Ecological Modelling 253 (2013) 97–106

Contents lists available at SciVerse ScienceDirect

Ecological Modelling

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

Evaluation of a soil greenhouse gas emission model based on Bayesian inference and MCMC: Model uncertainty

Gangsheng Wang[a,b,*], Shulin Chen[a,*]

[a] Department of Biological Systems Engineering, Washington State University, Pullman, WA 99164-6120, USA
[b] Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831-6301, USA

ARTICLE INFO

Article history:
Available online 6 October 2012

Key words:
Bayesian inference
Greenhouse gas (GHG)
Markov Chain Monte Carlo (MCMC)
Metropolis–Hastings algorithm
Model uncertainty

ABSTRACT

We combined the Bayesian inference and the Markov Chain Monte Carlo (MCMC) technique to quantify uncertainties in the process-based soil greenhouse gas (GHG) emission models. The Metropolis–Hastings sampling was examined by comparing four univariate proposal distributions (UPDs: symmetric/asymmetric uniform and symmetric/asymmetric normal) and one multinormal proposal distribution (MPD). Almost all the posterior parameter ranges from the MPD could be reduced to 1 order of magnitude. The simulation errors in CO_2 fluxes were much greater than those in N_2O fluxes, which resulted in a greater importance in model structure than in model parameters for CO_2 simulations. We suggested deriving the covariance matrix of parameters for MPD from the sampling results of a UPD; and generating a Markov chain by updating a single parameter rather than updating all parameters at each time. The method addressed in this paper can be used to evaluate uncertainties in other GHG emission models.

1. Introduction

Process-based mathematical models have been developed to simulate the greenhouse gas (GHG) emissions as an important part of the carbon–nitrogen dynamics in soils (Chen et al., 2008; Ma and Shaffer, 2001; Smith et al., 1997; Wu and McGechan, 1998). However, studies on the uncertainties in these models and model applications are limited (Wang and Chen, 2012). A subjective interpretation of uncertainty is "the degree of confidence that a decision maker has about possible outcomes and/or probabilities of these outcomes" (Refsgaard et al., 2007). Uncertainty assessment is therefore important when models are used for decision-making (Yohe and Oppenheimer, 2011). Uncertainty analysis not only gives the uncertainty from different sources (i.e., model parameters, model structure, model inputs and outputs), but also gives an evaluation of model performance and limitations.

Some studies on GHG models are concentrated on sensitivity analysis, which is one of the methods described in Refsgaard et al. (2007). The uncertainty of the PnET-N-DNDC model was evaluated by examining the sensitivity of the model outputs to such environmental factors as temperature, precipitation, photosynthetically active radiation (PAR) and model input variables, e.g., N-concentration in precipitation, litter mass, soil organic carbon (SOC), pH, and soil texture (Stange et al., 2000). The sensitivity analysis was conducted by changing one factor at a time while keeping all others constant.

Monte Carlo is the most widely used method in uncertainty anaylsis. Thorsen et al. (2001) used Monte Carlo to analyze the propagation of uncertainty from input data to model output, and found that the magnitude of uncertainty was closely associated with the investigated spatial scale, i.e., smaller output uncertainty on catchment scale than on grid level. The Monte Carlo technique was also applied to assess the uncertainty of DenNit model output with regard to parameterization (Reth et al., 2005). A Gaussian distribution within one standard error of mean was used for parameter sampling. In the comparison of carbon and nitrogen dynamics under conditions of conventional and diversified rotations, 64 parameter combinations were identified to test their impacts on model outcomes in Tonitto et al. (2007). An uncertainty analysis tool for the DNDC model allows for the selection of either the Monte Carlo or the Most Sensitive Factor (MSF) method (Li et al., 1992a). The MSF method involves running the model twice for each spatial unit (e.g., grid cell or polygon) with the maximum and minimum parameter values. These two runs generate two gas fluxes to form an interval, which is assumed to cover the real gas flux with a high probability. Using this uncertainty model, several highly sensitive factors influencing DNDC model were identified (Li et al., 1992a,b). The Monte Carlo analysis was also adapted to assess uncertainties in soil N_2O simulations from model input and structure of DAYCENT (Del Grosso et al., 2010). Probability distribution functions (PDFs) of major model inputs (weather, soil texture, and N applications)

* Corresponding authors at: Department of Biological Systems Engineering, Washington State University, Pullman, WA 99164-6120, USA.
E-mail addresses: wangg@ornl.gov (G. Wang), chens@wsu.edu (S. Chen).

