Ecological Modelling 301 (2015) 98–105

Contents lists available at ScienceDirect

Ecological Modelling

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

Fast calibration of a dynamic vegetation model with minimum observation data

Sebastian Lehmann [a,*], Andreas Huth [a,b]

[a] Helmholtz Centre for Environmental Research – UFZ, Department of Ecological Modelling, Permoserstr. 15, D-04318 Leipzig, Germany
[b] Institute for Environmental Systems Research (USF), University of Osnabrück, Barbarastr. 12, D-49069 Osnabrück, Germany

ARTICLE INFO

Article history:
Received 1 February 2014
Received in revised form 16 January 2015
Accepted 17 January 2015

Keywords:
Inverse modelling
Calibration
Dynamic vegetation model
Stochastic optimization

ABSTRACT

The estimation and uncertainty analysis of parameters for dynamic vegetation models is a complex process. If one is mainly interested in parameter estimation, this can be done with simple global stochastic search methods, while uncertainty analysis is carried out with traditional first-order analysis, which significantly reduces the number of needed model evaluations. Within a nonlinear regression framework, where the misfit between model and observations is expressed as a sum of weighted squares, we model the dynamics of tropical forest with a size-structured Sinko–Streifer model and demonstrate the general calibration procedure on a virtual data set. A second case study on real data for a single species shows that surprisingly total stem number, basal area and aboveground biomass are the minimum observations needed for successful calibration. A third case study on real data for a three species group shows the prediction of successional states while only using the former reduced set of observations for calibration. The methodology is well suited for time consuming models, where only limited amount of forest site observations are available.

1. Introduction

The calibration of dynamic vegetation models is a complex and time consuming process. Calibration refers to the procedure of adjusting the models parameters in such a way that the models response matches experimental observations. In the past this often has been done through manual calibration, making this a difficult task, especially if these models are complex and parameters may affect many processes at once. Recent research has applied Monte Carlo techniques under a Bayesian context to estimate parameter properties (Van Oijen et al., 2005, 2013; Hartig et al., 2013), but this needs a large amount of model evaluations to make reliable estimates of the underlying posterior distributions, because ecological models often have strong parameter correlations. If one is primary interested in parameter estimates and uncertainty assessment is less important, methods of global stochastic optimization in a frequentist context offer an alternative solution. The here presented methods are widely used in other research fields, particularly in the field of hydrological models (Duan et al., 1993; Tolson and Shoemaker, 2007; Gallagher and Doherty, 2007) although their application is new in the field of ecology.

In this study we give an overview over the steps needed for successful calibration of vegetation models and answer three specific questions. (1) Are methods of stochastic optimization able to correctly identify parameters values for a model of tropical rain forest. (2) What is the minimum amount of observation data needed to make reasonable estimates. (3) Are the parameter estimates made from observations at one point in time, able to make predictions about the temporal behaviour of the vegetation model in the past for a multi-species version of the model (succession).

2. Material and methods

In this section we give, among others, a more general introduction into methods, assumptions, possibilities and expected results for ecological model calibration within a statistical framework. This should help readers with a basic understanding to broaden their view on the underlying concepts.

2.1. Nonlinear regression model

We assume that the observed data vector y with dimension $|y| = n$ can be written in the following parametric form

$$y = m(\theta) + \epsilon \quad (1)$$

where $m(x)$ is the output of a (deterministic) model m with the to-be-estimated parameter vector $\theta \in \Theta$ with dimension

* Corresponding author. Tel.: +49 3412351064; fax: +49 341235451064.
E-mail addresses: sebastian.lehmann@ufz.de (S. Lehmann), andreas.huth@ufz.de (A. Huth).

http://dx.doi.org/10.1016/j.ecolmodel.2015.01.013
0304-3800/

$|\theta| = k$ and $\epsilon \sim N(0, S)$ is an additive multivariate normal distributed error term including systematic and measurement error. The distribution of y is the result of convoluting the error with the point mass $\delta_{m(\theta)}$, therefore, $y \sim N(m(\theta), S)$. We further independent observation, which means that $S = \sigma^2 S_W$, where $S_W = \mathrm{diag}(1/w_1, 1/w_2, \ldots, 1/w_n)$ is a diagonal matrix representing relative weights.

2.2. Parameter estimation

The unknown parameter vector θ gets estimated through maximum-likelihood estimation (MLE). The likelihood of y can be written as

$$L(\theta|y) = \frac{1}{\sqrt{(2\pi)^n |S|}} \exp\left(-\frac{1}{2} u^T S^{-1} u\right) \quad (2)$$

where $u = y - m(\theta)$. Taking the logarithm of $L(\theta|y)$ and leaving out constants that do not affect the location of the maximum results in the objective function

$$Q(\theta) = u^T W u \quad (3)$$

$$Q(\theta) = \sum_{i=1}^{n} w_i (y_i - m(x)_i)^2 \quad (4)$$

where $W = S_W^{-1} = \mathrm{diag}(w_1, w_2, \ldots, w_n)$ and $m(x)_i$ is the i-th component of the model output $m(x)$. Hence maximizing $L(\theta|y)$ is equivalent to minimizing $Q(\theta)$ and the parameter estimate $\hat{\theta}$ gets now determined through

$$\hat{\theta} = \underset{\theta \in \Theta}{\arg\min} Q(\theta) \quad (5)$$

The first question that arises is, how to assign values to the weights w_i, if the variances in the observation data are unknown beforehand? Generally little is known about the distribution of measuring error in ecological settings, albeit the use of the normal distribution is often justified from either the central limit theorem or as an approximative distribution. The situation gets more difficult, if one is trying to calibrate a model against different types of data, like stem-diameter-distribution, basal area or biomass. Also subjective requirements of the modeller may play an important role, like robustness (Huber, 2011) of the estimates or the minimization of a certain goodness of fit measure.

If nothing is known about the variances a method known as IRLS (iteratively recursive least squares) is applicable (Green, 1984). As an iterative procedure one starts with constant weights which get successively refined after each calibration step until the estimated parameters eventually converge. The refinement is based on the residuals where, in its simplest form, the squared residuals $u^T u$ of one estimation step serve as inverse weights for the next step. It can be shown that this results in the highest likelihood for the likelihood function (Eq. (2)).

Following such an approach multiple full calibration steps are needed, which are time consuming and impractical for the here considered case of complex models.

A simpler approach is to assume an existing functional relationship like $w_i = |y_i|^{-\xi}$ with $\xi \geq 0$ between the observations and the weights. Three values for ξ are frequently adopted. For $\xi = 0$ we have the case of constant variance, which is often assumed or acceptable to use after a suitable transformation (e.g. Box and Cox (1964)). For Poisson distributed field data with large rate parameter λ (>10), $\xi = 1$ serves as an good approximation. In Van Oijen et al. (2005, 2013) $\xi = 2$ is used, which implies that the standard deviation of the errors are proportional to the observations.

The second question is how to find the minimum of Q? Traditional nonlinear optimization algorithms like gradient descent or quasi Newton methods are only local search methods which additionally rely on the ability of computing first and/or second-order derivatives. The first property enforces one to make multiple optimization trials for multimodal problems and the second one is impractical due to the need for additional costly model evaluations. Hence we use a set of randomized search methods which do not exhibit the mentioned problems and do not take much effort to be implemented in an appropriate programming language.

- Adaptive Simulated Annealing (ASA): A parameter-free variant of the well known simulated annealing algorithm (Kirkpatrick, 1984). In contrast to the original annealing algorithm, a candidate solution gets sampled from a cauchy distribution and the cooling schedule is adapted to a fixed number of allowable function evaluations (Ingber, 1993, 1996).
- Dynamically dimensioned search (DDS): A simple algorithm originally developed for the calibration of watershed models, where a random candidate solution is drawn around the current best solution with the search radius r being the only free parameter. Through the successive reduction of the number of simultaneously perturbed parameters, DDS is also efficient for high dimensional problems (Tolson and Shoemaker, 2007).
- Adaptive Differential Evolution (JADE): An improved variant of the population based differential evolution algorithm (Storn and Price, 1997). Except the population size NP all other parameters are adjusted dynamically to the problem (Zhang and Sanderson, 2009).

2.3. General calibration procedure

Under the assumption of an existing simulation model the general calibration procedure is as follows:

First individual parameter ranges define the parameter space Θ, including expert knowledge regarding the ecological meaningfulness. An error model is specified according to available observation data or suitable assumptions about the error distribution. A random candidate solution $\theta \in \Theta$ is sampled from the parameter space, its objective function value $Q(\theta)$ is evaluated and successively improved through one of the stochastic search methods (Section 2.2) until a maximum number of simulations is reached.

For further analyses the determined maximum likelihood estimate $\hat{\theta}$ can be used to access uncertainty and identifiability properties. Either by estimating confidence intervals for $\hat{\theta}$ and the prediction $m(\hat{\theta})$, which requires a numerical approximation of the Jacobian at $\hat{\theta}$, or by estimating profile likelihoods using multiple calibration trials. See Appendix A for a detailed description and Appendix B for the proper handling of identifiability issues.

For selecting a model from a set of candidate models we use the Akaike information criterion (AIC) (Appendix C).

2.4. Modelling the dynamics of tropical forest

For modelling the dynamics of tropical forest we use a size-structured Sinko–Streifer model (Muller-Landau et al., 2006; Kohyama, 1991; Condit et al., 1998; Moorcroft et al., 2001). Let $n_i(t, x)$ ($\mathrm{cm}^{-1}\,\mathrm{m}^{-2}$) be the number of trees per area of species i at time t with a diameter at breast height (DBH) of x cm. The time evolution of species i is then described through the following partial differential equation (PDE):

$$\frac{\partial n_i(t,x)}{\partial t} = -\frac{\partial g_i(t,x) n_i(t,x)}{\partial x} - m_i(t,x) n_i(t,x) \quad (6)$$

With the boundary conditions $n_i(0, x) = 0\,\mathrm{cm}^{-1}\,\mathrm{m}^{-2}$ and $n_i(t, x_0) g_i(t, x_0) = r_i(t)$, which implies that we start with fallow land and assume a seedling input rate at the smallest stem diameter x_0 from outside. Where $g_i(t, x)$ ($\mathrm{cm\,a}^{-1}$) is the stem diameter grow function, $m_i(t, x)$