Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER Applied Mathematics and Computation 167 (2005) 419–434
APPLIED MATHEMATICS AND COMPUTATION
www.elsevier.com/locate/amc

Model-based data assessment for terrestrial carbon processes: implications for sampling strategy in FACE experiments

Luther W. White [a,*], Yiqi Luo [b]

[a] *Department of Mathematics, University of Oklahoma, Norman, Oklahoma 73019, United States*
[b] *Department of Botany and Microbiology, University of Oklahoma, Norman, Oklahoma 73019, United States*

Abstract

The value of different types of data in the estimation of different carbon transfer parameters is investigated. A carbon accounting model is used with different observation operators to generate data. The effectiveness of the inversion is assessed by observing relative errors of estimators and likelihood ratios. It is demonstrated that for an observation operator that relative errors vary widely with the sample test problems. An effective strategy to test types of data is to test the effectiveness of corresponding observation operators on an ensemble of sample problems for which parameters are selected from the space of admissible parameters. The selection is carried out under the assumption that the test parameters themselves are random variables uniformly distributed over the space of admissible parameters.

* Corresponding author.
E-mail address: lwhite@ou.edu (L.W. White).

0096-3003/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.amc.2004.07.016

1. Introduction

We consider an initial value problem modeling the sequestration and transfer of carbon (C) among various pools in a forest ecosystem. The model is derived from conservation considerations as C passes among 7 pools: nonwoody biomass, woody biomass, metabolic litter, structural litter, microbes, slow organic matter (SOM), and passive SOM. The solution of the model equations is a time-dependent vector with seven components expressing the quantity of C in a particular pool as a function of time. We refer to this vector as the state of the system. The vector of initial pool sizes, system C influx function, the functions describing of moisture and temperature effect, and C transfer coefficients are parameters needed to formulate the model. In a paper by Luo et al. [1] the development of this model is described and used with an output-least-squares estimation procedure to invert six data sets related to forest C processes in order to estimate transfer coefficients. In the present work we also consider only the transfer coefficients as parameters to be estimated with the other parameters being regarded as given and certain. The consideration of the consequences of their uncertainty is a topic of future studies. The inversion exercise then seeks to determine a set of transfer coefficients that may be considered as scaling multipliers within the model equation. In White and Luo [4] a similar 12 pool model is used to analyze data to estimate C transfer coefficients. In the present paper, however, we use the 7-pool model as in Luo et al. [1] since it captures the salient features for our studies. However, we still use the finite difference approximations developed in White and Luo [4] to determine numerical solutions of the initial value problems. Both of the above papers use available data and the underlying C sequestration model on which to base the estimation routines.

Instead of using an available data set and obtaining the estimators based on that data, we are interested in this work in studying information on the identity of transfer coefficients contained by various types of data. It is useful in this regard to take the probabilistic point of view of inversion of Tarantola [3]. From this view, a priori estimates on parameters belonging to an admissible set are expressed as a uniform distribution defined on the set of admissible parameters. The information contained in different data sets may be deduced by comparing the a posteriori distribution obtained by including the data with the a priori distribution. It is then possible to deduce information such as likelihood intervals and estimators of the parameters, for example, from the resulting probability density function (pdf). In this context the inversion procedure coincides with the Bayesian paradigm, Tarantola [3].

The consideration of the information contained in various data sets is motivated by sampling issues in C research. Quantification of C sequestration in terrestrial ecosystems involves uncertainties from several sources. For example, we are usually limited to measurement of a set or a subset of C processes. It is,

therefore, critical to know which set or subset of measurements would provide estimates of C sinks with the least uncertainty. Second, ecosystem C processes involve complex sets of changes in pools sizes and fluxes with heterogeneous turnover rates. Carbon turnover in the foliage pool, for example, is much faster than that in recalcitrant soil organic matter. It is essential to understand durations of experimental measurements that are required to constrain parameter estimation of C transfer coefficients from different pools. The third sampling issue is related to the discrepancy between desirable parameters for model prediction and measurable processes in experiments. Many of the parameters in C models are not necessarily measurable. How to constrain those desirable, but not measurable, parameters is another great challenge in C research. We use this modeling approach to address these issues.

From the perspective taken here, data is associated with certain observational mappings that take the system state to a data space where measurements are made. Our interest is to use the model to investigate the information associated with these mappings. Our approach is to use the model to generate data by specifying a vector of parameters consisting of C transfer coefficients. By solving the model equations we obtain the associated state. Data is then constructed by applying the observational operator to this state. Using data derived in this manner, we then generate a joint pdf on the set of admissible parameters. This pdf contains a priori information on the parameters as well as the information from the data. From this pdf we calculate marginal pdfs, likelihood intervals, and estimators. We may then compare estimators based on our procedure with the generating parameter.

