Probabilistic inversion of a terrestrial ecosystem model: Analysis of uncertainty in parameter estimation and model prediction

Tao Xu,[1] Luther White,[2] Dafeng Hui,[1,3] and Yiqi Luo[1]

Received 26 January 2005; revised 2 January 2006; accepted 20 January 2006; published 5 May 2006.

[1] The Bayesian probability inversion and a Markov chain Monte Carlo (MCMC) technique were applied to a terrestrial ecosystem model to analyze uncertainties of estimated carbon (C) transfer coefficients and simulated C pool sizes. This study used six data sets of soil respiration, woody biomass, foliage biomass, litterfall, C content in the litter layers, and C content in mineral soil measured under both ambient CO_2 (350 ppm) and elevated CO_2 (550 ppm) plots from 1996 to 2000 at the Duke Forest Free-Air CO_2 Experiment (FACE) site. A Metropolis-Hastings algorithm was employed to construct a posterior probability density function (PPDF) of C transfer coefficients on the basis of prior information of model parameters, model structure, and the six data sets. The constructed PPDFs indicated that the transfer coefficients from pools of nonwoody biomass, woody biomass, and structural litter were well constrained by the six data sets under both ambient and elevated CO_2. The data sets also gave moderate information to the transfer coefficient from the slow soil C pool. However, the transfer coefficients from pools of metabolic litter, microbe, and passive soil C were poorly constrained. The poorly constrained parameters were attributable to either the lack of experimental data or the mismatch of timescales between the available data and the parameters to be estimated. Cumulative distribution functions were constructed for simulated C pool sizes on the basis of the six data sets, showing that on average the ecosystem would store 16,616 g C m^{-2} at elevated CO_2 by the year 2010, significantly higher than 13,426 g C m^{-2} at ambient CO_2 with 95% confidence. This study shows that the combination of a Bayesian approach and MCMC inversion technique is an effective method to synthesize information from various sources for assessment of ecosystem responses to elevated CO_2.

Citation: Xu, T., L. White, D. Hui, and Y. Luo (2006), Probabilistic inversion of a terrestrial ecosystem model: Analysis of uncertainty in parameter estimation and model prediction, *Global Biogeochem. Cycles*, *20*, GB2007, doi:10.1029/2005GB002468.

1. Introduction

[2] The prevention of dangerous anthropogenic interference with climate system requires quantification of carbon (C) sinks in land and ocean. The latest Intergovernmental Panel of Climate Change (IPCC) reports that the terrestrial C sink will continue to sequester 5–10 Gt ($\times 10^{15}$ g) C per year by the end of the 21st century [*Houghton et al.*, 2001]. This range is estimated by mostly using terrestrial biosphere models – a major tool developed in the past decades to describe terrestrial C cycles [e.g., *Parton et al.*, 1987; *Luo and Reynolds*, 1999; *Cramer et al.*, 2001; *McGuire et al.*, 2001]. Although these models are extensively used to predict C sequestration in terrestrial ecosystems, uncertainty in association with model parameters and predictions has not been carefully analyzed. If the uncertainty issue is not adequately addressed, C sink potentials cannot be fully understood. Some of the C sinks may be underestimated while others may be overestimated, even to the extent that contradictory results may appear. In such situation, policies based on current understanding to stabilize CO_2 concentrations will fall short in meeting targets of environmental mitigation.

[3] Having realized the importance of uncertainty analysis on policy making, the global change research community has recently directed considerable attention to studying the stochasticity and uncertainty in ecosystem processes and effects of various sources of randomness on prediction of ecosystem changes [*Clark*, 2005; *Murphy et al.*, 2004; *Dose and Menzel*, 2004; *Forest et al.*, 2002; *Wang et al.*, 2001]. Expert-specified probability density function (PDF) [e.g., *Murphy et al.*, 2004] has been used to quantify key uncertain properties of climate change simulations. The Bayesian paradigm has been introduced to incorporate a priori probabilistic density functions (PDF) with measurements to generate *a posteriori* PDFs for parameters of ecosystem models [*Braswell et al.*, 2005; *Knorr and Kattge*,

[1]Department of Botany and Microbiology, University of Oklahoma, Norman, Oklahoma, USA.
[2]Department of Mathematics, University of Oklahoma, Norman, Oklahoma, USA.
[3]Department of Biology, Duke University, Durham, North Carolina, USA.

0886-6236/06/2005GB002468$12.00

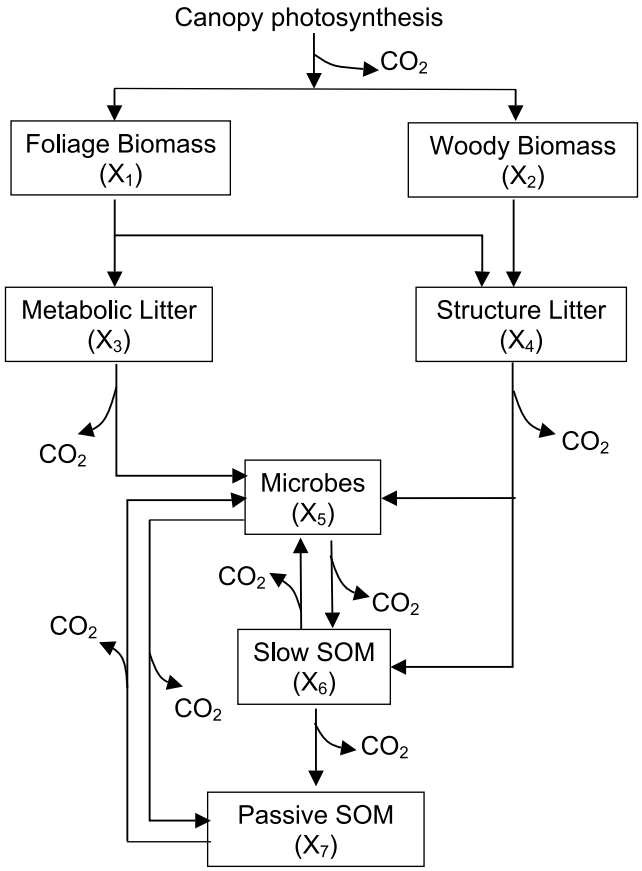

Figure 1. Diagram of the carbon process of the Duke Forest FACE site on which model equation (1) is based. The model has seven pools, and therefore the matrix A in section 2.1 is 7×7, vector B is 7×1, and vector X is 7×1. There are seven transfer coefficients, c_1, c_2, ... c_7, connecting the seven pools (Table 1). SOM stands for soil organic matter.

2005]. With a probabilistic approach, *Mastrandrea and Schneider* [2004] presented a cumulative probability function (CDF) to assess dangerous anthropogenic interference and showed its utility by applying it to analysis of uncertainty in model predictions of future changes. On a global scale, the Bayesian approach has been applied to constrain parameters in biosphere models against atmospheric CO_2 concentration data and to assess the biosphere C fluxes and uncertainties [*Kaminski et al.*, 2002; *Rayner et al.*, 2005].

[4] This study was designed to assess uncertainty in parameter estimation and model prediction with a terrestrial ecosystem (TECOS) model. The model was applied to the Duke Forest ecosystem, where a Free-Air CO_2 Experiment (FACE) has been in progress since August 1996, with a deterministic inversion for parameter estimation in a previous study [*Luo et al.*, 2003]. In this study, we conducted probabilistic inversion within a Bayesian framework by using the same data sets and the same model to facilitate a methodological comparison with the deterministic inversion. Within the Bayesian framework, the measurements were treated as random variables with certain probability distributions. A joint probability density function (PDF) was constructed for model parameters to analyze information content of observed data sets. Samples were taken from the joint PDF using a Markov chain Monte Carlo (MCMC) technique, which is appropriate for sampling high-dimensional PDFs of model parameters and widely used in inverse problems in engineering and geosciences [e.g., *Andersen et al.*, 2003; *Dosso and Wilmut*, 2002; *Oh and Kwon*, 2001; *Geman and Geman*, 1984]. The samples were used to construct marginal distributions for model parameters, to calculate parameter correlations, and to make CDFs for simulated pool sizes in forward modeling.

2. Methods

2.1. Carbon Cycling Model and Data Sources

[5] The model that we used in the study is a terrestrial ecosystem (TECOS) model that is a variation of the CENTURY model developed by *Parton et al.* [1987, 1988]. The TECOS model has a seven-pool compartmental structure (Figure 1) and has been applied to the Duke Forest FACE to study C sequestration process [*Luo et al.*, 2003]. In the model, C enters the ecosystem via canopy photosynthesis and is partitioned into nonwoody and woody biomass. Dead plant material goes to metabolic and structural compartments and is decomposed by microbes. Part of the litter C is respired and the rest is converted to soil organic matter (SOM) in slow and passive soil C pools. Carbon transfer coefficients are rate variables that quantify amounts of C per unit mass leaving each of the pools per day (Table 1). The inverses of the transfer coefficients are the mean C residence times, which are the key parameters measuring the C sequestration capacity of the ecosystem when combined with primary production [*Barrett*, 2002; *Luo et al.*, 2003]. Mathematically, the model is given by the following first-order ordinary differential equation:

$$\begin{aligned} \frac{dX(t)}{dt} &= \xi(t)ACX(t) + BU(t) \\ X(0) &= X_0, \end{aligned} \tag{1}$$

where $X(t) = (X_1(t), X_2(t), \ldots X_7(t))^T$ is a 7×1 vector describing C pool sizes, A and C are 7×7 matrices given by

$$A = \begin{pmatrix} -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0.712 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0.288 & 1 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0.45 & 0.275 & -1 & 0.42 & 0.45 \\ 0 & 0 & 0 & 0.275 & 0.296 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0.004 & 0.03 & -1 \end{pmatrix}$$

$$C = diag(c) \tag{2}$$

where $diag(c)$ denotes a 7×7 diagonal matrix with diagonal entries given by vector $c = (c_1, c_2, \ldots, c_7)^T$, components c_j, $(j = 1, 2, \ldots 7)$ represent C transfer coefficients associated with pool X_j, $(j = 1, 2, \ldots 7)$ (Table 1), $B = (0.25\ 0.30\ 0\ 0\ 0\ 0\ 0)^T$ is a vector that

Table 1. Description of Carbon Transfer Coefficients Among Carbon Pools Shown in Figure 1

Carbon Transfer Coefficients, g C g^{-1} d^1	Description
c_1	From pool "foliage biomass" (X_1) to pools "metabolic litter" (X_3) and "structure litter" (X_4)
c_2	From pool "woody biomass" (X_2) to pool "structure litter" (X_4)
c_3	From pool "metabolic litter" (X_3) to pool "microbes" (X_5)
c_4	From pool "structure litter" (X_4) to pools "microbes" (X_5) and "slow SOM" (X_6)
c_5	From pool "microbes" (X_5) to pools "slow SOM" (X_6) and "passive SOM" (X_7)
c_6	From pool "slow SOM" (X_6) to pools "microbes" (X_5) and "passive SOM" (X_7)
c_7	From pool "passive SOM" (X_7) to pool "microbes" (X_5)

partitions the photosynthetically fixed C to nonwoody biomass and woody biomass, $\xi(.)$ is a scaling function accounting for temperature and moisture effects on C decomposition, $U(.)$ is system input of photosynthetically fixed C given by a canopy photosynthetic model, and X_0 is an initial condition.

