Ecological Informatics 6 (2011) 333–340

Contents lists available at SciVerse ScienceDirect

Ecological Informatics

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

Selection of likelihood parameters for complex models determines the effectiveness of Bayesian calibration

Karl-Heinz Rahn [b], Klaus Butterbach-Bahl [b], Christian Werner [a,b,*]

[a] LOEWE Biodiversity and Climate Research Centre (BIK-F), Senckenberganlage 25, 60325 Frankfurt am Main, Germany
[b] Karlsruhe Institute of Technology, Institute for Meteorology and Climate Research, Atmospheric Environmental Research, Kreuzeckbahnstr. 19, 82467 Garmisch-Partenkirchen, Germany

ARTICLE INFO

Article history:
Received 16 March 2011
Received in revised form 1 August 2011
Accepted 2 August 2011
Available online 10 August 2011

Keywords:
MCMC
Bayesian calibration
Objective function
Data likelihood
Hierarchical Bayes

ABSTRACT

Assessing the parameter uncertainty of complex ecosystem models is a key challenge for improving our understanding of real world abstractions, such as those for explaining carbon and nitrogen cycle at ecosystem scale and associated biosphere-atmosphere-hydrosphere exchange processes. The lack of data about the variance of measurements forces scientists to revisit assumptions used in estimating the parameter distribution of complex ecosystem models.

An increasingly used tool for assessing parameter uncertainty of complex ecosystem models is Bayesian calibration. In this paper, we generate two data sets which may represent a seasonal temperature curve or the seasonality of soil carbon dioxide flux and a single high peak put on a low background signal as is e.g. typical for soil nitrous oxide emission. Based on these examples we illustrate that commonly used assumptions for measurement uncertainty can lead to a sampling of wrong areas in the parameter space, incorrect parameter dependencies, and an underestimation of parameter uncertainties. This step needs particular attention by modelers as these issues lead to erroneous model simulations a) in present and future domains, b) misinterpretations of process feedback and functioning of the model, and c) to an underestimation of model uncertainty (e.g. for soil greenhouse gas fluxes). We also test the extension of the Bayesian framework with a model error term to compensate the effects caused by the false assumption of a perfect model and show that this approach can alleviate the observed problems in estimating the model parameter distribution.

© 2011 Elsevier B.V. All rights reserved.

1. Introduction

Computer based modeling is a key tool for integrating knowledge of different sources, e.g. field site measurements, theoretical assumptions, and laboratory investigations and for understanding the underlying processes (see Arhonditsis et al., 2008b). In order to gain inference from these models, real world applications have to prove the correctness and therefore the usability of these models. As ecological models always inherit structural errors or uncertainties of parameter values they can never map the real world perfectly. By using Bayesian analysis, we can quantify the uncertainty of model simulations, reduce the a priori uncertainty of model parameters, consider structural errors (see Arhonditsis et al., 2008b), update model parametrization after gaining additional knowledge (see Sivia, 2006) or gain arguments to rank and to choose between different models (see Gilks et al., 1996; Sivia, 2006; van Oijen et al., 2011). Therefore, Bayesian analysis is widely applicable throughout various scientific fields.

We work with biogeochemical models, which are increasingly used to simulate ecosystem carbon (C) and nitrogen (N) turnover processes in the plant-soil system as well as the associated exchange processes between the biosphere, atmosphere and hydrosphere (see Butterbach-Bahl et al., 2004; Li et al., 2001). As a Tier 3 methodology, defined by the International Panel on Climate Changes (IPCC) guidelines for national greenhouse gas (GHG) inventories, these biogeochemical models have been used for national inventories (see Del Grosso et al., 2006; Kesik et al., 2005) as they are capable to simulate soil greenhouse gas emissions based on the current understanding of the underlying biological and physico-chemical processes. When coupled to detailed spatial databases biogeochemical models can a) be used to model the pronounced spatial and temporal variability of GHG fluxes (Butterbach-Bahl et al., 2004; Werner et al., 2007), b) evaluate various management effects (e.g., fertilization or ploughing) on biosphere-atmosphere exchange processes (Butterbach-Bahl et al., 2004) or c) analyze effects, feedback and functioning of future climate conditions on the biogeochemical cycling of C and N (Kesik et al., 2006). Lokupitiya and Paustian (2006) state that such models provide more robust and accurate estimates of ecosystem GHG emissions and removals. However, these models require greater diligence in documentation, transparency, and uncertainty assessment to ensure comparability between countries.