The particular types of data that we consider result from the application of different observation operators. In this regard, various measurements are made on the forest ecosystem. These data observers are detailed in Luo et al. [1] and we include them in a discussion in the Section 3. We analyze the information contained from those observations based on an observation time interval of a fixed length.

While considering data from a single generating parameter vector of C transfer coefficients is useful to assess different data types, it is of interest to associate a measure with the observation operator itself not depending on a particular generating parameter. We show that results can vary greatly depending on the choice of generating parameter vector. In fact we demonstrate that, for a given observation operator, there are generating coefficients for which a particular transfer parameter may be recovered very well while there may be other generating coefficients that those parameters are not well recovered. In other words the ability of a given observation operator to recover the coefficients depends on the generating coefficient. It would be useful to compare observation operators while removing dependence on a particular generating parameter vector. Thus, in order to assess the effectiveness of a given observation operator, it is desirable to consider a property dependent on the entire

sample space and not on a particular generating coefficient. Hence, in order to focus on the information resulting from a particular observation operator, we introduce an ensemble of transfer coefficients with which to compare relative error of recovered parameters.

The organization of this paper is as follows. In Section 2 we describe the underlying initial value problem that serves as the C sequestration model as well as its approximation. We discuss in Section 3 the observational operators and the formulation of probability density functions. In Section 4 we consider an example by specifying initial conditions, C influx functions, the parameter vector $\mathbf{c}_0$ of transfer coefficients used to generate data, as well as the parameter bounds used in formulating the parameter admissible set. We then present results for various estimators for comparison with the generating parameter. These results are obtained using all available data. Different observational operators are then considered. Since we have constructed the example it is possible to calculate relative error between estimators and the generating coefficients. The selection of $\mathbf{c}_0$ in this example is random, and it should be noted that in fact the actual results are not particularly good. This motivates the question of whether such is the case in general or whether the selected $\mathbf{c}_0$ is simply a bad choice. In Section 4 we demonstrate that, for a given observation operator, it is possible to choose $\mathbf{c}_0$ so that relative error of the estimator for a particular component is small. This means that to determine a measure for the efficiency of a given operator to recover data it is necessary to consider an ensemble composed of a collection of $\mathbf{c}_0$ parameters. Our approach is to construct an ensemble viewing the generating $\mathbf{c}_0$ itself as a random variable that is uniformly distributed over the set of admissible transfer coefficient vectors. We then determine expected errors that are obtained by integrating over the test $\mathbf{c}_0$ coefficients. We obtain expected relative errors for mean estimators and likelihood intervals. Finally, in Section 5, we discuss conclusions based on our results.

2. Underlying system and approximation

In this section we present the underlying model and the finite difference approximations that are used for its numerical approximation. In subsequent applications the numerical model will be used to generate what we take to be the "true" state. The underlying model is a system of 7 differential equations with initial conditions given by

$$\mathrm{d}/\mathrm{d}t\ \mathbf{x}(t) = \xi(t)AC\mathbf{x}(t) + \mathbf{b}u(t), \quad \mathbf{x}(0) = \mathbf{x}_0. \tag{1}$$

The matrix C is a diagonal matrix whose diagonal entries consist of the components of the vector $\mathbf{c}$

$$C = \mathrm{diag}(\mathbf{c}).$$

The vector **c** constitutes the parameter vector and is a vector in $\mathbf{R}^7$ giving the transfer coefficients among carbon pools. Here $\mathbf{x}(t)$ is a column 7-vector giving carbon pool sizes as a function of time. The 7×7 matrix A gives interaction weights among pools. The scalar-valued function $\xi(t)$ takes into account moisture and temperature effects. The function $u(t)$ is a real-valued C influx function while **b** is a column 7-vector determining the specific pools directly effected by the input functions. The system Eq. (1) is derived based on the analysis of carbon transfer see Luo et al. [1].

To study the problem, a finite difference approximation of the initial value problem (1) is introduced in White and Luo [4]. Thus, we consider the following system of difference equations. Setting

$$\mathbf{x}_j = \mathbf{x}(\mathbf{c})(t_j), \xi_j = \xi(t_j) \quad \text{and} \quad u_j = u(t_j)$$

the difference approximation to Eq. (1) that we use is given by

$$(\mathbf{x}_{j+1} - \mathbf{x}_j)/\Delta t = AC(\xi_{j+1}\mathbf{x}_{j+1} + \xi_j\mathbf{x}_j)/2 + \mathbf{b}(u_{j+1} + u_j)/2.$$