[6] This study used six data sets: foliage biomass growth, woody biomass growth, litterfall, C content in the litter layers, C content in mineral soil, and soil respiration, collected from year 1996 to 2000 at the Duke Forest, North Carolina, USA, where FACE has been in progress since 1996 [*Luo et al.*, 2003]. Measurement methods of the six data sets were described in papers by *DeLucia et al.* [2002], *Finzi et al.* [2001], *Schlesinger and Lichter* [2001], and *Andrews and Schlesinger* [2001]. The experiment was set on a 15-year-old loblolly pine plantation with six plots, each with a size of 30 m in diameter. The CO_2 concentration in the three treatment plots has been maintained at 200 ppm above ambient, and the other three control plots have been fumigated with ambient air [*Hendrey et al.*, 1999]. The initial pool size $X_0 = (469\ 4100\ 64\ 694\ 123\ 1385\ 923)^T$ was based on experimental data at the start of the FACE experiment (year 1996). The photosynthetically fixed C inputs $U(.)$ at both ambient CO_2 and elevated CO_2 were estimated with the mechanistic canopy model MAESTRA [*Luo et al.*, 2001] for the period 1996–2000 (Figure 2). The cumulative C inputs simulated by the MAESTRA model from year 1996 to 2000 are 6535 g C m^2 and 8823 g C m^2 for ambient and elevated CO_2 respectively, making a cumulative difference of about 2288 g C m^2 over a five year period.

[7] In the study, equation (1) was numerically solved with a finite difference method to give C pool sizes at each time t. In line with the time steps in $\xi(.)$ and $U(.)$, time difference dt was set to one day. The observation mapping operator $\Phi = (\varphi_1^T, \varphi_2^T, \ldots, \varphi_6^T)^T$ maps the modeled pool sizes at time t to observations by $\Phi X(t)$, and

$$\begin{aligned}
\varphi_1 &= (0 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0)\\
\varphi_2 &= (0.75 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0)\\
\varphi_3 &= (0.75c_1 \quad 0.75c_2 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0)\\
\varphi_4 &= (0 \quad 0 \quad 0.75 \quad 0.75 \quad 0 \quad 0 \quad 0)\\
\varphi_5 &= (0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 1 \quad 1)\\
\varphi_6 &= (0.25c_1 \quad 0.25c_2 \quad 0.55c_3 \quad 0.45c_4 \quad 0.7c_5 \quad 0.55c_6 \quad 0.55c_7).
\end{aligned} \tag{3}$$

Each φ_i, $i = 1, 2, \ldots 6$ maps simulated values in the state space to one of the field observations as $\varphi_1 X(t)$ for woody biomass, $\varphi_2 X(t)$ for foliage biomass, $\varphi_3 X(t)$ for litterfall, $\varphi_4 X(t)$ for C in forest floor, $\varphi_5 X(t)$ for C in forest mineral soil, and $\varphi_6 X(t)$ for soil respiration. For example, φ_5 directly maps the modeled total C content in the three soil pools to the observations.

2.2. Application of Bayes' Theorem

[8] A complete description of Bayesian probabilistic inversion approach can be found in Appendix A. In the context of this study, Bayes' theorem states that the posterior probability density function (PPDF) $p(c|Z)$ of C transfer coefficients (i.e., model parameters c) can be obtained from prior knowledge of parameters c represented by a prior probability density function $p(c)$ and information contained in the six data sets represented by a likelihood function

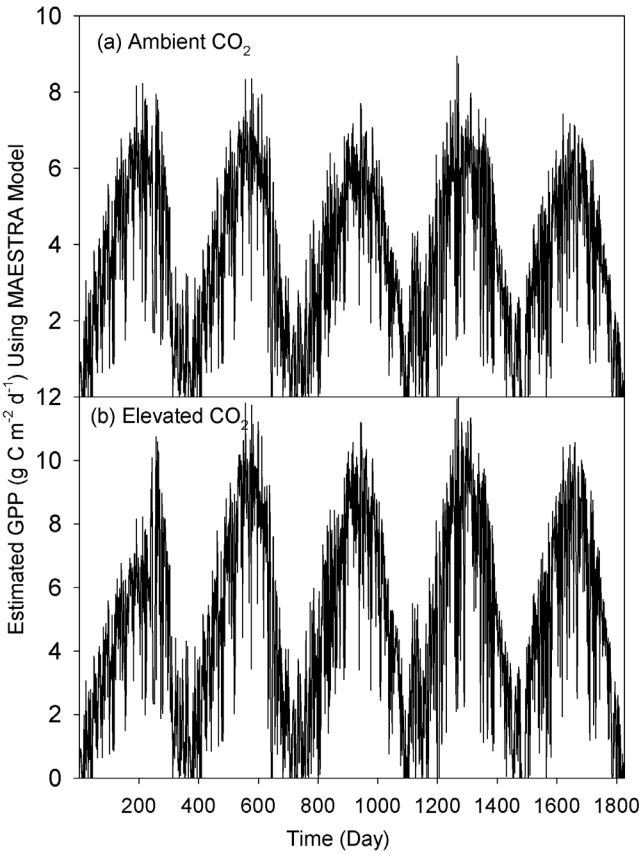

Figure 2. Simulated canopy photosynthesis (carbon input to the system) using MAESTRA model under ambient and elevated CO_2 from year 1996 to year 2000.

Table 2. Standard Deviation of Errors of the Six Data Sets and Normalizing Factors[a]

	Standard Deviations σ		Normalizing Factors *w*	
	Ambient	Elevated	Ambient	Elevated
Soil respiration (σ_1), g C m^{-2} d^{-1}	0.84	0.95	1.40	1.84
Woody biomass (σ_2), g C m^{-2}	377	490	284,304	481,943
Foliage biomass (σ_3), g C m^{-2}	35.0	45.6	2,453	4,159
litterfall (σ_4), g C $m^{-2}yr^{-1}$	49.4	100	4,894	20,345
Soil carbon (σ_5), g C m^{-2}	66	157	8,712	49,612
Mineral carbon (σ_6), g C m^{-2}	134	340	36,180	231,200

[a]Normalizing factors are from *Luo et al.* [2003]. Standard deviation σ is calculated from the normalizing factor *w* using $\sigma = \sqrt{w/2}$.

$p(Z|c)$. To apply Bayes' theorem, we first specified the prior PDF $p(c)$ by giving a set of limiting intervals for parameters c, then constructed the likelihood function $p(Z|c)$ on the basis of the assumption that errors in the observed data followed Gaussian distributions.

[9] The prior probability density function $p(c)$ of the parameters was specified as a uniform distribution over the following intervals:

$$\begin{aligned}
1.76\times10^{-4} &\le c_1 \le 2.95\times10^{-3}\\
5.48\times10^{-5} &\le c_2 \le 2.74\times10^{-4}\\
5.48\times10^{-3} &\le c_3 \le 2.74\times10^{-2}\\
5.48\times10^{-4} &\le c_4 \le 2.74\times10^{-3}\\
2.74\times10^{-3} &\le c_5 \le 6.85\times10^{-3}.\\
2.28\times10^{-5} &\le c_6 \le 2.84\times10^{-4}\\
1.37\times10^{-6} &\le c_7 \le 9.13\times10^{-6}
\end{aligned} \tag{4}$$

[10] These lower and upper limits were chosen as the same parameter limits over which the cost function of *Luo et al.* [2003] was minimized. In the Bayesian framework of this study, these limits are our prior knowledge about the approximate ranges of the parameters. We assumed a uniform distribution $p(c)$ for parameters c with an emphasis on the equal probability of all parameter values occurring in the limits. This may be the best prior to choose in the absence of any other knowledge regarding parameter distributions. The parameter space was defined as the product of the above intervals and denoted as Ω.

[11] The likelihood function was specified according to distributions of observation errors. Errors $e(t)$ in each observation $Z(t)$ at time t is expressed by

$$e(t) = Z(t) - \Phi X(t), \tag{5}$$

where $\varphi X(t)$ is the modeled value, which is a product of $X(t)$ from equation (1) and Φ from equation (2). For the six data sets used in this study, equation (5) is expanded as

$$e(t) = (e_1(t), e_2(t), \ldots, e_6(t))^T. \tag{6}$$

Corresponding to each of the data sets, there is one random error component $e_i(t) = Z_i(t) - \varphi_i X(t)$ with φ_i given in equation (2), $t \in obs(Z_i)$, and $obs(Z_i)$ being the sequence of observation times of the ith data set. We assumed that $e(t)$ followed a multivariate Gaussian distribution with a zero mean. This assumption is commonly made in many studies [*Braswell et al.*, 2005; *Raupach et al.*, 2005] mostly because a Gaussian distribution, in general, can well approximate errors of various sources due to the central limit theorem [*von Mises*, 1964]. With the Gaussian distribution, the probability density function of $e(t)$ at time t is given by

$$P(e(t)) \propto \exp\left\{-\frac{1}{2}[Z(t) - \Phi X(t)]^T \mathrm{cov}(e_t)^{-1}[Z(t) - \Phi X(t)]\right\}, \tag{7}$$

where $\mathrm{cov}(e_t)$ is a covariance matrix of vector $e(t)$. In the study, the nondiagonal elements in matrix $\mathrm{cov}(e_t)$ measuring error correlations are assumed nil while the diagonal elements specifying variances of the components of $e(t)$ were calculated from the normalizing factors of *Luo et al.* [2003] which were estimated from observations (Table 2). With an assumption that each component $e(t)$ being independently and identically distributed over the observation times, the likelihood function $p(Z|c)$ is then the multiplication of the distributions of $e_i(t)$, $i = 1,\ldots,6$ (equation (7)) at all observation times:

$$P(Z|c) \propto \exp\left\{-\sum_{i=1}^{6}\frac{1}{2\sigma_i^2}\sum_{t\in obs(Z_i)}[Z_i(t) - \varphi_i X(t)]^2\right\}, \tag{8}$$

where constants $\sigma_1^2, \sigma_2^2, \ldots, \sigma_6^2$ are the error variances of soil respiration, woody biomass, foliage biomass, litterfall, soil C and mineral C respectively (Table 2). Then, with Bayes' theorem, the PPDF of parameters c (Appendix A, equation (A2)) is given by

$$p(c|Z) \propto p(Z|c)p(c). \tag{9}$$

2.3. Sampling With the Metropolis-Hastings (M-H) Algorithm and Convergence Test

[12] The M-H algorithm is a Markov chain Monte Carlo (MCMC) technique revealing high-dimensional probability density functions of random variables via a sampling procedure [*Metropolis et al.*, 1953; *Hastings*, 1970; *Geman and Geman*, 1984; *Gelfand and Smith*, 1990]. To generate a Markov chain in the parameter space, we ran the M-H algorithm by repeating two steps: a proposing step and a moving step. In each proposing step, the algorithm generates a new point c^{new} on the basis of the previously accepted point $c^{(k-1)}$ with a proposal distribution $q(c^{new}|c^{(k-1)})$. In each moving step, point c^{new} is tested against the Metropolis criterion to examine if it should be accepted or rejected (see Appendix B for a detailed description of the M-H algorithm).