0304-3800/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.ecolmodel.2012.09.010

were assigned to quantify the uncertainty due to model inputs. The structural uncertainty was a synthesis of parametric uncertainty and model residual errors derived by an empirically based linear mixed effect model (Ogle et al., 2007).

The aforementioned methods usually did not take into account the probability distribution of parameters. Even if a distribution was considered, e.g., a Gaussian distribution in Reth et al. (2005), it was a priori. The posterior distribution of a parameter is more informative than a priori for modelers to use and evaluate a data-driven model (Hartig et al., 2011). The Bayesian inference based on Markov Chain Monte Carlo (MCMC) has been used to calibrate parameters and quantify parametric uncertainty in the N_2O submodel of CERES-EGC (Lehuger et al., 2009), where the uniform probability distributions were assigned as a priori for 11 global parameters.

The model structure in Del Grosso et al. (2010) referred to both model parameters and structure (quantified by simulation errors); whereas model parameters and structure were usually regarded as two distinct sources for total uncertainty (Refsgaard, 1997; Wang et al., 2009). Lehuger et al. (2009) focused on parametric uncertainty and used the observation errors (standard deviations) of N_2O fluxes to represent the simulation errors in the likelihood estimator. However, the simulation errors may be treated as a latent variable and incorporated into the Bayesian framework with model parameters for quantifying the model structural uncertainty (Wang and Chen, 2012).

The objective of this paper was to evaluate model uncertainties due to model parameters and structure by coupling the Bayesian theory with MCMC method. Based on Bayes' theorem, the posterior distribution of both model parameters and model output variance can be derived from the prior distribution and observed outputs, and the 95% confidence intervals (CIs) of any output variables due to parameter uncertainty and model structure uncertainty can be estimated. The uncertainties from these two sources were also compared with that due to observation errors in the GHG fluxes. A soil GHG emission model (Appendix A) was used to test the proposed uncertainty analysis method.

2. Material and methods

2.1. Bayesian inference

According to Bayesian inference (Hartig et al., 2011), the posterior distribution $\pi(\Theta|y_t)$, i.e., the likelihood function $L(\Theta|y_t)$, of parameter set Θ can be generated from the prior distribution $f(\Theta)$ conditioned on observations y_t:

$$\pi\left(\Theta|y_t\right) = \frac{f\left(y_t|\Theta\right)\cdot f\left(\Theta\right)}{\int f\left(y_t|\Theta\right)\cdot f\left(\Theta\right)\cdot d\Theta} \propto f\left(y_t|\Theta\right)\cdot f\left(\Theta\right) \quad (1)$$

where $\Theta(\theta, \sigma_y)$ is a vector including the model parameter set (θ) and the standard deviation (σ_y) depicting simulation errors; $f(y_t|\Theta)$ is the distribution function of model output variable y_t conditioned on Θ; and t is a time index. Generally, y_t is a transformation of the model output Y_t to obtain a homoscedastic variance for the simulation errors (Kuehl, 1999). The square-root transformation (Engeland et al., 2005), i.e., $y_t = \sqrt{Y_t}$, is adopted in this study.