Combining the terms we find that

$$[I - \Delta t\xi_{j+1}AC/2]\mathbf{x}_{j+1} = [I + \Delta t\xi_j AC/2]\mathbf{x}_j + \Delta t\mathbf{b}(u_{j+1} + u_j)/2,$$

for $j = 0, 1, \ldots, NT - 1$. Set

$$B_j = I - \Delta t\xi_j AC/2,$$

$$B'_j = I + \Delta t\xi_j AC/2$$

and

$$\mathbf{f}_j = \Delta t\mathbf{b}(u_{j+1} + u_j)/2.$$

It follows that the system of difference equations is

$$\begin{aligned} &B_1\mathbf{x}_1 = B'_0\mathbf{x}_0 + \mathbf{f}_0, \\ &B_2\mathbf{x}_2 = B'_1\mathbf{x}_1 + \mathbf{f}_1, \\ &\cdots \\ &B_{NT}\mathbf{x}_{NT} = B'_{NT-1}\mathbf{x}_{NT-1} + \mathbf{f}_{NT-1}. \end{aligned} \tag{2}$$

Define the column vectors of length 7(NT)

$$\mathbf{X} = \begin{pmatrix} \mathbf{x_1}. \\ \vdots \\ \mathbf{x}_{NT} \end{pmatrix},$$

$$\mathbf{F} = \begin{pmatrix} \mathbf{f}_0 \\ \vdots \\ \mathbf{f}_{NT-1} \end{pmatrix}$$

and

$$\mathbf{F}_0 = \begin{pmatrix} B_0\mathbf{x}_0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

and the $7(\mathrm{NT}) \times 7(\mathrm{NT})$ matrix

$$\underline{B}(\mathbf{c}) = \begin{pmatrix} B_1 & 0 & \dots & & & 0 \\ -B_1' & B_2 & 0 & \dots & \dots & 0 \\ 0 & \dots & \dots & \dots & \dots & 0 \\ 0 & & \dots & & -B_{NT-1}' & B_{NT} \end{pmatrix}.$$

The approximating system is then given by

$$\underline{B}(\mathbf{c})\mathbf{X}(\mathbf{c}) = \mathbf{F} + \mathbf{F}_0. \tag{3}$$

Thus, given a transfer coefficient $\mathbf{c}$, the matrix $\underline{B}(\mathbf{c})$ is defined. Assuming invertibility of $\underline{B}(\mathbf{c})$, we may solve the above equation for $\mathbf{X}(\mathbf{c})$. The mapping $\mathbf{c} \to \mathbf{X}(\mathbf{c})$ is defined from the prescribed set of admissible transfer coefficients Q_{ad} to the state vector $\mathbf{X}(\mathbf{c})$.

3. Data observation operators and probability density functions

In this section we describe the measurement models used to construct data. These operators map the model state obtained in the previous section to observable quantities corresponding to data. We assume that a vector of transfer coefficients $\mathbf{c}$ is specified and that the column m-vector $\mathbf{x}_j(\mathbf{c})$ gives the state associated with $\mathbf{c}$ at time t_j. Observation operators are generally of the form

$$\boldsymbol{\phi}(\mathbf{c}) = \Phi\mathbf{c} + \boldsymbol{\phi},$$

where Φ is an 7×7 matrix and $\boldsymbol{\phi}$ is a column 7-vector. A measurement at t_j takes the form

$$z_j(\mathbf{c}) = \boldsymbol{\varphi}(\mathbf{c})^{\mathrm{T}}\mathbf{x}_j(\mathbf{c}) + \zeta_j,$$

where ζ_j is a term independent of the pool-size vector. We consider the following data sets and their observation operators.

Soil respiration

$$\Phi = \mathrm{diag}([0, 0, 0.55, 0.45, 0.7, 0.55, 0.55])$$

and

$$\boldsymbol{\varphi} = 0,$$

with the observation given by

$$z_j(\mathbf{c}) = \mathbf{c}^{\mathrm{T}} \Phi \mathbf{x}_j(\mathbf{c}) + 0.25(1 - b_1 - b_2)u_j,$$

where b_1 and b_2 are the first two components of the vector $\mathbf{b}$ and $\zeta_j = 0.25(1 - b_1 - b_2)u_j$.

Woody biomass

$$\Phi = 0$$

and

$$\boldsymbol{\phi} = [0, 1, 0, 0, 0, 0, 0]^{\mathrm{T}},$$

with the observation given by

$$z_j(\mathbf{c}) = \boldsymbol{\phi}^{\mathrm{T}} \mathbf{x}_j(\mathbf{c}).$$

Litterfall

$$\Phi = \mathrm{diag}([1, 0, 0, 0, 0, 0, 0])$$

and

$$\boldsymbol{\phi} = 0,$$

with the observation given by

$$z_j(\mathbf{c}) = \mathbf{c}^{\mathrm{T}} \Phi \mathbf{x}_j(\mathbf{c}).$$

Foliage biomass

$$\Phi = 0$$

and

$$\boldsymbol{\phi} = [0.75, 0, 0, 0, 0, 0, 0]^{\mathrm{T}},$$

with the observation given by

$$z_j(\mathbf{c}) = \boldsymbol{\phi}^{\mathrm{T}} \mathbf{x}_j(\mathbf{c}).$$

To compare data we define a fit-to-data functional as follows. For the kth data set above, let

$$J_k(\mathbf{c}) = \varepsilon_k \sum_{i=1}^{NTk} \left[\boldsymbol{\phi}(\mathbf{c})^{\mathrm{T}} \mathbf{x}_j(\mathbf{c}) + \zeta_j z_i^k \right]^2,$$

where

$$\varepsilon_k = 1/\big((N_{Tk} - 1)\mathrm{var}(z^k)\big).$$

For each k the fit-to-data functional $J_k(\mathbf{c})$ gives an expression of the squared error between the data vector z^k and the corresponding output associated with a given transfer coefficient $\mathbf{c}$, $\{\Phi^k(\mathbf{c})x_i(\mathbf{c})\}_{i=1}^{NTk}$. For our purposes $NTk = NT$ corresponds to the number of time steps used in solving the initial value problem above. If observations occur at other times, then NTk represents the number of observations at those times. The fit-to-data functional is given by

$$J(\mathbf{c}) = 1/5\sum_{k=1}^{5} J_k(\mathbf{c}).$$

Instead of considering a minimization problem, as done in Luo et al. [1], to find that coefficient vector from among an admissible set minimizing the fit-to-data functional, we take a probabilistic view. Towards this end, we introduce the function

$$f_m(\mathbf{c}) = K\exp(-J(\mathbf{c})/2),$$

where the constant K is a normalization constant such that the integral of $f_m(\mathbf{c})$ over the sample space Q is one. The function $f_m(\mathbf{c})$ is a pdf that carries the information from the measurements and the model equation. A priori bounds on parameters are incorporated into the probabilistic formulation by assuming the components of the parameter vector are independent random variables. Thus, the information on bounds is specified and the a priori joint pdf on $Q_{\mathrm{ad}} \subset Q$ is obtained as the product of uniform distributions defined on the intervals constituting the bounds on each parameter. The set of admissible parameters Q_{ad} is defined as a Cartesian product of the constituent bounding intervals. We designate this pdf by $f_a(\mathbf{c})$. The information from the model/data and the prior bounds is combined as a conjunction of information as in Tarantola [3] to obtain a pdf $f(\mathbf{c})$ as

$$f(\mathbf{c}) = f_m(\mathbf{c})f_a(\mathbf{c}).$$

From this joint probability density function we obtain the following:

(1) Quantification of system information based on the model, prior information, and data.
(2) Marginal distributions for individual parameters,

$$f_i(\mathbf{c}_i) = \int_{Q'_i} f(\mathbf{c})\,\mathrm{d}\mathbf{c}'_i,$$

for the marginal pdf where integration is with respect to all parameters except $\mathbf{c}_i$ and $Q'i$ designates the parameter space excluding the $\mathbf{c}_i$ variable. The marginal cumulative distribution is given by

$$F_i(s) = \int_{P_i(s)} f(\mathbf{c})\,\mathrm{d}\mathbf{c},$$

where

$$P_i(s) = \{\mathbf{c} : c_i \leqslant s\}.$$

(3) Mean and maximum likelihood estimators are given by

$$c_{\text{imean}} = \int_Q c_i f(\mathbf{c})\,\mathrm{d}\mathbf{c}$$
$$c_{\text{imax}} = \max\{f_i(c_i) : c_i \in Q_i\}.$$

(4) Likelihood intervals and bounds are determined as follows. Let $\overline{c}_i$ and $\underline{c}_i$ be defined by

$$F_i(\overline{c}_i) = 0.95 \quad \text{and} \quad F_i(\underline{c}_i) = 0.05,$$

as left and right endpoints of a 90% likelihood interval $[\underline{c}_i, \overline{c}_i]$.