[13] The proposal distribution $q(c^{new}|c^{(k-1)})$ can strongly affect the efficiency of the M-H algorithm. To find an effective proposal distribution, we first made a test run of

the algorithm with 20,000 simulations, using a uniform proposal distribution centered at the currently accepted point: $c^{new} = c^{(k-1)} + r \times (c^{\max} - c^{\min})/D$, where r is a random number uniformly distributed between -0.5 and $+0.5$, $c^{\max}$ and $c^{\min}$ are the upper and lower limits of parameter vector c, D is a value controlling the proposing step size. This study set D = 5 so that the maximum step size is 1/10 of the range between the upper and lower limits of parameters c. Out of 20,000 simulations, the test run accepted about 1,200 updated samples. On the basis of the test run, we constructed a Gaussian distribution $N(0,\mathrm{cov}^0(c))$, where $\mathrm{cov}^0(c)$ is a diagonal matrix with its diagonal being set to the estimated variances of the parameters c from the initial test run and zeros elsewhere, and then we adopted the following proposal distribution to formally execute the consecutive MCMC simulations:

$$c^{new} = c^{(k-1)} + N\left(0, \mathrm{cov}^0(c)\right) \tag{10}$$

On the basis of equation (10), in each proposing step of the M-H algorithm a new point c^{new} is generated from its predecessor $c^{(k-1)}$ from a Gaussian distribution with mean $c^{(k-1)}$, constant variances estimated from the test run and zero parameter covariances.

[14] We formally made five parallel runs of the M-H algorithm with the proposal distribution in equation (10). The five runs started at dispersed initial points in the parameter space Ω and each run simulated 15,000 times. We monitored the trace plots of samples and calculated the running means and standard deviations of the parameters as simulation progressed. The initial number of samples (about 2,500 samples in the burn-in period) was discarded after the running means and standard deviations were stabilized. The acceptance rates for the newly generated samples were about 30 $\sim$ 40% for the five runs. For statistical analysis of the parameters, we used the union of the samples of the five runs (about 60,000 samples in total) after their burn-in periods.

[15] Theoretically, the M-H algorithm converges to a stationary distribution as guaranteed by the ergodicity theorem in Markov chain theory. In practice, the convergence of the sampling chains is often tested by the Gelman-Rubin (G-R) diagnostic method (Appendix C). In this study, we applied the G-R test and calculated the G-R statistics to examine the convergence of the five parallel runs. Only after the G-R test satisfied the convergence (G-R statistics approaches to 1) were the samples used for statistical inferences.

2.4. Parameter Estimation

[16] We estimated parameter statistics of maximum likelihood estimators (MLEs), means, and correlations on the basis of the union of the five-run samples. Histograms and cumulative distributions (CDFs) were constructed from the series of samples to display distributions of parameters in the parameter space Ω. Uncertainties of estimated parameters were quantified with a 95% highest-probability density interval – the interval of the minimum width containing 95% of the area of the marginal distribution. MLEs were made by observing the parameter values corresponding to the peaks of the marginal distributions. Means of parameters c_i ($E(c_i)$, i = 1, …, 7) were estimated by

$$E(c_i) = \frac{1}{k}\sum_{n=1}^{k} c_i^{(n)}, \tag{11}$$

where k is the number of samples given by the M-H algorithm. Correlations between parameters ($corr(c)$) were estimated by

$$corr(c) = \left(\frac{\mathrm{cov}(c_i, c_j)}{\sqrt{\mathrm{cov}(c_i, c_i)\mathrm{cov}(c_j, c_j)}}\right) i,j = 1, 2, \ldots 7, \tag{12}$$

where $\mathrm{cov}(c_i,c_j)$ is covariance between parameter c_i and c_j and estimated by

$$\mathrm{cov}(c_i, c_j) = \frac{1}{k}\sum_{n=1}^{k}\left[c_i^{(n)} - E(c_i)\right]\left[c_j^{(n)} - E(c_j)\right]. \tag{13}$$

Components in matrix $corr(c)$ are within -1 and $+1$. A value of $+1$ (-1) indicates perfect positive (negative) correlation and near-zero values indicates little correlation. By definition, the diagonal components are +1.

2.5. Simulated Pool Sizes

[17] We used Monte Carlo simulation to propagate the parameter uncertainty as expressed by PPDF $p(c|Z)$ (equation (9)) forward and constructed CDFs for simulated pools sizes using equation (1) with a time step set to one day. The model simulation was made over ten years from 2000 to 2010. Time courses of photosynthetic C input $U(.)$ and environmental scalar $\xi(.)$, from 1996 to 2000 were replicated two times from 2000 to 2010. At the end of the simulation (year 2010) we collected numerical solutions of equation (1) with input of 12,000 samples of $p(c|Z)$ in forward simulation. Cumulative distribution functions (CDFs) of $X(t)$ were constructed from the 12,000 numerical solutions to quantify uncertainty of C pool sizes.

3. Results

[18] Our inversion results are presented in Figure 3 for ambient CO_2 and Figure 4 for elevated CO_2. Figures 3a–3g and 4a–4g show 10,000 samples from the sampling series of the M-H simulation. Figures 3h–3n and 4h–4n show histograms of all 60,000 samples generated by the five runs since the five runs converged as indicated by the G-R statistic (Table 3). Figures 3o–3u and 4o–4u are cumulative distribution functions (CDFs) constructed from the histograms of each of the five runs for transfer coefficients c. At both ambient and elevated CO_2, parameters c_1, c_2, and c_4 were well constrained within their prespecified ranges (Figures 3, 4h, 4i, and 4k). Comparison of parameter distributions shows that parameter c_1 is much higher at elevated than ambient CO_2 (Figures 3 and 4h). Distributions of parameter c_2 were about the same at both elevated and ambient CO_2 (Figures 3 and 4i). In contrast, parameters c_3, c_5 and c_7 were poorly constrained (Figures 3, 4j, 4l, and 4n) at both CO_2 treatments. To examine their PPDFs in broader ranges, we decreased the lower limits defined in equation

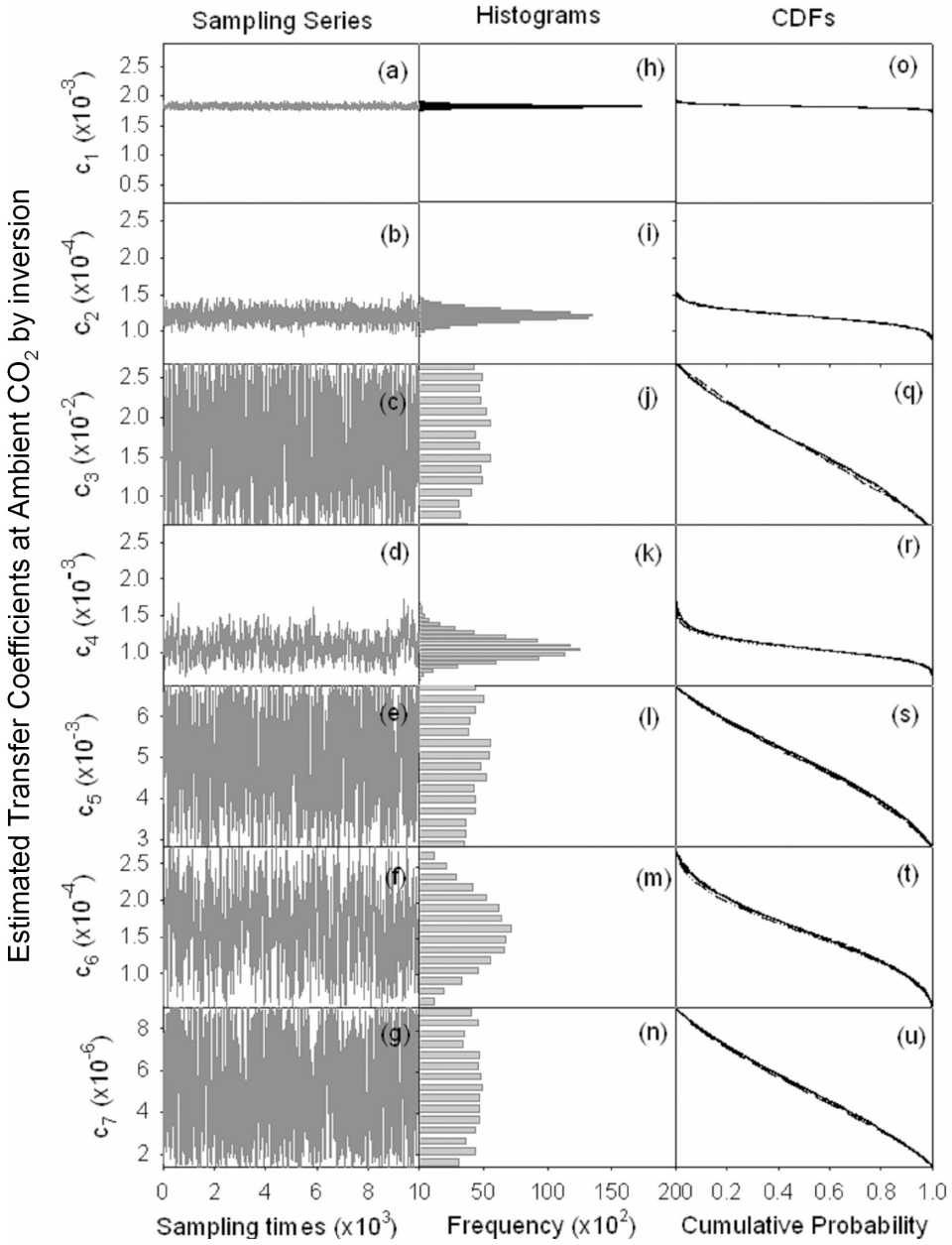

Figure 3. Inversion results under ambient CO_2 showing 60,000 samples from M–H simulation, the histograms of all samples from the five runs, and the CDFs constructed from each of the five runs. The *y* axes represent the prespecified limits of the parameters.

(4) by 1/5 and increased the upper limits in equation (4) by fivefold for parameters c_3, c_5 and c_7 at ambient CO_2. Similar to Figures 3j, 3l, and 3n, histograms in Figures 5c, 5e, and 5g still did not show statistically meaningful distributions.

[19] Histograms of parameter c_6 (Figures 3 and 4m) appear to contain more information on parameter constraint than parameters c_3, c_5 and c_7 but less than parameters c_1, c_2, and c_4 at ambient and elevated CO_2. The low information content on parameter c_6 may be caused partly by a limited number of data points for soil C content and partly by large variation of the soil C measurements. To increase the weight of the limited data points, we decreased the original variances of C contents in the forest floor and mineral soil by half (i.e., increased weighing factors of the two data sets in equation (8) by 100%) and reran the M-H algorithm to construct marginal distributions. Histograms of parameter c_6 with the reduced variances were much more concentrated than those with the original variances at both ambient and elevated CO_2 (Figure 6). No significant changes in marginal distributions of other parameters were observed (data not presented). The distribution of parameter c_6 was well constrained at ambient CO_2 with the reduced variances (Figure 6c) and edge hitting at elevated CO_2 with either the original or reduced variances (Figures 6b and 6d).

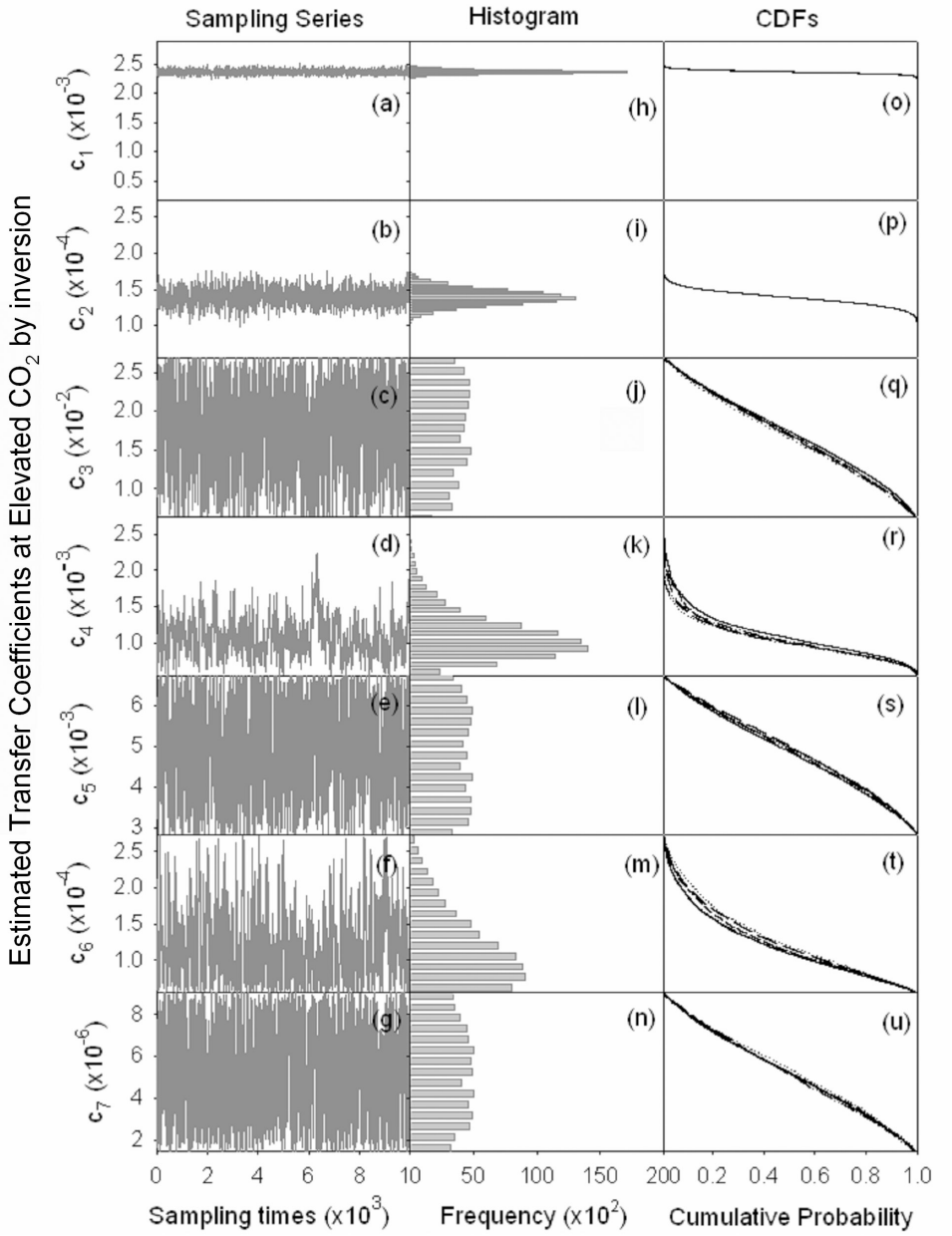

Figure 4. Inversion results under elevated CO_2 showing 60,000 samples from M–H simulation, the histograms of all samples from the five runs, and the CDFs constructed from each of the five runs. The *y* axes represent the prespecified limits of the parameters.

Elevated CO_2 shifted the marginal distribution downward in comparison to that at ambient CO_2 (Figure 6d versus Figure 6c).

[20] For parameters c_1, c_2, c_4 and c_6, the maximum likelihood estimates (MLEs, Table 3) were identified by observing the parameter values corresponding to the peaks of their marginal distributions (Figures 3, 4h, 4i, 4k, and 4m). There are no distinctive modes to indicate MLEs for parameters c_3, c_5 and c_7 (Figures 3, 4j, 4l, and 4n) and thus their MLEs were not identified (Table 3). Nevertheless, we were able to give the mean estimates for all parameters c_i, $i = 1,\ldots,7$ by calculating the sample means. The 95% probability intervals were estimated from the CDFs (Figure 3 and 4o–4u) to quantify parameter uncertainty (Table 3). Among the parameters, c_1 has the least variability relative to its range, followed by c_2, c_4 and c_6 (mostly symmetric with distinctive modes), while parameters c_3, c_5 and c_7 have the largest variability (widely spread marginal distributions). In general, the 95% confidence intervals cover estimated values by *Luo et al.* [2003] for all transfer coefficients except c_1 at elevated CO_2.

[21] Under both ambient and elevated CO_2, our cross-correlation analysis based on equation (12) showed that the seven parameters are not significantly intercorrelated with

Table 3. Maximum Likelihood Estimates (MLEs), Mean Estimates, 95% High-Probability Intervals (Lower Limit, Upper Limit), and G-R Statistics[a]

Parameters, g C g^{-1} d^{-1}	MLE	Mean	95% High-Probability Interval	G-R Statistics	*Luo et al.* [2003]
			Ambient		
c_1 ($\times 10^{-3}$)	1.82	1.82	(1.72, 1.89)	1.0	1.76
c_2 ($\times 10^{-4}$)	1.21	1.21	(0.99, 1.42)	1.0	1.00
c_3 ($\times 10^{-2}$)	NA	1.70	(0.66, 2.70)	1.0	2.15
c_4 ($\times 10^{-3}$)	1.04	1.04	(0.80, 1.34)	1.0	0.845
c_5 ($\times 10^{-3}$)	NA	5.10	(3.10, 6.85)	1.0	8.530
c_6 ($\times 10^{-4}$)	1.70	1.70	(0.55, 2.65)	1.0	0.898
c_7 ($\times 10^{-6}$)	NA	5.25	(1.51, 9.00)	1.0	3.1
			Elevated		
c_1 ($\times 10^{-3}$)	2.34	2.34	(2.25, 2.46)	1.0	2.17
c_2 ($\times 10^{-4}$)	1.25	1.25	(1.19, 1.52)	1.0	1.41
c_3 ($\times 10^{-2}$)	NA	1.71	(0.65, 2.71)	1.0	2.268
c_4 ($\times 10^{-3}$)	1.03	1.10	(0.50, 1.71)	1.0	0.965
c_5 ($\times 10^{-3}$)	NA	4.84	(2.90, 6.80)	1.0	2.534
c_6 ($\times 10^{-4}$)	0.55	0.66	(0.50, 2.40)	1.0	0.558
c_7 ($\times 10^{-6}$)	NA	5.19	(1.60, 9.00)	1.0	2.700

[a]As a comparison, the result of *Luo et al.* [2003] was also listed. NA means not available. The G-R statistics were calculated from the five sequences after the burn-in periods.

each other (Figure 7) except for the pair c_3 and c_4. Parameters c_3 and c_4 were negatively correlated with a correlation coefficient of -0.25 at ambient CO_2 and -0.15 at elevated CO_2.

[22] Under both ambient and elevated CO_2, the simulated and observed data sets using mean value estimates (Table 3) fitted closely with R^2 generally between 0.7 and 1, but mostly more than 0.8 (Figure 8). The fittings are similar to those shown by *Luo et al.* [2003]. Among the comparisons between the simulated values and observed data, large deviation existed between the simulated and observed foliage biomass (Figures 8c and 8d), probably due to both model assumptions and observation errors as discussed by *Luo et al.* [2003].

