Environmental Modelling & Software 75 (2016) 273–316

Contents lists available at ScienceDirect

Environmental Modelling & Software

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

Markov chain Monte Carlo simulation using the DREAM software package: Theory, concepts, and MATLAB implementation

Jasper A. Vrugt [a, b, c, *]

[a] Department of Civil and Environmental Engineering, University of California Irvine, 4130 Engineering Gateway, Irvine, CA, 92697-2175, USA
[b] Department of Earth System Science, University of California Irvine, Irvine, CA, USA
[c] Institute for Biodiversity and Ecosystem Dynamics, University of Amsterdam, Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 17 February 2015
Received in revised form
13 August 2015
Accepted 13 August 2015
Available online xxx

Keywords:
Bayesian inference
Markov chain Monte Carlo (MCMC) simulation
Random walk metropolis (RWM)
Adaptive metropolis (AM)
Differential evolution Markov chain (DE-MC)
Prior distribution
Likelihood function
Posterior distribution
Approximate Bayesian computation (ABC)
Diagnostic model evaluation
Residual analysis
Environmental modeling
Bayesian model averaging (BMA)
Generalized likelihood uncertainty estimation (GLUE)
Multi-processor computing
Extended metropolis algorithm (EMA)

ABSTRACT

Bayesian inference has found widespread application and use in science and engineering to reconcile Earth system models with data, including prediction in space (interpolation), prediction in time (forecasting), assimilation of observations and deterministic/stochastic model output, and inference of the model parameters. Bayes theorem states that the posterior probability, $p(H|\tilde{\mathbf{Y}})$ of a hypothesis, H is proportional to the product of the prior probability, $p(H)$ of this hypothesis and the likelihood, $L(H|\tilde{\mathbf{Y}})$ of the same hypothesis given the new observations, $\tilde{\mathbf{Y}}$, or $p(H|\tilde{\mathbf{Y}}) \propto p(H)L(H|\tilde{\mathbf{Y}})$. In science and engineering, H often constitutes some numerical model, $\mathscr{F}(\mathrm{x})$ which summarizes, in algebraic and differential equations, state variables and fluxes, all knowledge of the system of interest, and the unknown parameter values, $\mathbf{x}$ are subject to inference using the data $\tilde{\mathbf{Y}}$. Unfortunately, for complex system models the posterior distribution is often high dimensional and analytically intractable, and sampling methods are required to approximate the target. In this paper I review the basic theory of Markov chain Monte Carlo (MCMC) simulation and introduce a MATLAB toolbox of the DiffeRential Evolution Adaptive Metropolis (DREAM) algorithm developed by Vrugt et al. (2008a, 2009a) and used for Bayesian inference in fields ranging from physics, chemistry and engineering, to ecology, hydrology, and geophysics. This MATLAB toolbox provides scientists and engineers with an arsenal of options and utilities to solve posterior sampling problems involving (among others) bimodality, high-dimensionality, summary statistics, bounded parameter spaces, dynamic simulation models, formal/informal likelihood functions (GLUE), diagnostic model evaluation, data assimilation, Bayesian model averaging, distributed computation, and informative/noninformative prior distributions. The DREAM toolbox supports parallel computing and includes tools for convergence analysis of the sampled chain trajectories and post-processing of the results. Seven different case studies illustrate the main capabilities and functionalities of the MATLAB toolbox.

1. Introduction and scope

Continued advances in direct and indirect (e.g. geophysical, pumping test, remote sensing) measurement technologies and improvements in computational technology and process knowledge have stimulated the development of increasingly complex environmental models that use algebraic and (stochastic) ordinary (partial) differential equations (PDEs) to simulate the behavior of a myriad of highly interrelated ecological, hydrological, and biogeochemical processes at different spatial and temporal scales. These water, energy, nutrient, and vegetation processes are often non-separable, non-stationary with very complicated and highly-nonlinear spatio-temporal interactions (Wikle and Hooten, 2010) which gives rise to complex system behavior. This complexity poses significant measurement and modeling challenges, in particular how to adequately characterize the spatio-temporal processes of the dynamic system of interest, in the presence of (often)