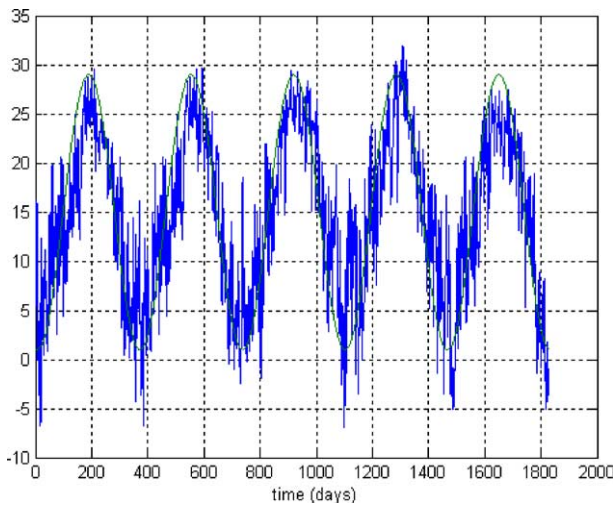

Fig. 1. Temperature time series and approximation.

4. Formulation of test problems

In this section we formulate a specific test problem in which time is measured in months. The coefficient function $\xi(t)$ is included to model seasonal environmental fluctuations in moisture and temperature. The function $\xi(t)$ is taken to be a product of the form

$$\xi(t) = \mathrm{Temp}(t)\mathrm{Mois}(t).$$

In Figs. 1 and 2 are presented the observed time series for Temp(t) and Mois(t) with trend functions $T_p(t)$ and $M_s(t)$ given by

$$T_p(t) = 15 + 14\sin(2\pi(30t + 266)/365)$$

and

$$M_s(t) = 0.27 + 0.14\sin(2\pi(30t + 46)/365).$$

For the model in this work we use the trend functions to express $\xi(t)$. Thus, we consider

$$\xi(t) = T_p(t)M_s(t)$$

and leave the inclusion of the uncertainty $\xi(t)$ due to the noise in the temperature and moisture measurements for a later analysis. We consider the C flux input as expressed by the trend function

$$u(t) = 6 + 5\sin(2\pi(30t + 269)/365).$$

The time series of influx measurements is given in Fig. 3.

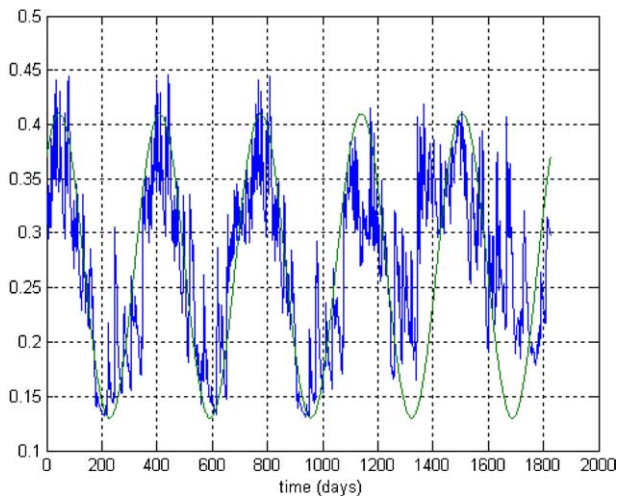

Fig. 2. Moisture time series and approximation.

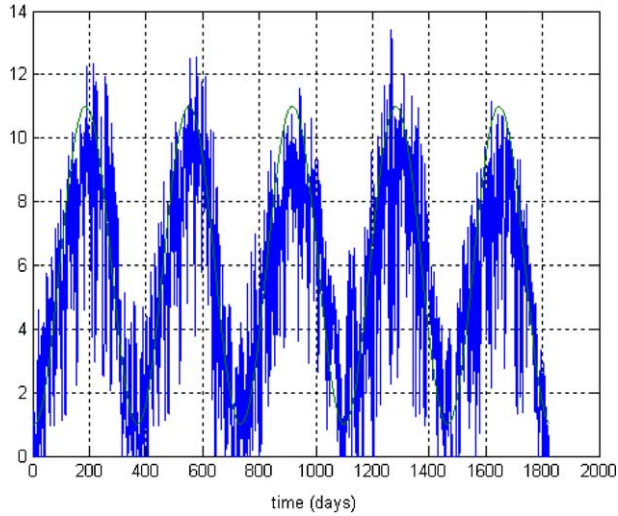

Fig. 3. CO_2 influx function.