[23] Simulated C pool sizes in foliage biomass, woody biomass, structure litter, slow SOM, and passive SOM have symmetric distributions in 2010 (Figures 9a, 9b, 9d, 9f, and 9g). Two C pools: metabolic litter and microbes, have left-skewed distributions (Figures 9c and 9d). The CDFs under elevated CO2 were right shifted except for passive SOM, suggesting that elevated CO_2 increased C sequestration in the forest ecosystem. Table 4 lists means and 95% confidence intervals of C pool sizes in all the seven compartments. Mean C contents increased by 1.5% in the passive soil C pool and by 39.2% in the slow soil C pool. The simulated C content in the whole forest ecosystem increased by 23.8% by 2010. The 95% confidence intervals of simulated C pool sizes were significantly shifted to the right for woody biomass and for total C in the system (Figures 9b and 9h). However, the distributions of simulated C pool sizes in several compartments were statistically overlapped (Figures 9c, 9e, and 9g).

[24] In our forward simulation, we extended the input from year 2001 to year 2010 by repeating C input to the system twice. The cumulative difference of C input between the ambient and elevated CO_2 treatments over the fifteen year period from 1996 to 2010 is 6863 g C m^{-2}, the simulated cumulative difference of soil respiration over the same period is 3,657 g C m^{-2}, and the difference in pool sizes at the end of the simulation period (year 2010) is about 3,190 g C m^{-2} on the average (Table 4). Thus the extra C that system stored (3,190 g C m^{-2} on average) and released (3,657 g C m^{-2} via soil respiration) nearly match the extra C input (6863 g C m^{-2}) over the fifteen year period. Note that the match is not exact because of the fact that the total C amounts (Table 4) are mean estimations derived from the two CDFs in Figure 9h that were constructed from empirical data and thus may contribute estimation error.

4. Discussion

4.1. Probabilistic Versus Deterministic Inversion

[25] When a Gaussian type of error is used in probabilistic inversion, maximum likelihood estimates (MLEs) of parameters are equivalent to optimal estimates from deterministic inversion using the least squares (LS) method [*Tarantola*, 1987; *Raupach et al.*, 2005]. *Luo et al.* [2003] used a LS criterion and a Levenburg-Marquardt method coupled with Quasi Monte Carlo (LMQMC) to search for the C transfer coefficients c. In LMQMC, search directions are calculated using gradient vectors and approximated Hessian matrices of a cost function, and quasi Monte Carlo steps are used to find the step size that gives the largest decrease in the cost function along the search direction. A stopping criterion is set to terminate the algorithm when the cost function could not be reduced significantly. With the probabilistic inversion in this study, we exploited the same parameter space and observed data sets as in the work by *Luo et al.* [2003] to construct a PDF for parameters c, from which we derive statistical inferences (e.g., MLEs, means, and 95% confidence intervals) of c. The MLEs of the relatively well constrained parameters c_1, c_2, c_4 and c_6 are generally in good agreement with those by the deterministic inversion as done by *Luo et al.* [2003] except for c_6 in the ambient CO_2 (Table 3). The well-constrained parameters in the probabilistic inversion are those parameters to which the cost function in deterministic inversion was mostly sensitive.

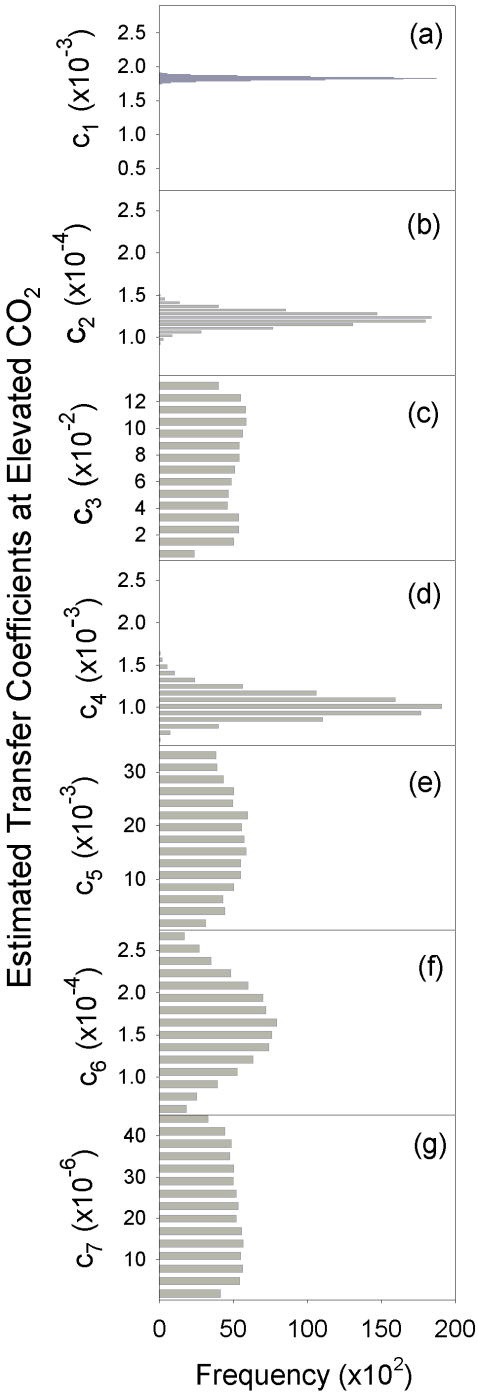

Figure 5. Marginal distributions of parameters c where the lower limit of parameters c_3, c_5, and c_7 are reduced by 1/5 and the upper limits are increased fivefold at ambient CO_2.

The MLEs of parameters c_3, c_5, c_7 were not comparable to those estimated by the deterministic inversion since they could not be uniquely determined in this study. However, the PDFs of the poorly constrained parameters in the probabilistic inversion offer broad 95% confidence intervals, which cover the optimal estimates by *Luo et al.* [2003]. Different estimates of parameter c_6 at ambient CO_2 between the probabilistic and deterministic inversions most likely

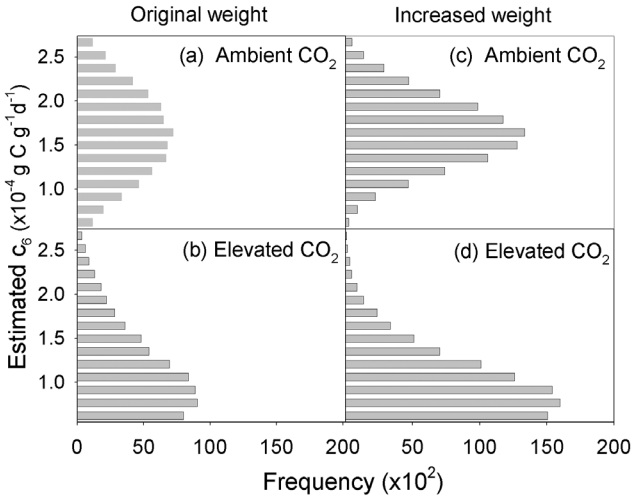

Figure 6. Sensitivity of marginal distribution to reduced error variances: (a and b) marginal distributions of c_6 with original variances and (c and d) marginal distributions constructed using reduced error variances of forest floor carbon data set and mineral carbon data set.

resulted from that the LMQMC had not updated its initial value significantly before the stopping criterion ended the search.

[26] The probabilistic approach employed by this study is advantageous over the deterministic approach of *Luo et al.* [2003] in at least three aspects: First, the probabilistic inversion constructs parameter distributions (such as in

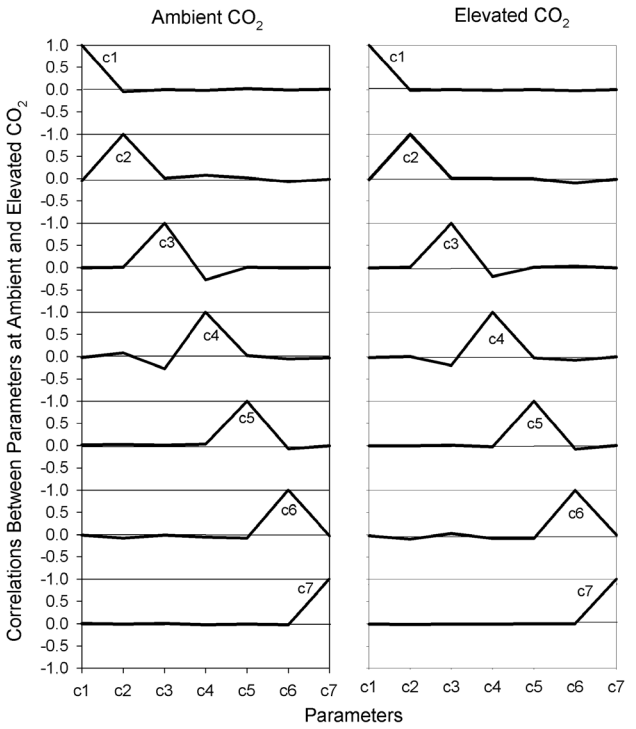

Figure 7. Correlations among model parameters c_1, c_2, . . ., c_7 under ambient and elevated CO_2.

