.03</td>
<td>0.5</td>
</tr>
<tr>
<td>6</td>
<td>R(_{root})</td>
<td>root respiration rate temperature dependence</td>
<td>0.01</td>
<td>0.2</td>
</tr>
<tr>
<td>7</td>
<td>GDD(_{s})</td>
<td>spring photosynthetic GDD maximum</td>
<td>5</td>
<td>1000</td>
</tr>
<tr>
<td>8</td>
<td>MaintR</td>
<td>fraction of GPP respired for maintenance</td>
<td>0.1</td>
<td>0.4</td>
</tr>
<tr>
<td>9</td>
<td>LMA</td>
<td>leaf mass per area (g C/m(^2))</td>
<td>50</td>
<td>90</td>
</tr>
<tr>
<td>i2</td>
<td>Rd</td>
<td>rate of dark respiration</td>
<td>0.001</td>
<td>0.1</td>
</tr>
<tr>
<td>j2</td>
<td>V_{cmax}</td>
<td>velocity of carboxylation (umol/mol)</td>
<td>60</td>
<td>150</td>
</tr>
<tr>
<td>l2</td>
<td>Mobc</td>
<td>mobile carbon</td>
<td>75</td>
<td>200</td>
</tr>
<tr>
<td>f2</td>
<td>Q_{rd}</td>
<td>temperature dependence of Rd</td>
<td>0.5</td>
<td>2.5</td>
</tr>
<tr>
<td>k2</td>
<td>Rsoil(_{3})</td>
<td>soil respiration scaling co-efficient (data set 3)</td>
<td>0.5</td>
<td>1.5</td>
</tr>
</tbody>
</table>

Notes: Both parameters and initial pool sizes were optimized conditional on the data constraints. Parameters are arranged in descending order of constraint (i.e., best-constrained parameters first to worst-constrained parameters last) to relate to Fig. 2. Abbreviations: SOM, soil organic matter; Lf, litterfall.
The strongest parameter correlations were between the basal rate and temperature dependence of root respiration (parameters 6 and \( q \), Table 2 and Fig. 4b) and between different parameters governing spring phenology (parameters \( I \) and \( I' \), Table 2 and Fig. 4b). Parameters that were poorly constrained (\( z-K_2 \), Fig. 4b) did not tend to show a better-defined correlation structure than parameters that were well constrained. This suggests that reducing correlations in the posterior parameter distributions does not imply a better-constrained model. The same is not true for parameter covariance, which was steadily reduced with the addition of each new data stream (Fig. 4c). Covariance scales the correlation by the standard deviation of the parameters, thus lowering the weight of parameters that have well-constrained posterior distributions. Parameters that were not well constrained when using all available data tended to show a strong covariance structure (Fig. 4d).

Well-constrained parameters had limited covariance, even though some were highly correlated, reflecting the narrow range of variability for those parameters. This implies that using data relevant to these parameters could lead to a better-constrained model.

**The effect of improved parameterization on future projections**

Reduced model uncertainty under current climate conditions (Fig. 1) translated to reduced uncertainty in modeled future projections (Fig. 5). However, uncertainty in future projections of net ecosystem exchange was most reduced by the use of the few data streams that had the largest impact on model uncertainty under current climate conditions. Parameter-based uncertainty (i.e., without consideration of process-based uncertainty) as to whether the system could be projected to be a source or a sink for atmospheric carbon for the next 100 years was also significantly reduced.
parameters remained unconstrained (Fig. 2). This was despite the fact that 14 model-based prediction uncertainty for net ecosystem exchange streams led to only a minor reduction in parameter-based uncertainty in both short- and long-term projections (when using adequate data constraints), however, suggests that uncertainty due to parameter misspecification can be effectively eliminated, leaving process representation as the remaining source of uncertainty. This is highly beneficial in that a model with well-constrained parameters and narrow confidence intervals is much easier to falsify (or prove wrong) than one with poorly constrained parameters and large uncertainties. The evaluation of process error in long-term model projections is nontrivial (Keenan et al. 2011b, 2012b, Medlyn et al. 2011, Migliavacca et al. 2012), and may require observations of long-term ecosystem processes (Luo et al. 2011) in combination with manipulation experiments (Lezinger and Thomas 2011, Templer and Reinmann 2011).