* Corresponding author at: LOEWE Biodiversity and Climate Research Centre (BIK-F), Senckenberganlage 25, 60325 Frankfurt am Main, Germany. Tel.: +49 69 7542 1865; fax: +49 69 7542 1800.
E-mail address: christian.werner@senckenberg.de (C. Werner).

1574-9541/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecoinf.2011.08.002

A major challenge is that these models need to simulate all major processes involved in ecosystem C, N, and water cycling. Thus, they generally have a large number of model parameters. The MOBILE-DNDC model, a recently developed model framework incorporating parts of the biogeochemical DNDC and Forest-DNDC models (see Bruijn et al., 2009), has 147 model parameters for the submodule soil-chemistry alone. Some of these parameters are "lumped parameters", i.e. parameters describing complex biological or physico-chemical processes (e.g., describing N_2O formation by microbial nitrification in a given soil layer), which cannot directly be measured in the field due to significant spatial and temporal variability and short-comes in measurement methodologies.

Setting up a Bayesian calibration framework (see Gelman et al., 2003; Klemedtsson et al., 2008; Lehuger et al., 2009; Reinds et al., 2008) for MOBILE-DNDC, we were facing serious difficulties, which can already be observed when using Bayesian analysis for much simpler models. In this article, we therefore use two simple models as minimal examples, representing a common class of observations like a temperature curve or time-series of the seasonality of soil fluxes of carbon dioxide (CO_2) or nitrous oxide (N_2O) (see Wu et al., 2010).

Since the assumptions of the likelihood function are essential in this methodology, we investigate their influence on a Markov Chain in detail. Note that we focus on the class of normal distributions, as they are widely used in the community. A one-dimensional normal distribution is characterized by two parameters, an expectation and a variance. Both unknown parameters are estimated using measurements. Unfortunately, in ecological modeling, a reliable approximation of the variance of each measurement, due to micro-site variability or incomplete understanding of small-scale feedback mechanisms, is not always available, since e.g. the experimental quantification of a parameter describing temperature effects on gaseous denitrification products is hampered by the significant variability in denitrification activities in different soils and technical constrains with regard to the quantification of N_2 fluxes (Groffman et al., 2006). Moreover, consistency of measurement datasets is often affected by site access, experimental costs or erroneous individual measurements.

One approach to overcome the lack of information is to use a multiple of the measurement value as the standard deviation (e.g., a multiple of one or a half of the value). Introduced by the authors of van Oijen et al. (2005), a similar approach is used by numerous studies in the ecological community (see for example Arhonditsis et al., 2007, 2008a; Karlberg et al., 2006; Klemedtsson et al., 2008; Patenaude et al., 2008; Reinds et al., 2008; Svensson et al., 2008) as well as in other scientific communities such as hydrology (see Conrad and Fohrer, 2009). The lack of information about the variance of data for their investigated field site led the authors of van Oijen et al. (2005) to choose 30% of the mean value as standard deviation. Their assumption implies that a linear relationship between measurement values and their standard deviations exists. As these assumptions are widely used, it is important to understand the effect on the estimation of the posterior parameter distribution.

Therefore, we studied three different approaches for dealing with unknown standard deviations and contrast them with the true variability. We illustrate their influence on the posterior distribution and the efficiency of the Markov Chain. Additionally, we relax the assumption of a perfect model by introducing a model error term (cf. Arhonditsis et al., 2008b).