As for the prior distribution, a common approach is to assume uniform priors (Iskrev, 2007), which means $f(\Theta)$ is a constant. In addition, the model output (y_t) is assumed to follow a normal distribution (Congdon, 2001):

$$f\left(y_t|\Theta\right) = 1/\left(\sqrt{2\pi}\sigma_y\right)\cdot\exp\left[-\left(y_t - y_t(\theta)\right)^2/\left(2\sigma_y^2\right)\right] \quad (2)$$

Thus

$$\pi\left(\Theta|y_t\right) \propto f\left(y_t|\Theta\right) \propto 1/\sigma_y\cdot\exp\left[-\left(y_t - y_t(\theta)\right)^2/\left(2\sigma_y^2\right)\right] \quad (3)$$

$$\frac{\pi\left(\Theta^*|y_t\right)}{\pi\left(\Theta^k|y_t\right)} = \frac{\sigma_y^k}{\sigma_y^*}\exp\left\{-\left[\frac{\left(y_t - y_t(\theta^*)\right)^2}{2\left(\sigma_y^*\right)^2} - \frac{\left(y_t - y_t(\theta^k)\right)^2}{2\left(\sigma_y^k\right)^2}\right]\right\} \quad (4)$$

When all the simulation errors are assumed to be independent, the likelihood of the model outcome can be expressed as the product of the likelihood of each individual outcome at each time step

$$\pi\left(\Theta|y\right) = \prod_t \pi\left(\Theta|y_t\right) \quad (5)$$

2.2. Metropolis–Hastings algorithm for MCMC

The Metropolis–Hastings (MH) algorithm is a typical Markov Chain Monte Carlo (MCMC) sampling method to randomly sample from the posterior distribution described by Eq. (5). The procedure of MH can be found in many reports (Chumbley et al., 2007; Hastings, 1970; Link and Barker, 2008; Mathe and Novak, 2007; Tiana et al., 2007). An important criterion in MH is the acceptance probability:

$$a\left(x^*|x^k\right) = \min\left\{1, \frac{\pi\left(x^*\right)J\left(x^k|x^*\right)}{\pi\left(x^k\right)J\left(x^*|x^k\right)}\right\} \quad (6)$$

where $a(x^*|x^k)$ denotes the acceptance probability; x^k is the current state of the chain; x^* is the new state of the chain generated from x^k using a specified irreducible proposal distribution $J(x^*|x^k)$; and $\pi(\bullet)$ is the posterior distribution function defined by Eq. (5).

If we draw a random number (Z) from the uniform distribution $U(0,1)$, then a new state (i.e., $k+1$) of the chain can be determined by

$$x^{k+1} = \begin{cases} x^*, & \text{if } a\left(x^*|x^k\right) \geq Z \\ x^k, & \text{if } a\left(x^*|x^k\right) < Z \end{cases} \quad (7)$$

Four univariate (symmetric and asymmetric uniform, symmetric and asymmetric normal) and one multivariate (symmetric multinormal) proposal distributions were examined (see Appendix B). In the transition from the current state to a new state, three strategies may be used (Hastings, 1970). (i) All elements in the parameter vector are changed. In this case, x^* and x^k are no longer the one-dimensional parameters as in the previous four distributions, they are vectors containing d parameters. (ii) Only one of the elements is randomly selected and changed. (iii) Only one element is changed, and this element is selected in a fixed, rather than a random, sequence.

It is often inefficient to sample small values in MCMC if a parameter ranges several orders of magnitude. Thus we conducted MCMC pertaining to the logarithmic transformation of parameter values. Hereafter the log-transformed parameter space is called the logarithmic parameter space.

2.3. Generating random samples from multivariate normal distribution

It is not as simple to implement a multivariate normal distribution as a univariate distribution. It is required to know the covariance matrix in advance, and to randomly generate a vector, not a single parameter value, from the distribution. Although the true covariance matrix is unknown, it can be approximately estimated from the parameter samples generated by MCMC using any of the four univariate proposal distributions (UPDs). The algorithm described in Hernadvolgyi (1998) was followed to generate random vectors from a multinormal proposal distribution (MPD). In summary, the application of MPD in the MH algorithm may include six steps: