Soil Biology & Biochemistry 90 (2015) 197–203

Contents lists available at ScienceDirect

Soil Biology & Biochemistry

journal homepage: www.elsevier.com/locate/soilbio

Model structure and parameter identification of soil organic matter models

Carlos A. Sierra*, Saadatullah Malghani, Markus Müller

Max Planck Institute for Biogeochemistry, Hans-Knöll-Str. 10, 07745 Jena, Germany

ARTICLE INFO

Article history:
Received 26 April 2015
Received in revised form
24 July 2015
Accepted 5 August 2015
Available online 18 August 2015

Keywords:
Soil carbon
Parameter estimation
Identifiability analysis
Soil incubations
Radiocarbon

ABSTRACT

Soil organic matter models with complex ecological mechanisms usually include a large number of parameters than simpler models that omit detailed processes. Finding parameter values for these complex models is challenging given the poor availability of comprehensive datasets that describe different processes. Depending on the type of data available, the estimation of parameters in complex models may lead to identifiability problems, i.e. obtaining different combinations of parameters that give equally good predictions in comparison with the observed data. In this manuscript, we explore the problem of identifiability in soil organic matter models, pointing out combinations of empirical data and model structure that can minimize identifiability issues. We found that only sets of up to 3 or 4 parameters may be uniquely identifiable, depending on the number of data constrains used for parameter identification. When only using data on soil respiration fluxes from soil incubations or mass loss from litter decay studies, up to 2 parameters can be uniquely identifiable independently on the model structure. For nonlinear microbial models, all parameters cannot be identified simultaneously with mass loss or respiration data, combined with additional constraints from isotopes. Parameter identifiability possess series challenges for proposing complex model structures in global soil carbon models given the limitation of comprehensive datasets at a global scale.

1. Introduction

Soil organic matter cycling models are important to integrate with different sources of information and predict rates of element cycling in soils (Swift et al., 1979; Paustian et al., 1997; Manzoni and Porporato, 2009; Sierra et al., 2014). They can also be used for predicting effects of land-use changes on carbon stocks or consequences of other environmental changes on soil organic matter and respiration fluxes. These models play an important role in Earth system models because soils are one of the major global reservoirs of carbon, but predictions among models disagree considerably (Friedlingstein et al., 2013; Todd-Brown et al., 2013); therefore, there is interest in improving the structure and parameterization of the current generation of soil carbon models.

Obtaining parameter values for soil carbon models can be challenging, but a new set of data-assimilation techniques are being used now to integrate data from soil incubation experiments or litter decomposition studies as well as soil carbon stocks (e.g., Tuomi et al., 2009; Schädel et al., 2013; Ahrens et al., 2014; Hararuk et al., 2015; Schädel et al., 2014; Hararuk and Luo, 2014). These new techniques offer a tremendous opportunity to better describe complex mechanisms of soil organic matter cycling using relatively simple to collect data such as respiration fluxes from laboratory incubations or mass loss from litter decomposition experiments. Furthermore, these techniques can help to determine the kinetic heterogeneity of the soil organic matter by statistically separating different pools without resorting to laboratory fractionations (Schädel et al., 2013; Ahrens et al., 2014).

A major challenge however, is to obtain identifiable parameter sets from data-assimilation procedures given the data available. A model is said to be identifiable if different values of the parameters generate different probability distributions of the observable variables. If multiple parameter sets generate similar probability distributions for the observed variables the model is said to be non-identifiable with respect to observations (Brun et al., 2001; Omlin et al., 2001; Luo et al., 2009; Soetaert and Petzoldt, 2010). In many cases, the number of parameters to identify in the model exceeds the number of parameters that can be uniquely identified from the available data. In these cases it is common to say that the

* Corresponding author.
E-mail address: csierra@bgc-jena.mpg.de (C.A. Sierra).

http://dx.doi.org/10.1016/j.soilbio.2015.08.012
0038-0717/

model is overparameterized with respect to the observations (Beck, 1987; Brun et al., 2001). This issue leads to problems in data assimilation such as poor convergence properties in optimization routines (Brun et al., 2001), and strong correlations among obtained posterior values in Bayesian optimization methods. This is similar to collinearity in regression; if a model is non-identifiable with respect to observations the obtained parameter values may compensate each other and the model output is insensitive to these changes in parameters when compared with the available data.

The identifiability problem, also called equifinality in other contexts, has been addressed previously in hydrology under the framework of the generalized likelihood uncertainty estimation (GLUE, Beven and Binley, 1992; Beven and Freer, 2001; Beven, 2006); and has been applied to soil organic matter modeling by Wetterstedt and Ågren (2011) and Sierra et al. (2012b). However, an alternative and more intuitive framework has also been proposed by Brun et al. (2001) to establish the set of parameters that can be uniquely identified given an existing dataset; where uniquely is understood in the sense of unique probability distributions generated from the parameter sets. The method consists of calculating a collinearity index given certain model structure and available data. The index informs about the degree of near linear dependence by which changes in parameter values predict the same dataset equally well. When the collinearity index is high, a change in the value of one parameter is compensated by a change in other parameters generating almost identical probability distributions for the observed variables.