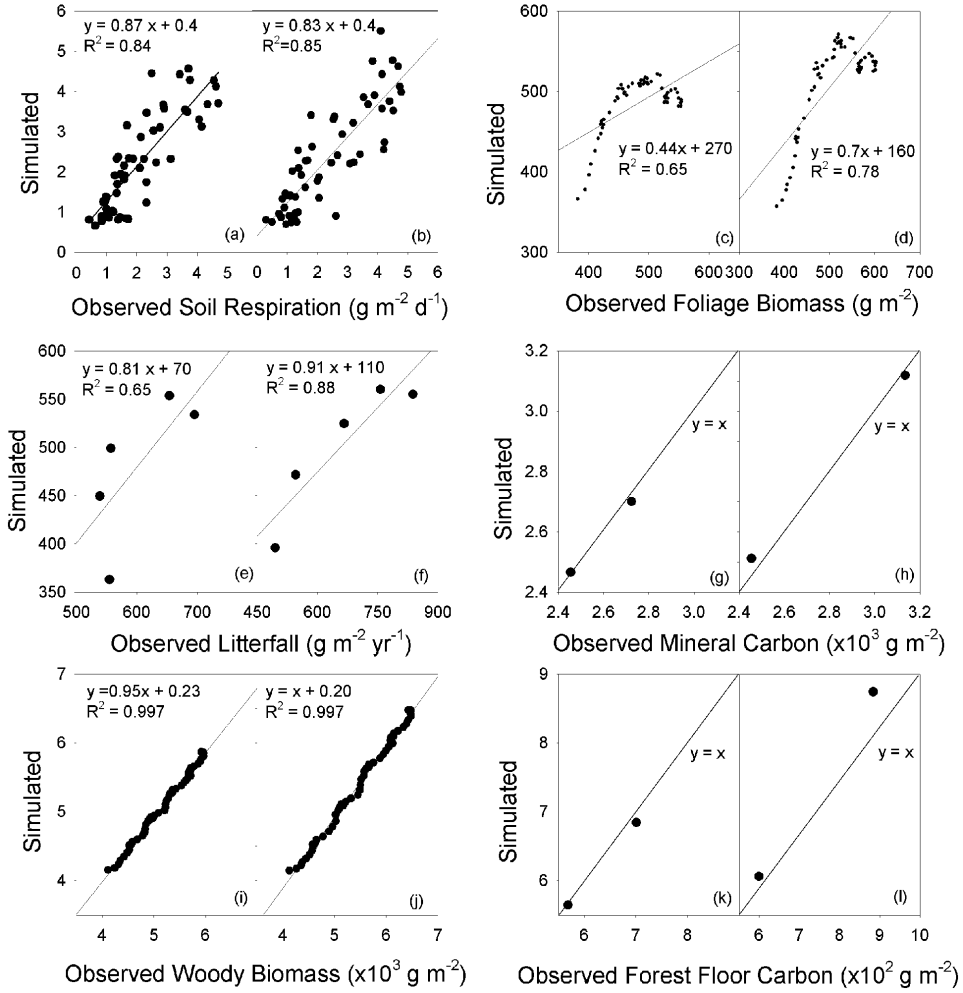

Figure 8. Comparison between the simulated data sets and the observed data sets under both ambient CO_2 and elevated CO_2. For each pair of plots, the left plot shows the matching at ambient CO_2, and the right plot shows the case at elevated CO_2.

Figures 3 and 4) while the deterministic inversion provides only point estimates. The parameter distributions can be used to quantify MLEs, means, and confidence intervals, and thus offer much richer information than the point estimates by the deterministic inversion. Second, the probabilistic approach reveals whether a parameter is well constrained or not (e.g., parameter c_7 versus c_1) whereas the deterministic inversion could not. From the degree to which parameters are constrained by data, we can assess parameter uncertainties as measured, for example, by confidence intervals. Third, the probabilistic approach can readily analyze correlations among parameters (e.g., Figure 7) while the deterministic inversion may not be always able to reveal such information. The probabilistic approach analyzes parameter correlations from the sampling series. The deterministic optimization approach usually estimates optimal parameter values without quantifying their correlations [e.g., *Barrett*, 2002; *Luo et al.*, 2003]. Although some applications estimated correlations among parameters from Hessian matrix [e.g., *Wang et al.*, 2001] at the optimal point, analytic solutions of Hessian matrices may not be easily obtainable, especially when models are given as a set of differential equations.

4.2. Constraints of Parameters by Data Sets

[27] The nature of inverse analysis is to exploit information content contained in data, model structure, and prior knowledge on parameters [*Raupach et al.*, 2005]. The six data sets used in this study contain enough information to constrain C transfer coefficients of nonwoody biomass, woody biomass, structural litter, and slow soil C (c_1, c_2, c_4 and c_6), but not enough for metabolic litter, microbial, and passive soil C (c_3, c_5 and c_7). The lack of microbial biomass data in this study may cause large uncertainty of c_5. We did an exercise by using modeled microbial biomass data (i.e., a virtual data set) in the inverse analysis. Parameter c_5 became well constrained (data not presented). That suggests that microbial biomass data are crucial in future inverse analysis. Parameter c_3 is the transfer coefficient from the metabolic litter pool, which is small and turnovers

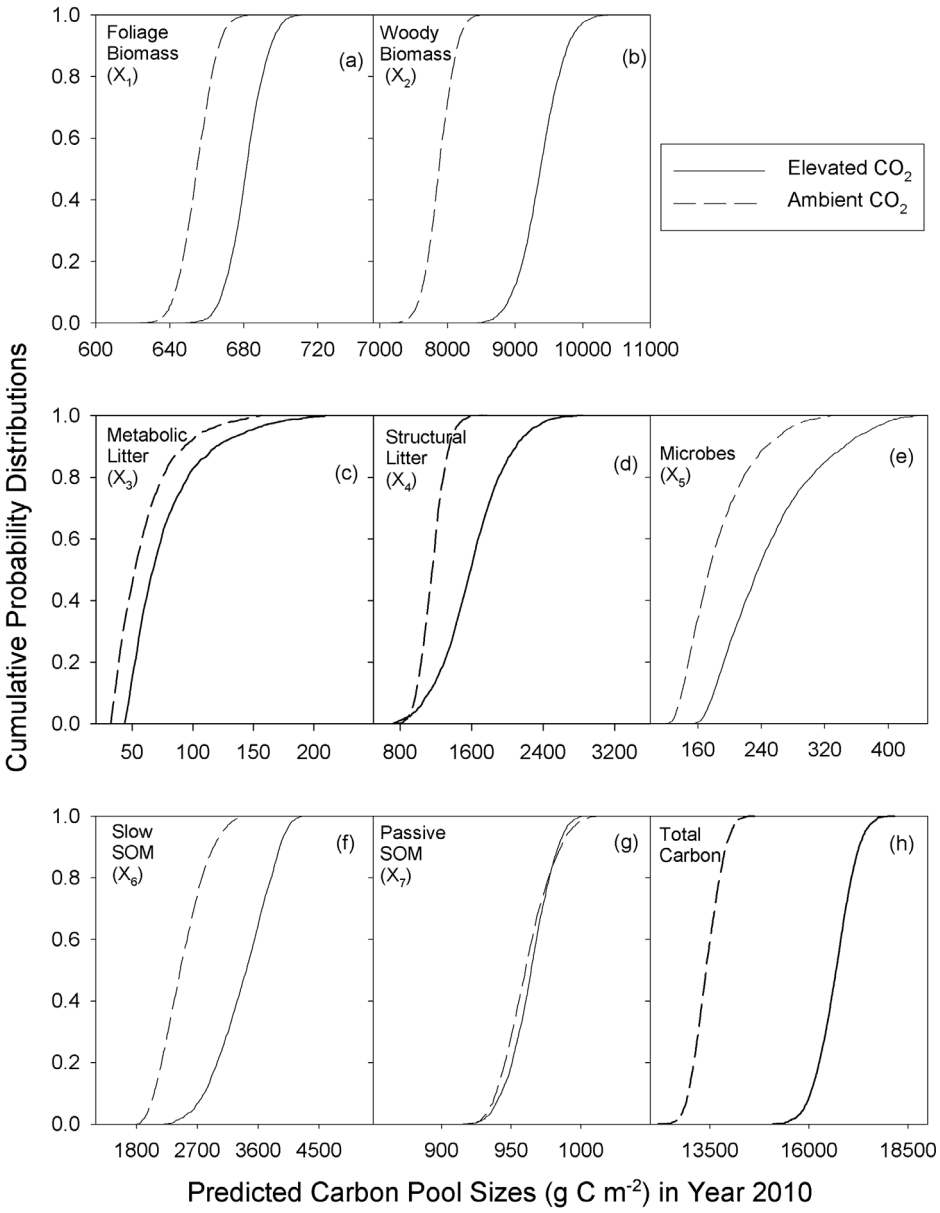

Figure 9. CDFs of simulated carbon pool sizes of year 2010.

fast. Although the concept of metabolic litter may be important in ecology [*Berg and McClaugherty*, 2003], we did not have data to constrain the transfer coefficient c_3 from the pool. We may have to merge this pool with the structural litter pool in future inverse analysis unless data of labile litter compounds become available. Also, seen from Figure 7, parameters c_3 and c_4 show posterior correlation, which indicates the information content of litterfall measurement is not sufficient to separate these coefficients, and therefore they cannot be constrained separately.

[28] Parameter c_7 describes C transfer from the passive soil organic matter pool, which has a residence time of hundreds or thousands of years. It can be hardly constrained by short-term observation. However, this pool is very critical to simulate long-term C dynamics in terrestrial ecosystems [*Parton et al.*, 1987]. We may explore C isotope data to constrain this parameter in a future study.

[29] Parameter correlations are part of the information revealed by the inverse analysis, which likely reflect relationships defined by model structure, correlations among data, or errors, or any combinations of the three. This study identified only one negative correlation between parameter c_3 and c_4 (-0.25 and -0.15). The negative correlation may suggest a complementary relationship between C transfer rates of metabolic litter and structure litter and is physically reasonable since the total litter amount was partitioned into the two pools. An increase in parameter c_3 is supposed to be accompanied with a decrease in parameter c_4 and vice versa. Unless we have data of labile versus structural components of litter,

Table 4. Summary Statistics for Simulated Carbon Pool Sizes for Year 2010

	Ambient CO_2		Elevated CO_2		
Pools, g C m^{-2}	Mean	95% Confidence Interval	Mean	95% Confidence Interval	Mean Increment of C Content, %
Foliage biomass (x_1)	656	(637, 672)	686	(662, 701)	4.6
Woody biomass (x_2)	7800	(7460, 8285)	9400	(8750, 10000)	20.5
Metabolic litter (x_3)	69	(32, 110)	77	(43, 142)	11.5
Structure litter (x_4)	1250	(900, 1460)	1700	(900, 2400)	36.5
Microbes (x_5)	210	(130, 300)	280	(160, 400)	33.3
Slow SOM (x_6)	2500	(1950, 3200)	3480	(2500, 4100)	39.2
Passive SOM (x_7)	954	(930, 992)	968	(932, 997)	1.5
Total Carbon	13426	(12700,14100)	16616	(15700,17550)	23.8

the inversion could not independently estimate these two parameters.

4.3. Data Properties and Parameter Uncertainties

[30] Data properties such as error distributions, cross correlations among multiple data sets, and the evolution of self-correlations or cross correlations with time are critical for evaluation of parameter uncertainties. In this study, we reduced the error variances of the forest floor C and the mineral C data sets by half to examine the sensitivity of model parameters to error variances. As shown in Figure 6, a reduction of the error variances substantially reduced the uncertainty of c_6. Our exercise demonstrated that error magnitudes in observations play an important role in determining parameter uncertainty. In general, error distributions determine the form of a likelihood function (equation (8)) and correspondingly PPDF $p(c|Z)$ (equation (9)). Although Gaussian distribution errors are often assumed in extant work (see *Raupach et al.* [2005] for a general discussion), other distributions such as skewed or lognormal may be more realistic for particular data. It is yet to examine key properties of uncertainty sources in association of those non-Gaussian distributions in the probabilistic inversion.

[31] It is a tremendously difficult task to obtain the properties of error distributions, cross correlations among multiple data sets, and the evolution of self-correlations or cross correlations with time in experimental observations. When they are not available as in most of the current studies, assumptions about uncertainty properties must be made, for example, with a constant σ across time, Gaussian distributions, or independent random errors among multiple data sets (e.g., *Braswell et al.* [2005] and this study) to proceed with inverse analysis. Currently there have been initial efforts toward specifying data properties related to key terrestrial C observations based on expert judgment [*Raupach et al.*, 2005], which are helpful in quantifying parameter and prediction uncertainties. In future experimental research, it is highly desirable to quantify those data properties from measurements.

4.4. Data-Model Fitness and Simulation of Pool Sizes

[32] Our probabilistic inversion improved the data-model fitness with higher R^2 values than that using the deterministic inversion (Figure 8 versus Figure 3 of *Luo et al.* [2003]). However, there were still plenty of unexplained variances (Figure 8). In particular, the systematic variation in foliage biomass was not explained by the model with parameter values estimated by the probabilistic inversion. *Luo et al.* [2003] suggested that a restricted search range for c_1 may be partially responsible for the systematic deviation. This study did allow the inversion to search for values of c_1 in a much broader range than that by *Luo et al.* [2003] and still did not make enough improvement of the fitness. The other reason of the discrepancy suggested by *Luo et al.* [2003] is the quality of foliage biomass data, which was indirectly estimated from diameter at breast height (DBH). It would be ideal to make direct measurements of foliage biomass if technique allows. The systematic deviation may also result from model structure, which may not accurately represent growth and senescence processes of foliage biomass. In addition, this inversion analysis used multiple data sets. We may have to explore various other setups to define the PPDF (e.g., varying weighing factors in equation (8) for different data sets) to improve the model ability to match the multiple data sets simultaneously.

[33] The predicted mean pool sizes were generally larger under elevated than ambient CO_2 (Table 4 and Figure 9). However, the 95% confidence intervals of most of the pools did not exhibit significant change except for the woody biomass pool. The CDFs for the passive SOM pool nearly coincided (Figure 9g), showing short-term simulation could not change the size of a long-term pool. The CDFs of simulated pool sizes were constructed by solving the model equation (1) with the sampling series. This approach incorporates information from the posterior parameter estimates, posterior correlations among parameters, and model structure into forward simulation. Even though parameters c_3, c_5 and c_7 were poorly constrained, simulated pool sizes for X_3, X_5 and X_7 are not uniformly distributed, suggesting meaningful information contained in the forward simulation. The information contained in the CDFs of forward simulation is derived from the model structure itself, in combination with the constrained parameters of other pools.

[34] Our simulation was solely based on existing knowledge of the uniform prior distribution over the limit intervals (equation (4)), the model structure (equation (1)) and the six measured data sets with the assumed Gaussian error properties (Table 2). The same data sets and the same model structure were used in this study to facilitate comparisons with results in paper by *Luo et al.* [2003]. However, there are more, longer data sets available at the Duke site, which will improve model projections. In addition, the C input $U(.)$ and environmental scalars $\xi(.)$ estimated from recorded micrometerological data during the period from 1996 to 2000 were extrapolated to the period from 2000 to 2010 by replicating the time series twice. Environmental conditions at the site are likely to be different during the two periods,

resulting in different $U(.)$ and $\xi(.)$. Other factors such as progressively limiting nutrient availability [*Finzi et al.*, 2006; *Luo et al.*, 2004] and forest stand development [*Hooker and Compton*, 2003] may further complicate projections. In spite of the fact that the quantitative results given by the forward model simulation can be improved, the approach of probabilistic inversion used in our study is very useful and informative for data-model integration in ecology.

5. Conclusions

[35] Using the Bayesian approach and a MCMC inversion technique in this study, we constructed probability distributions of the model parameters (Figures 3 and 4), made statistical estimates (Table 3), analyzed the correlations among the parameters (Figure 7), and developed cumulative probabilistic distributions of simulated pool sizes (Figure 9). Thus the probabilistic inversion provides much more informative outputs than the deterministic inversion. Our study showed that at both ambient and elevated CO_2, the transfer coefficients from pools of nonwoody biomass (c_1), woody biomass (c_2), structural litter (c_4), and slow soil C (c_6) were well constrained by the six data sets. In contrast, the transfer coefficients from pools of metabolic litter (c_3), microbe (c_5), and passive soil C (c_7) were poorly constrained. The simulated distributions of pool sizes indicated that elevated CO_2 stimulated C sequestration in the forest ecosystem. The 95% confidence intervals were significantly higher in the woody biomass and total ecosystem at elevated than ambient CO_2.

[36] The Bayesian approach offers a rigorous method to assess uncertainty of model predictions. Nevertheless, its applications to ecological research are still at an infant stage and yet to be developed. For example, correlations among model parameters due to model structure have to be appropriately accounted for in the probabilistic inversion. Uncertainties of the estimated parameters and model projections are sensitive to error variances (Figure 6), other data properties, and assumptions on forms of distributions. Although initial assessments have been made for properties of observational data related to inverse analysis of terrestrial C processes [*Raupach et al.*, 2005], a comprehensive understanding of uncertainty sources and data properties is required to rigorously carry out for the global C cycle.

Appendix A: Bayes' Theorem

[37] A general description of the Bayesian probabilistic inversion is given by Bayes' theorem [e.g., *Box and Tiao*, 1973; *Tarantola*, 1987; *Gill*, 2002; *Leonard and Hsu*, 1999] in a form of

$$p(c|Z) = \frac{p(Z|c)p(c)}{p(Z)}, \tag{A1}$$

where $p(c)$ is the prior probability density function (PDF) representing prior knowledge about parameters c, $p(Z|c)$ is the conditional probability density of observations Z on c (also called the likelihood function of parameters c), $p(Z)$ is the probability of observations Z, and $p(c|Z)$ is the posterior probability density function (PPDF) of parameters c. The theorem states that the posterior information of model parameters c represented by $p(c|Z)$ can be obtained from the prior information represented by $p(c)$ and the observed information given by $p(Z|c)$. $p(c|Z)$ is often written in the following form:

$$p(c|Z) \propto p(Z|c)p(c); \tag{A2}$$

that is, $p(c|Z)$ is proportional to $p(Z|c)p(c)$.

[38] From the Bayesian viewpoint, $p(c|Z)$ in (A1) represents the solution to an inverse problem since it gives a probabilistic description of parameters c over parameter space. The interpretation of $p(c|Z)$ leads to the following integrals:

$$E(c) = \int cp(c|Z)dc \tag{A3}$$

$$\mathrm{cov}(c) = \int (c - E(c))(c - E(c))^T p(c|Z)dc \tag{A4}$$

$$p(c_i|Z) = \int p(c|Z)dc_1 \ldots dc_{i-1}dc_{i+1} \ldots dc_m, \tag{A5}$$

which are the expected value, the covariance and the marginal distributions of c, respectively. These are some of the statistics describing parameter uncertainties.

Appendix B: Metropolis-Hastings Algorithm

[39] In practice, except for situations where $p(c|Z)$ have very simple forms, it is not always possible to draw samples easily from $p(c|Z)$ (as is the case in this study). The sampling problem had hindered the applications of the Bayesian theory for a long period of time in history until later solved by the MCMC techniques, after the foundational work of *Metropolis et al.* [1953], *Hastings* [1970], *Geman and Geman* [1984], and the synthesizing paper by *Gelfand and Smith* [1990]. The basic idea for the MCMC sampling is to design a Markov chain with $p(c|Z)$ as the targeted stationary distribution. Once the chain has simulated for sufficiently long period samples in the chain will follow the stationary distribution, then one can collect the samples from the simulation and calculate various statistics associated with the PPDF from them. One of the mostly used techniques for MCMC is the Metropolis-Hastings (M-H) algorithm, which is briefly described below.

[40] For simplicity of notation, we denote $L(c)$ as the targeted stationary distribution $p(c|Z)$. A computer implementation of the M-H algorithm consists the following steps: [*Spall*, 2003].

[41] Step 1: Choose an arbitrary initial point $c^{(0)}$ in the parameter space.

[42] Step 2: (Proposing step). Propose a candidate point c^{new} according to a proposal distribution $q(c^{new}|c^{(k-1)})$.

[43] Step 3: (Moving step). Calculate $P(c^{(k-1)},c^{new}) = \min\left\{1, \frac{L(c^{new})q\left(c^{(k-1)}|c^{new}\right)}{L(c^{k-1})q(c^{new}|c^{k-1})}\right\}$, and compare the value with a

random number U from the uniform distribution U[0, 1] that is defined on interval [0, 1]. Set $c^{(k)} = c^{new}$ if $U \leq P(c^{(k-1)}, c^{new})$; otherwise set $c^{(k)} = c^{(k-1)}$. This test criterion is also called the Metropolis criterion.

[44] Step 4: Repeat steps 2 and 3 until enough samples are obtained.

[45] In most applications, the proposal distribution $q(c^{new}|c^{(k-1)})$ is usually set as either a uniform distribution or a symmetric Gaussian distribution centered at the current point. The Gaussian distribution may also take into any prior knowledge about the parameters (e.g., estimated covariance) into account. The proposing efficiency of $q(c^{new}|c^{(k-1)})$ affects the efficiency of the algorithm, and hence should be properly designed to ensure a moderate sample-acceptance rate. *Robert and Rosenthal* [1998] indicated that a rate of 23% is sometimes an optimal acceptance rate. In practice, it is often desirable to make ''test runs'' of the algorithm and adjust parameters in the proposal distribution on the basis of the test run until the acceptance rate is approximately 23%. In general, the acceptance rate can be adjusted between $20 \sim 50\%$.

Appendix C: Convergence of MCMC

[46] Since the Markov chain generated by the M-H algorithm is reversible, the standard ergodicity theorem in Markov chain theory states that if it is irreducible and aperiodic, the chain converges to a unique stationary distribution [*Spall*, 2003]. This means samples $c^{(k)}$ as k becomes sufficiently large are draws from the stationary distribution and can be used to make statistical inferences for the random variable.