**Parameter uncertainty**

One predominant goal of studies that aim to inform models with data is to identify model parameters. Early attempts in the field of terrestrial carbon cycling...
reported a limited number of parameters could be identified when using only eddy-covariance data (Wang et al. 2001, 2007, Knorr and Kattge 2005). Recent efforts using multiple constraints (Rayner 2010) report a much larger proportion of identifiable parameters. Richardson et al. (2010) reported 11 out of 12 parameters were well constrained when using six different data constraints with a simple model, while two studies (Weng and Luo 2011, Keenan et al. 2012b) constrained roughly half of the model parameters with comparatively more complex models. Here we show that improving parameter constraint is not solely a matter of using more data, but of selecting the correct data to use. Four of the available data sets (net ecosystem exchange, soil carbon turnover, soil respiration, woody litterfall) constrained 16 (64%) of the total parameters constrained (Fig. 2). Many parameters remained unconstrained even when using all data streams together. The fact that these parameters were not identifiable, while model projections were well constrained, may suggest that they are redundant in the current model structure (when accounting for parameter covariance; see next paragraph). Simplifying process representation for model aspects that cannot be parameterized could aid in
reducing the complexity of current models. Invoking “Occam’s razor” in this fashion (making models only as complex as justified by the data) would minimize the common problem of model over-fitting, and could be considered a necessary step to avoid the development of excessively complex models.

**Equifinality and parameter covariance**

“Equifinality” is defined as the situation where different parameter combinations or model structures can yield similar model performance (Beven 2006). In the case of parameters, equifinality can be detected by assessing correlation and covariance in posterior parameter distributions. Here we find that the level of equifinality depends on the number of different measurement types used to constrain the model. When using few data constraints, large equifinality allowed for divergent future projections of carbon cycling (Fig. 5). When using sufficient constraints, however, a lower level of equifinality was reached that did not prove detrimental to model performance and did not necessarily lead to an increase in model uncertainty over time. The model parameters that were least constrained tended to be those that had higher covariance (Fig. 4d). This implies that trade-offs between these parameters allowed the model to get equivalent results with varied parameter values.

Strong parameter correlations were observed for both well and poorly constrained parameters. For example, despite being very well constrained, parameters governing the basal respiration rates and temperature sensitivity of different soil organic matter layers were highly correlated (parameters $a$–$c$, Fig. 4b). Similarly, parameters controlling the rate of root turnover, and the size of the root carbon pool were correlated, with higher values of one compensated for by lower values in the other (parameters $o$ and $k$, Fig. 4b). Eight out of 14 parameters that were poorly constrained showed strong correlation with other parameters. The majority of these correlative pairs were with other parameters that were already relatively well constrained (i.e., all except pairings photosynthetic potential ($j_2$) with the fraction
of photosynthesis respired for maintenance (8, Fig. 4b). Some poorly constrained parameters were not correlated with other parameters (e.g., the soil respiration scaling parameter, $k_\text{s}$). In our analysis, the introduction of additional data constraints increased parameter correlations, implying that apparently uncorrelated parameters may have high-dimensional parameter relationships that are not detected by simple 1:1 correlative analysis (Richardson and Hollinger 2006, Trudinger et al. 2009, Ricciuto et al. 2011). Strong posterior parameter correlation is often interpreted as an indicator that the constraining data were not sufficient to distinguish between counteracting processes in the model (e.g., Ricciuto et al. 2011). Here we show that strong, non-detrimental correlations can persist even in a well-constrained model, and may be an inevitable consequence of model structure. This correlation is not necessarily reduced by the use of additional data. Parameter covariance, however, was continuously reduced with the use of additional data.

**What data are most useful?**

Previous studies have demonstrated the success of using additional data streams in conjunction with eddy-covariance flux data to improve estimates of ecosystem carbon exchange at different time scales (e.g., Williams et al. 2005, Xu et al. 2006, Moore et al. 2008, Richardson et al. 2010, Ricciuto et al. 2011, Weng and Luo 2011, Keenan et al. 2012b). The majority of studies emphasized the combination of stocks with fluxes, though no guidance is available as to what is the most appropriate or informative data to use. Our results show highly informative measurements at both ends of the cost-of-acquisition spectrum (e.g., senescence dates or leaf litterfall vs. eddy covariance or soil carbon turnover times). Coarse (woody) litterfall and leaf litterfall are often-overlooked measurements, but are ranked highly here. The results also show that some measurements, which have been the focus of much interest, are of low relative importance for modeling the carbon cycle. It should be kept in mind that we have not included all measurements that can possibly be made. Other measurements could include, for example, nonstructural carbohydrate reserves, nutrient stoichiometry, leaf-angle distributions, transfer rates between carbon pools, bole respiration, and so forth. All data sources are almost never available at the same site, but studies using synthetic data could be performed by those interested in quantifying the relative value of different data (e.g., for proposed measurement campaigns).