* Department of Civil and Environmental Engineering, University of California Irvine, 4130 Engineering Gateway, Irvine, CA, 92697-2175, USA.
E-mail address: jasper@uci.edu.
URL: http://faculty.sites.uci.edu/jasper

http://dx.doi.org/10.1016/j.envsoft.2015.08.013
1364-8152/

incomplete and insufficient observations, process knowledge and system characterization. This includes prediction in space (interpolation/extrapolation), prediction in time (forecasting), assimilation of observations and deterministic/stochastic model output, and inference of the model parameters.

The use of differential equations might be more appropriate than purely empirical relationships among variables, but does not guard against epistemic errors due to incomplete and/or inexact process knowledge. Fig. 1 provides a schematic overview of most important sources of uncertainty that affect our ability to describe as closely and consistently as possible the observed system behavior. These sources of uncertainty have been discussed extensively in the literature, and much work has focused on the characterization of parameter, model output and state variable uncertainty. Explicit knowledge of each individual error source would provide strategic guidance for investments in data collection and/or model improvement. For instance, if input (forcing/boundary condition) data uncertainty dominates total simulation uncertainty, then it would not be productive to increase model complexity, but rather to prioritize data collection instead. On the contrary, it would be naive to spend a large portion of the available monetary budget on system characterization if this constitutes only a minor portion of total prediction uncertainty.

Note that model structural error (label 4) (also called epistemic error) has received relatively little attention, but is key to learning and scientific discovery (Vrugt et al., 2005; Vrugt and Sadegh, 2013).

The focus of this paper is on spatio-temporal models that may be discrete in time and/or space, but with processes that are continuous in both. A MATLAB toolbox is described which can be used to derive the posterior parameter (and state) distribution, conditioned on measurements of observed system behavior. At least some level of calibration of these models is required to make sure that the simulated state variables, internal fluxes, and output variables match the observed system behavior as closely and consistently as possible. Bayesian methods have found widespread application and use to do so, in particular because of their innate ability to handle, in a consistent and coherent manner parameter, state variable, and model output (simulation) uncertainty.

If $\tilde{\mathbf{Y}} = \{\tilde{y}_1, \ldots, \tilde{y}_n\}$ signifies a discrete vector of measurements at times $t = \{1,\ldots,n\}$ which summarizes the response of some environmental system $\Im$ to forcing variables $\mathbf{U} = \{\mathbf{u}_1,\ldots,\mathbf{u}_n\}$. The observations or data are linked to the physical system.

$$\tilde{\mathbf{Y}} \leftarrow \Im(\mathbf{x}^*) + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{x}^* = \{x_1^*, \ldots, x_d^*\}$ are the unknown parameters, and $\boldsymbol{\varepsilon} = \{\varepsilon_1,\ldots,\varepsilon_n\}$ is a n-vector of measurement errors. When a hypothesis, or simulator, $\mathbf{Y} \leftarrow \mathscr{F}(\mathbf{x}^*, \tilde{\mathbf{u}}, \tilde{\psi}_0)$ of the physical process is available, then the data can be modeled using

$$\tilde{\mathbf{Y}} \leftarrow \mathscr{F}\left(\mathbf{x}^*, \tilde{\mathbf{U}}, \tilde{\psi}_0\right) + \mathbf{E}, \tag{2}$$

where $\tilde{\psi}_0 \in \boldsymbol{\Psi} \in \mathbb{R}^\tau$ signify the τ initial states, and $\mathbf{E} = \{e_1,\ldots,e_n\}$ includes observation error (forcing and output data) as well as error due to the fact that the simulator, $\mathscr{F}(\cdot)$ may be systematically different from reality, $\Im(\mathbf{x}^*)$ for the parameters $\mathbf{x}^*$. The latter may arise from numerical errors (inadequate solver and discretization), and improper model formulation and/or parameterization.