We use here this alternative identifiability framework to determine the type of model structures, and number of parameters for each model structure, that can be uniquely identified with easy-to-collect datasets in soil organic matter studies. First, we performed a theoretical analysis to determine combinations of parameters and available datasets that lead to identifiability problems in soil incubation experiments. Second, we used data from an incubation with data on soil CO_2 release rates to obtain parameter values for different models and determine the cases associated with identifiability issues. Our objective is to provide recommendations about the complexity of the models that can be identified with specific empirical data.

2. Methods

Two types of analyses are presented here; first, we show a theoretical analysis of the type of models and number of parameters that can be estimated with data from soil incubation experiments under the assumption of perfect data; i.e. artificially generated data from a known model. We consider two cases, linear compartment models and nonlinear microbial-explicit models. Second, we use observed data from an incubation experiment that contains measurement uncertainty and perform a parameter estimation procedure for different model structures. The first analysis relies on the calculation of a collinearity index while the second analysis uses inverse parameter estimation techniques. These are described in the following sections.

2.1. Collinearity index

The identifiability analysis used here and the calculation of the collinearity index was initially described in Brun et al. (2001), applied in Omlin et al. (2001) and Medlyn et al. (2005), and implemented in the UNCSIM computer program (Reichert, 2005) and in R package FME (Soetaert and Petzoldt, 2010). The procedure consist of two-steps, first a sensitivity matrix for all model parameters is calculated, and second the collinearity index is calculated from this sensitivity matrix for different combinations of parameter sets. This matrix, which its elements consists of dimensionless sensitivities of model output with respect to parameters, is calculated according to the expression

$$S_{i,j} = \frac{\partial r_i}{\partial \Theta_j} \cdot \frac{w_{\Theta_j}}{w_{r_i}}, \tag{1}$$

where $S_{i,j}$ represents each entry of the matrix, r_i are model residuals calculated from a cost (objective) function evaluated at the time point where observations are available, Θ_j is a model parameter, w_{r_i} is the scaling of r_i, and w_{Θ_j} is the scaling of parameter Θ_j (Soetaert and Petzoldt, 2010).

For different combinations of columns of the sensitivity matrix, which express parameters that will be identified with available data, the collinearity index γ is defined as

$$\gamma = \frac{1}{\sqrt{\min\left(\mathrm{EV}\left[\widehat{S}^{\top}\widehat{S}\right]\right)}} \tag{2}$$

with

$$\widehat{S}_{ij} = \frac{S_{ij}}{\sqrt{\sum_j S_{ij}^2}} \tag{3}$$

where $\widehat{S}$ contains the columns of the sensitivity matrix that correspond to the parameters included in the set, and EV represents the eigenvalues of the matrix $\widehat{S}^{\top}\widehat{S}$. If $\gamma = 1$, the columns are orthogonal and the parameter set is identifiable. If $\gamma \to \infty$, the parameters are linearly dependent (Brun et al., 2001; Soetaert and Petzoldt, 2010). The value of the collinearity index γ has an intuitive interpretation; a change in the residuals caused by a change in one of the parameters can be compensated by a proportional change $1/\gamma$ in another parameter. For example, a value of $\gamma = 20$ means that a change in residuals caused by a change in one parameter value can be compensated by changing another parameter value by 5%. As the value of γ increases, smaller changes in parameter values are required to compensate changes among each other (Brun et al., 2001).

2.2. Expected empirical data

For the identifiability analysis of soil incubation data, we expect that the soil carbon stock at the beginning of the experiment and the CO_2 evolution rates over time are always available. In addition, we assume that isotopic data may be available from the experiment either in the form of stable or radioactive isotopes. For simplicity, we assume that $\Delta^{14}C$ values of the respired CO_2 over time, and the $\Delta^{14}C$ value of the carbon stock may also be available.

Two distinct groups of models were considered, linear compartment models with first order rates, and a two-pool nonlinear model with an explicit microbial pool and Michaelis–Menten kinetics.

The set of linear models can be expressed in matrix form as (Sierra et al., 2012a)

$$\frac{d\boldsymbol{C}}{dt} = \boldsymbol{I} + \mathbf{A}\cdot\boldsymbol{C}, \tag{4}$$

where $\boldsymbol{C}$ is a vector of carbon pools, $\boldsymbol{I}$ a vector of carbon inputs to soil, and $\mathbf{A}$ a matrix of decomposition and transfer rates among compartments. This matrix can be split as $\mathbf{A} = \mathbf{T}\cdot\mathbf{K}$, i.e. a matrix of transfer coefficients among compartments and a matrix with decomposition rates k_i in the diagonal. The matrix of transfer coefficients contains -1 in the diagonal, and in the off-diagonal