To pose our sample problem, we define the following matrices and vectors

$$A = \begin{pmatrix} -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0.7123 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0.2877 & 1 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0.45 & 0.275 & -1 & 0.42 & 0.45 \\ 0 & 0 & 0 & 0.275 & 0.296 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0.004 & 0.03 & -1 \end{pmatrix},$$

$$\mathbf{b} = [0.25, 0.3, 0, 0, 0, 0, 0]^{\mathrm{T}}$$

and

$$\mathbf{x}_0 = [469, 4100, 64, 694, 123, 1385, 923]^{\mathrm{T}},$$

as the vector of initial pool sizes. To formulate the set of admissible parameters, we specify the following bounds on the components of the parameter vector $\mathbf{c}$.

$$\begin{aligned} 5.24 \times 10^{-3} &\leqslant c_1 \leqslant 5.91 \times 10^{-3}, \\ 1.63 \times 10^{-3} &\leqslant c_2 \leqslant 8.30 \times 10^{-3}, \\ 1.63 \times 10^{-1} &\leqslant c_3 \leqslant 8.28 \times 10^{-1}, \\ 1.63 \times 10^{-3} &\leqslant c_4 \leqslant 8.30 \times 10^{-2}, \\ 8.14 \times 10^{-2} &\leqslant c_5 \leqslant 2.59 \times 10^{-1}, \\ 1.63 \times 10^{-3} &\leqslant c_6 \leqslant 8.30 \times 10^{-3}, \\ 4.07 \times 10^{-5} &\leqslant c_7 \leqslant 2.77 \times 10^{-4}. \end{aligned}$$

The admissible set Q_{ad} is obtained as the Cartesian product of the intervals above. A vector of transfer coefficients is randomly chosen

$$\mathbf{c}_0 = [5.28 \times 10^{-2},\ 3.0 \times 10^{-3},\ 6.44 \times 10^{-1},\ 2.53 \times 10^{-2},\ 2.56 \times 10^{-1},\ 2.69 \times 10^{-3},\ 9.29 \times 10^{-5}]^{\mathrm{T}}.$$

The vector $\mathbf{c}_0$ is used to generate a system state. The system is solved numerically using the algorithm of the previous section with time step sizes of 1 month. We then use the observation operators specified in the previous section to operate on that state in order to generate data. These data are used to construct a joint probability density function as was outlined above.

The appropriate integrals are numerically calculated to obtain the marginal cumulative distribution functions and pdfs. A critical step in this program involves sampling the space of admissible parameters. Integration is approximated by using a formula such as

$$\int_{Q_s} f(\mathbf{c})\,\mathrm{d}\mathbf{c} = 1/N \sum_{i=1}^{N} \chi_{Qi}(i) f(\mathbf{c}^i)\Delta\mathbf{c},$$

see Niederreiter [2]. The number N corresponds to the number of samples $\mathbf{c}^i$ generated in the simulations from the parameter space. The quantity $\Delta\mathbf{c}$ represents the differences in the parameter bounds, and $\chi_{Qi}(i)$ is a characteristic function that is one if $\mathbf{c}^i \in Q_s$ and zero otherwise. The set Q_s designates the suitable parameter set for the particular computation.

To compare different observation operators, we next compare data in two ways for the case in which data is observed over 120 months. We present comparison using calculated relative error. Since we have constructed an example, we know the solution. Thus, we can calculate a relative error between the estimated coefficients and the model coefficients as a comparison of different data sets. For the ith coefficient, we calculate

$$\text{Relative error } (i) = |c_{\text{esti}} - c_{\text{oi}}|/c_{\text{oi}}.$$

The number of sample simulations N effects the accuracy of integrations, see Niederreiter [2]. We tried varying N from 50 to 2000 samples chosen as equidistributed sequences within Q_{ad}. We noted that the relative errors for maximum likelihood estimators are more sensitive to the number of samples while the relative errors for mean estimators are stable. The relative errors for mean estimators of $\mathbf{c}_0$ using 50 sample simulations are given by

$$0.35, 0.65, 0.15, 0.94, 0.34, 0.54 \text{ and } 0.77$$

and remain stable independent of the number of simulations. The relative errors of c_1, c_3, and c_5 are smaller than the other four coefficients, reflecting the fact that those C pools turn over much faster than the other pools.

Ecologically, c_1–c_7 represent C transfer coefficients from seven plant and soil pools. Specifically c_1 and c_2 describe litter fall from nonwoody and woody plant biomass, respectively. Coefficients c_3 and c_4 quantify decomposition of labile and structural component, respectively, of litter. Coefficient c_5 represents the turnover rate of microbial biomass while c_6 and c_7 are decomposition of soil organic matter in slow and passive pools, respectively. The seven transfer coefficients differ by a few degrees of magnitude Luo et al. [1]. The larger a coefficient is, the faster the pool that the coefficient represents turns over and reaches equilibrium. Since C transfers from woody plant biomass (c_2), structural litter pool (c_4), slow and passive soil organic matter pools (c_6 and c_7) is slow, the turnover rates of C in those pools ranges from tens to thousands of years. It usually requires long-term data sets to constrain the estimates of those coefficients.

In reality, plant C pools and fluxes are relatively easily measured in comparison to soil processes. Those data sets provide little constraint on below ground C transfer. Thus, the relative error of the coefficients c_3 is also high even though the turnover rate of C in that pool is fast.