The weight assigned to each measurement potentially has a large impact on the ranking of different data. In our optimization framework, we chose to weight each data stream equally, independent of the number of observations, to ensure that the optimization did not favor model performance for one aspect of the ecosystem over another. We also weight each data stream by its associated uncertainty to account for the quality of the information contained therein. This choice, however, could affect the ranking of data streams. Other alternatives include giving each measurement equal weight, instead of each data stream. The problem boils down to information content: theoretically, an observation should be given weight relative to the information it contributes to the optimization. When using multiple constraints, the problem of quantifying the relative information is well exemplified by, say, quantifying the contribution of one estimate of soil carbon, compared to one half-hourly measurement of net ecosystem carbon exchange. This is particularly relevant when using high-frequency measurements of net ecosystem exchange – given 10 000 estimates of net ecosystem exchange, one additional NEE estimate does not necessarily contribute new information, while one estimate of the soil carbon stock does. Our chosen approach is in keeping with the philosophy that a model should predict all observations within measurement uncertainty, independent of the number of measurements available. Clearly, a detailed assessment of the real information content of observations, and an associated scheme for adequately weighting different data streams is an area in need of much research.

Turnover times of soil carbon pools have been suggested to be of utmost importance for accurately modeling the carbon cycle (Matamala et al. 2003, Strand et al. 2008, Gaudinski et al. 2010, Richardson et al. 2010). They have been inferred by model inversion approaches (Barrett 2002, Luo et al. 2003, Xu et al. 2006, Zhou and Luo 2008, Zhang et al. 2010), though measurements are rarely available to test different model structures and parameterizations (but see Gaudinski et al. 2009, Riley et al. 2009). Here we show that, after net ecosystem carbon exchange, turnover rates of the different soil carbon pools have the largest impact for improving model performance. Turnover times of different soil carbon pools (e.g., Gaudinski et al. 2010) and non-structural carbohydrate reserves (Richardson et al., in press), are not commonly available for model testing and should greatly aid in generating better-informed models in the future.

**Conclusions**

Financial resources in the field of earth system science are highly limited, and field campaigns expensive, so it is imperative to identify what measurements are of most use for a specific question. Here we presented results using a method by which to quantify the value of a diverse range of ecological data for improving models of the terrestrial carbon cycle. Using a hierarchical framework, we showed that relatively few data streams contribute to the largest reduction in uncertainty in model performance. In the presence of these data streams, which are distributed across the cost-of-acquisition spectrum, other measurement sources become redundant. For example, bud-burst dates, and carbon stock sizes, were of relatively little value for
constraining model performance in the presence of more informative measurements. Our results highlight the importance of estimates of carbon-stock turnover times, in conjunction with soil respiration and net ecosystem carbon-exchange measurements. These data sources should be given priority in future efforts. Using this framework together with information on the cost of measurement acquisition would help project managers to develop more efficient and effective measurement campaigns.

Acknowledgments

Carbon flux and biometric measurements at HFEMS have been supported by the Office of Science (BER), U.S. Department of Energy (DOE), and the National Science Foundation Long-Term Ecological Research Programs. T. F. Keenan, A. D. Richardson, and J. W. Munger acknowledge support from NOAA’s Climate Program Office, Global Carbon Cycle Program, under award NA11OAR4310054. T. F. Keenan and A. D. Richardson acknowledge support from the Northeastern States Research Cooperative, and the Office of Science (BER), U.S. Department of Energy, through the Northeastern Regional Center of the National Institute for Climatic Change Research. We especially thank the many participants who have sustained the long-term data collection, and in particular the summer students engaged in collecting field data, supported by NSF Research Experience for Undergraduates (REU) program, and the Harvard Forest Woods Crew for logistical and maintenance support. The computations in this paper were run on the Odyssey cluster supported by the FAS Science Division Research Computing Group at Harvard University.

Literature Cited