[47] There are various techniques for monitoring convergence of MCMC simulation in practice, for example, run several parallel chains and visually inspecting the trace plots and autocorrelation sequences, monitor the running means and standard deviations, or apply the Gelman-Rubin (G-R) diagnostic method. The idea of G-R test is that if the simulated Markov chain has reached convergence, the within-run variation should be roughly equal to the between-run variation [*Gelman and Rubin*, 1992]. Specifically, denoting for each parameter component c_i of vector c the samples from K parallel M-H runs of length N as $c_i^{n,k}$ ($n = 1, 2, \ldots, N$; $k = 1, 2, \ldots, K$), then the between and within-run variances are defined as

$$B_i = \frac{N}{K-1}\sum_{k=1}^{K}\left(\overline{c}_i^{\cdot,k} - \overline{c}_i^{\cdot\cdot}\right)^2$$

$$W_i = \frac{1}{K(N-1)}\sum_{k=1}^{K}\sum_{n=1}^{N}\left(c_i^{n,k} - \overline{c}_i^{\cdot,k}\right)^2. \quad \text{(C1)}$$

The G-R scale reduction statistics is given by

$$GR_i = \sqrt{\frac{W_i(N-1)/N + B_i/N}{W_i}}. \quad \text{(C2)}$$

Once convergence is reached GR_i should approximately equal one.

[48] **Acknowledgments.** We greatly appreciate the two anonymous reviewers for their insightful comments. This research was supported by grants from the Terrestrial C Program at the Office of Science (Biological and Environmental Research, or BER), U.S. Department of Energy (DE-FG03-99ER62800), from the National Institute of Global Environmental Change South Central Regional Center, and from the National Science Foundation (DEB 0092642 and DEB 0444518). Research at the Duke Forest FACE (Free-Air Carbon Dioxide Enrichment) facility was supported by the Office of Science (BER) program, U.S. Department of Energy.

References

Andersen, K., S. Brooks, and M. Hansen (2003), Bayesian inversion of geoelectrical resistivity data, *J. R. Stat. Soc., Ser. B*, *65*, 619–642.

Andrews, J. A., and W. H. Schlesinger (2001), Soil CO_2 dynamics in a temperate forest with experimental CO_2 enrichment, *Global Biogeochem. Cycles*, *15*, 149–162.

Barrett, D. J. (2002), Steady state turnover time of carbon in the Australian terrestrial biosphere, *Global Biogeochem. Cycles*, *16*(4), 1108, doi:10.1029/2002GB001860.

Berg, B., and C. McClaugherty (2003), *Plant Litter: Decomposition, Humus Formation, Carbon Sequestration*, Springer, New York.

Box, G. E. P., and G. C. Tiao (1973), *Bayesian Inference in Statistical Analysis*, Addison-Wesley, Boston, Mass.

Braswell, B. H., W. J. Sacks, E. Linder, and D. S. Schimel (2005), Estimating diurnal to annual ecosystem parameters by synthesis of a carbon flux model with eddy covariance net ecosystem exchange observations, *Global Change Biol.*, *11*, 335–355.

Clark, J. S. (2005), Why environmental scientists are becoming Bayesians, *Ecol. Lett.*, *8*, 2–14.

Cramer, W., et al. (2001), Global response of terrestrial ecosystem structure and function to CO_2 and climate change: Results from six dynamic global vegetation models, *Global Change Biol.*, *7*, 357–373.

DeLucia, E. H., K. George, and J. G. Hamilton (2002), Radiation-use efficiency for a forest exposed to elevated concentrations of atmospheric carbon dioxide, *Tree Physiol.*, *22*, 1003–1010.

Dose, V., and A. Menzel (2004), Bayesian analysis of climate change impacts in phenology, *Global Change Biol.*, *10*, 259–272.

Dosso, S. E., and M. J. Wilmut (2002), Quantifying data information content in geoacoustic inversion, *IEEE J. Oceanic Eng.*, *27*(2), 296–304.

Finzi, A. C., A. S. Allen, E. H. DeLucia, D. S. Ellsworth, and W. H. Schlesinger (2001), Forest litter production, chemistry, and decomposition following two-years of free-air CO_2 enrichment, *Ecology*, *82*, 470–484.

Finzi, A. C., et al. (2006), Progressive nitrogen limitation of ecosystem processes under elevated CO_2 in a warm-temperate forest, *Ecology*, *87*, 15–25.

Forest, C. E., P. H. Stone, A. P. Sokolov, M. R. Allen, and M. D. Webster (2002), Quantifying uncertainties in climate system properties with the use of recent climate observations, *Science*, *295*, 113–114.

Gelfand, A. E., and A. F. M. Smith (1990), Sampling-based approaches to calculating marginal densities, *J. Am. Stat. Assoc.*, *85*(410), 398–409.

Gelman, A., and D. B. Rubin (1992), Inference from iterative simulation using multiple sequences, *Stat. Sci.*, *7*, 457–511.

Geman, S., and D. Geman (1984), Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images, *IEEE Trans. Pattern Anal. Machine Intel.*, *6*, 721–741.

Gill, J. (2002), *Bayesian Methods—A Social and Behavioral Approach*, CRC Press, Boca Raton, Fla.

Hastings, W. K. (1970), Monte Carlo sampling methods using Markov chain and their applications, *Biometrika*, *57*, 97–109.

Hendrey, G. R., D. S. Ellsworth, K. F. Lewin, and J. Nagy (1999), A free-air enrichment system for exposing tall forest vegetation to elevated atmospheric CO_2, *Global Change Biol.*, *5*, 293–310.

Hooker, T. D., and J. E. Compton (2003), Forest ecosystem carbon and nitrogen accumulation during the first century after agricultural abandonment, *Ecol. Appl.*, *13*, 299–313.

Houghton, J. T., Y. Ding, D. J. Griggs, M. Noguer, P. J. van der Linden, and D. Xiaosu (Eds.) (2001), *Climate Change 2001: The Scientific Basis*, pp. 1–896, Cambridge Univ. Press, New York.

Kaminski, T., W. Knorr, P. J. Rayner, and M. Heimann (2002), Assimilating atmospheric data into a terrestrial biosphere model: A case study of the seasonal cycle, *Global Biogeochem. Cycles*, *16*(4), 1066, doi:10.1029/2001GB001463.

Knorr, W., and J. Kattge (2005), Inversion of terrestrial ecosystem model parameter values against eddy covariance measurements by Monte Carlo sampling, *Global Change Biol.*, *11*, 1333–1351.

Leonard, T., and J. S. J. Hsu (1999), *Bayesian Methods—An Analysis for Statistics and Interdisciplinary Researchers*, Cambridge Univ. Press, New York.

Luo, Y., and J. F. Reynolds (1999), Validity of extrapolating field CO_2 experiments to predict carbon sequestration in natural ecosystems, *Ecology*, *80*, 1568–1583.

Luo, Y., B. Medlyn, D. Hui, D. Ellsworth, J. F. Reynolds, and G. Katul (2001), Gross primary productivity in the Duke Forest: Modeling synthesis of the free-air CO_2 enrichment experiment and eddy-covariance measurements, *Ecol. Appl.*, *11*, 239–252.

Luo, Y., L. W. White, J. G. Canadell, E. H. DeLucia, D. S. Ellsworth, A. Finzi, J. Lichter, and W. H. Schlesinger (2003), Sustainability of terrestrial carbon sequestration: A case study in Duke Forest with inversion approach, *Global Biogeochem. Cycles*, *17*(1), 1021, doi:10.1029/2002GB001923.

Luo, Y. Q., et al. (2004), Progressive nitrogen limitation of ecosystem responses to rising atmospheric carbon dioxide, *BioScience*, *54*(8), 731–739.

Mastrandrea, M. D., and S. H. Schneider (2004), Probabilistic integrated assessment of "dangerous climate change", *Science*, *304*, 571–575.

McGuire, A. D., III, I. C. Prentice, N. Ramankutty, T. Reichenau, A. Schloss, H. Tian, L. J. Williams, and U. Wittenberg (2001), Carbon balance of the terrestrial biosphere in the twentieth century: Analyses of CO_2, climate and land use effects with four process-based ecosystem models, *Global Biogeochem. Cycles*, *15*, 183–206.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller (1953), Equation of state calculation by fast computer machines, *J. Chem. Phys.*, *21*(6), 1087–1092.

Murphy, J. M., D. M. H. Sexton, D. N. Barnett, G. S. Jones, M. J. Webb, M. Collins, and D. A. Stainforth (2004), Quantification of modeling uncertainties in a large ensemble of climate change simulations, *Nature*, *430*, 768–772.

Oh, S.-H., and B.-D. Kwon (2001), Geostatistical approach to Bayesian inversion of geophysical data: Markov chain Monte Carlo method, *Earth Planets Space*, *53*, 777–791.

Parton, W. J., D. S. Schimel, C. V. Cole, and D. S. Ojima (1987), Analysis of factors controlling soil organic matter levels in Great Plains grasslands, *Soil Sci. Soc. Am. J.*, *51*, 1173–1179.

Parton, W. J., W. B. Stewardt, and C. V. Cole (1988), Dynamics of C, N, P and S in grassland soils: A model, *Biogeochemistry*, *5*, 109–1131.

Raupach, M. R., P. J. Rayner, D. J. Barrett, R. S. Defries, M. Heimann, D. S. Ojima, S. Quegan, and C. C. Schmullius (2005), Model-data synthesis in terrestrial carbon observation: Methods, data requirements and data uncertainty specifications, *Global Change Biol.*, *11*, 378–397.

Rayner, P. J., M. Scholze, W. Knorr, T. Kaminski, R. Giering, and H. Widmann (2005), Two decades of terrestrial carbon fluxes from a carbon cycle data assimilation system (CCDAS), *Global Biogeochem. Cycles*, *19*, GB2026, doi:10.1029/2004GB002254.

Schlesinger, W. H., and J. Lichter (2001), Limited carbon storage in soil and litter of experimental forest plots under increased atmospheric CO_2, *Nature*, *411*, 466–469.

Spall, J. C. (2003), *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*, *Wiley-Intersci. Ser. Discrete Math. Optim.*, Wiley-Interscience, Hoboken, N. J.

Tarantola, A. (1987), *Inverse Problem Theory: Methods for Data Fitting and Model Parameter Estimation*, Elsevier, New York.

Von Mises, R. (1964), *Mathematical Theory of Probability and Statistics*, Elsevier, New York.

Wang, Y., R. Leuning, H. Cleugh, and P. Coppin (2001), Parameter estimation in surface exchange models using nonlinear inversion: How many parameters can we estimate and which measurements are most useful?, *Global Change Biol.*, *7*, 495–510.

D. Hui, Y. Luo, and T. Xu, Department of Botany and Microbiology, University of Oklahoma, 770 Van Vleet Oval, Norman, OK 73019-0245, USA. (yluo@ou.edu)

L. White, Department of Mathematics, University of Oklahoma, Norman, OK 73019, USA.