We give estimated coefficients where data from all data sets are used over 120 months of data and 500 sample simulations are used to evaluate the integrals. In Fig. 4, the solid curve gives model coefficient parameters that are used to generate data. The dotted curve gives maximum likelihood estimators and the dashed curve is obtained as the mean estimator. Although the transfer coefficients themselves differ by a few magnitudes, the estimators reasonably match with the test values. In Table 1 is portrayed the relative errors for this example using individual data sets. It can be seen, while the relative errors are by no means small, that some data are better than others for different coefficients.

An issue of interest is how dependent are the results on the choice of the example. The results so far are obtained for a randomly chosen, but physically reasonable, $\mathbf{c}_0$. It is of interest to consider the effect of changing the value of the generating $\mathbf{c}_0$. In Table 2 we see that for any particular measurement operator there are values of $\mathbf{c}_0$ for which the relative error of the mean estimator is less than 13%. It follows that to determine an indicator of the effectiveness of a particular data type toward estimating a coefficient, we should consider an ensemble of $\mathbf{c}_0$ vectors. Towards this end, we generate an ensemble as an equi-distributed sequence from the admissible parameter set and view $\mathbf{c}_0$ as a random variable that is uniformly distributed over the admissible set of parameters Q_{ad}. The relative error is then viewed as a random variable defined on Q_{ad}. Our results indicate, for example, that data of woody biomass itself led to 40% of samples in the modeling study that have relative errors for c_2 less than 13%. Similarly, data of soil carbon help reduce relative errors for c_4 and c_6, see Table 2.

In Table 3 we present the mean relative error with respect to the choice of the example. That is, the relative error itself is considered a random variable

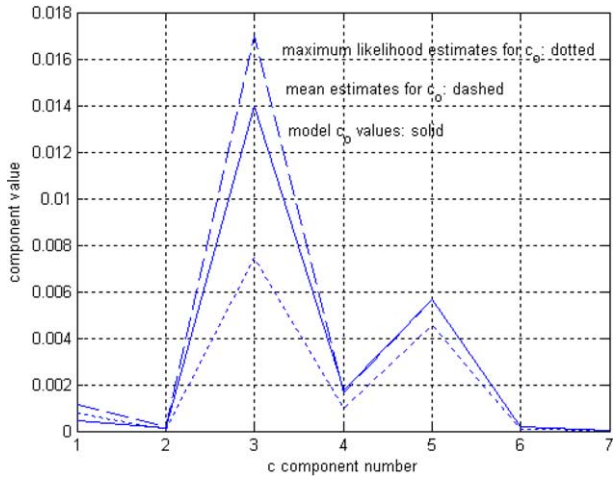

Fig. 4. Comparison of generating model coefficient and estimated coefficients.

Table 1
Relative error from mean estimator coefficients

	c_1	c_2	c_3	c_4	c_5	c_6	c_7
Soil respiration	0.36	0.65	0.20	1.07	0.30	0.88	0.79
Woody biomass	0.35	0.59	0.21	1.02	0.33	0.90	0.78
Foliage biomass	0.28	0.71	0.20	1.02	0.33	0.90	0.78
Litterfall	0.28	0.71	0.20	1.02	0.33	0.90	0.78
Soil carbon	0.36	0.65	0.19	0.91	0.30	0.32	0.74

Table 2
Fraction of samples with relative error less than 0.13

	c_1	c_2	c_3	c_4	c_5	c_6	c_7
Soil respiration	0.30	0.28	0.27	0.33	0.29	0.25	0.28
Woody biomass	0.32	0.40	0.24	0.25	0.24	0.25	0.25
Foliage biomass	0.27	0.22	0.23	0.23	0.25	0.23	0.26
Litterfall	0.27	0.23	0.24	0.24	0.26	0.25	0.26
Soil carbon	0.25	0.24	0.30	0.54	0.28	0.58	0.21

of the $\mathbf{c}_0$ and its mean is calculated. From the results in Table 3 we deduce that foliage biomass and litter fall data are effective for the estimation of c_1, woody biomass is effective data for c_2, and soil carbon is effective data for c_2, c_4, c_6. While the modeling results are consistent with our intuition, they also illustrate limitations of individual data sets in constraining parameter estimation for

Table 3
Mean relative error

	c_1	c_2	c_3	c_4	c_5	c_6	c_7
Soil respiration	0.6797	0.4462	0.4756	0.4564	0.3194	0.4846	0.5784
Woody biomass	0.7300	0.2616	0.4895	0.4932	0.3478	0.4894	0.5765
Foliage biomass	0.5077	0.4850	0.4900	0.4941	0.3475	0.4899	0.5779
Litterfall	0.5077	0.4850	0.4900	0.4941	0.3475	0.4899	0.5779
Soil carbon	0.7133	0.3442	0.4886	0.2379	0.3440	0.1514	0.5750