By adopting a Bayesian formalism the posterior distribution of the parameters of the model can be derived by conditioning the spatio-temporal behavior of the model on measurements of the observed system response

$$p\left(\mathbf{x}\middle|\tilde{\mathbf{Y}}\right) = \frac{p(\mathbf{x})p\left(\tilde{\mathbf{Y}}\middle|\mathbf{x}\right)}{p\left(\tilde{\mathbf{Y}}\right)}, \tag{3}$$

where $p(\mathbf{x})$ and $p\left(\mathbf{x}\middle|\tilde{\mathbf{Y}}\right)$ signify the prior and posterior parameter distribution, respectively, and $L\left(\mathbf{x}\middle|\tilde{\mathbf{Y}}\right) \equiv p\left(\tilde{\mathbf{Y}}\middle|\mathbf{x}\right)$ denotes the likelihood function. The evidence, $p(\tilde{\mathbf{Y}})$ acts as a normalization constant (scalar) so that the posterior distribution integrates to unity

$$p\left(\tilde{\mathbf{Y}}\right) = \int_{\chi} p(\mathbf{x})p\left(\tilde{\mathbf{Y}}\middle|\mathbf{x}\right)\mathrm{d}\mathbf{x} = \int_{\chi} p\left(\mathbf{x},\tilde{\mathbf{Y}}\right)\mathrm{d}\mathbf{x}, \tag{4}$$

over the parameter space, $\mathbf{x} \in \chi \in \mathbb{R}^d$. In practice, $p(\tilde{\mathbf{Y}})$ is not required for posterior estimation as all statistical inferences about $p\left(\mathbf{x}\middle|\tilde{\mathbf{Y}}\right)$ can be made from the unnormalized density

$$p\left(\mathbf{x}\middle|\tilde{\mathbf{Y}}\right) \propto p(\mathbf{x})L\left(\mathbf{x}\middle|\tilde{\mathbf{Y}}\right) \tag{5}$$

If we assume, for the time being, that the prior distribution, $p(\mathbf{x})$

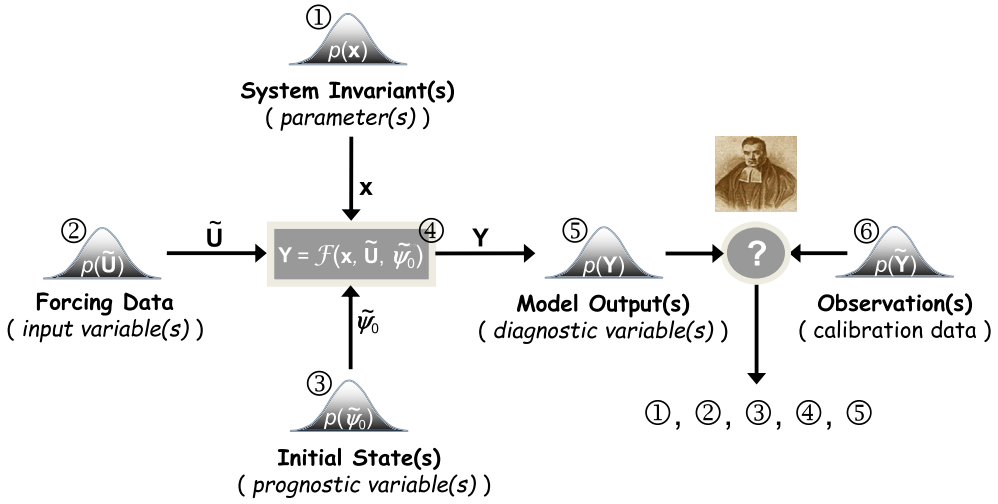

Fig. 1. Schematic illustration of the most important sources of uncertainty in environmental systems modeling, including (1) parameter, (2) input data (also called forcing or boundary conditions), (3), initial state, (4) model structural, (5) output, and (6) calibration data uncertainty. The measurement data error is often conveniently assumed to be known, a rather optimistic approach in most practical situations. Question remains how to describe/infer properly all sources of uncertainty in a coherent and statistically adequate manner.