2. Notation

Using a deterministic model f depending on a parameter set described by a vector $\theta \in \mathbb{R}^n$, we derive for each input variable $x \in \mathbb{R}$ a model result $y \in \mathbb{R}$:

$$y := f(x;\theta), \tag{1}$$

which we compare to N data points (e.g. measurements) stored in a data vector D for N inputs x_i with $i \in 1,\ldots,N$. We assume that the difference between the data point of the ith location and the model output (given a parameter) y_i is normal distributed with expectation 0 and standard deviation $\sigma(i)$. Hence, the likelihood function of the difference is defined as (note that y_i depends on θ: $y_i = y_i(\theta)$):

$$l(\theta; D(i)) = \frac{1}{\sigma(i)\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{(D(i)-y_i)-0}{\sigma(i)}\right)^2}. \tag{2}$$

A common simplification when real variance information is unavailable is to set $\sigma(i)$ a) to a constant value for all measurements $D(i)$ or b) to a multiple of each measurement, e.g. $\sigma(i) = 1 \cdot D(i)$ (cf. Svensson et al., 2008; van Oijen et al., 2005, using a relative error of 30%). Since we consider the measurements to be independent, the likelihood function is simply the product of all individual likelihoods.

3. Simulations

With two simple models, we show the impact of different assumptions for an unknown σ on the posterior parameter distribution. We start with a simple sine curve (Model A) which could represent an annual temperature curve or the seasonality of soil GHG emissions. Subsequently a second model (Model B) is discussed, representing a single peak emission put on a low background as is e.g. typical for soil N_2O emission (see Wolf et al., 2010). We explore three different approaches for Model A, dealing with unknown σ and contrast the results with the true σ of our generated data. Note, that we could also use real data from ecological observations (e.g. soil emissions of N_2O or CO_2). But using a synthetic model has the advantage, that σ of each synthetic measurement (data point) is exactly known. Thus, we avoid introducing estimation errors of the measurements standard deviations.

In the first approach, we define σ as a multiple of the absolute data points (multiplicative sigma approach). In the second example, we additionally introduce a minimum value as a lower threshold for σ (capped sigma approach). This is motivated by the way Svensson et al. (2008) tried to minimize the impact of small data points on the total likelihood. A constant σ of 1.0 for all data points is used in the third example (constant sigma approach), whereas the last example is run with the exact σ (true sigma approach). The results of Model B were limited to the multiplicative sigma approach and the true sigma approach. We then relax the implicit assumption of a perfect model by introducing a model error term to Model A. Therefore we extend the Bayesian framework to a hierarchical Bayesian framework.

3.1. Model A

Consider a model with only two parameters a and b.

$$y := f(x;a,b) = a \cdot \sin(x) + b, \qquad x \in (0, 2\pi). \tag{3}$$

We generate a vector D with 100 synthetic data points for a given parameter set $\theta = (a,b) := (10,5)$ with standard normal distributed noise. To generate higher variances with higher absolute values we simply multiply the noise with the absolute value of sine + 1.0. Since the sine is bounded by -1.0 and 1.0, the absolute value plus 1.0 ranges between 1.0 and 2.0. Hence the standard normal noise is amplified by a factor between 1.0 (not amplified) and 2.0 (doubled).

$$D(i) = 10 \cdot \sin(x_i) + 5 + \epsilon_i \cdot (|\sin(x_i)| + 1) \tag{4}$$

$$x_i = \frac{i \cdot 2\pi}{100}, \qquad i = 1,\ldots,100 \tag{5}$$

$$\epsilon_i \sim N(0,1) \tag{6}$$

It directly follows that the standard deviation of data point $D(i)$ equals $|sin(x_i)| + 1$.[1] This synthetic dataset is generated once and used

[1] cf. Fisz (1989) $Var(aX+b) = a^2 Var(X)$.

Download English Version:

https://daneshyari.com/en/article/4375117

Download Persian Version:

https://daneshyari.com/article/4375117

Daneshyari.com