Table 4
90% likelihood ratio

	c_1	c_2	c_3	c_4	c_5	c_6	c_7
Soil respiration	0.9963	0.9925	0.9894	0.9686	0.9682	0.9970	0.9986
Woody biomass	0.9986	0.8840	1.000	0.9998	0.9996	0.9975	0.9984
Foliage biomass	0.9098	1.000	1.000	0.9987	0.9988	0.9995	1.000
Litterfall	0.9098	1.000	1.000	0.9987	0.9988	0.9995	1.000
Soil carbon	0.9962	0.9073	0.9947	0.7695	0.9894	0.5573	0.9942

coefficients that are not directly related to the measurements. All data seem to contain the same quantity of information for c_3, c_5, and c_7.

As a second indicator, we consider likelihood bounds for 90% likelihood intervals. Upper and lower bounds are calculated as $\mathbf{c}_0$ is varied. Means over the ensemble of these bounds are then calculated. As an indicator of the increase of information resulting from the introduction of data, we calculate the ratio of the length of the 90% likelihood interval determined in the presence of a given data type divided by the length of the 90% likelihood interval without the data. These results are presented in Table 4.

From this table we see that the likelihood is reduced for c_1 by foliage biomass and litter fall data. Woody biomass is useful in estimating c_2. Also, soil carbon data are effective in reducing the likelihood interval for c_2, c_4 and c_6. We note that c_3, c_5, and c_7 likelihood intervals are not reduced by any data sets. Correlations between the mean values of the relative error and the mean values of the 90% likelihood ratio are

$$c_1 = 0.99, \quad c_2 = 0.98, \quad c_3 = 0.91, \quad c_4 = 1.0, \quad c_5 = 0.98,$$
$$c_6 = 1.0, \quad c_7 = 1.0.$$

5. Conclusions

In this paper we have discussed the initial value problem and its approximation modeling the distribution of carbon among 7 pools. It is of interest to

estimate the coefficients describing the transfer among these pools from data obtained from measurements on various attributes of the system. In particular, 5 different types of observations are considered. A sample problem is considered in which data is generated from a randomly chosen $\mathbf{c}_0$. From that data a vector of transfer coefficients is estimated and a relative error is calculated. Results are given for mean estimators. It is shown that the relative error is dependent on the choice of $\mathbf{c}_0$. Thus, in order to determine an indicator of the information obtained from a given type of data, we view the generating values of $\mathbf{c}_0$ as uniformly distributed random variables as well as the relative error and likelihood intervals obtained from the data. We also use the ratio of the lengths of the 90% likelihood intervals for a posteriori distributions and a priori distributions as an indicator of the information obtained from the introduction of different data. We find that foliage biomass and litter fall are effective data for obtaining information for c_1, woody biomass is effective for estimation of c_2, while soil carbon is effective for obtaining information c_2, c_4 and c_6. The data presented does not seem to be effective in determining coefficients c_3, c_5, and c_7 due to the discrepancy between measurable processes in experiment and desirable parameter (e.g., c_3) in the model and a mismatch between short-term data availability and long-term processes (e.g., c_7). Results from this analysis call for more evaluation of effectiveness of experimental measurements on constraints of model parameter estimation.

Acknowledgments

This study was financially supported by the Office of Science (BER), US Department of Energy, Grant No. DE-FG03-99ER62800. This research is also part of the Forest-Atmosphere Carbon Transfer and Storage (FACTS-1) project at Duke Forest. The FACTS-1 project is supported by the US Department of Energy, Office of Biological and Environmental Research, under DOE contract DE-FG05-95ER62083 at Duke University and contract DE-AC02-98CH10886 at Brookhaven National Laboratory.

References

[1] Y.L. Luo, L. White, J. Canadell, E. DeLucia, E. Ellsworth, A. Finzi, J. Lichter, W. Schlesinger, Sustainability of terrestrial carbon sequestration: a cases study in Duke Forest with an inversion approach, Global Biogeochem. Cycles, vol. 17, no. 1, 1021 2003, pp. 21–34.

[2] H. Niederreiter, Random Number Generation and Quasi-Monte Carlo Methods, SIAM, Philadelphia, PA, 1992.

[3] A. Tarantola, Inverse Problem Theory, Elsevier, New York, 1987.

[4] L. White, Y. Luo, Estimation of carbon transfer coefficients using Duke Forest free-air CO_2 enrichment data, Appl. Math. Comput. 130 (2002) 101–120.