Environmental Modelling & Software 60 (2014) 45–56

Contents lists available at ScienceDirect

Environmental Modelling & Software

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

A computer program for uncertainty analysis integrating regression and Bayesian methods

Dan Lu [a,*], Ming Ye [b], Mary C. Hill [c], Eileen P. Poeter [d], Gary P. Curtis [e]

[a] Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA
[b] Department of Scientific Computing, Florida State University, Tallahassee, FL 32306, USA
[c] U.S. Geological Survey, Boulder, CO 80303, USA
[d] Integrated Ground Water Modeling Center, Department of Geology and Geological Engineering, Colorado School of Mines, Golden, CO 80401, USA
[e] U.S. Geological Survey, Menlo Park, CA 94025, USA

ARTICLE INFO

Article history:
Received 18 November 2013
Received in revised form
1 June 2014
Accepted 2 June 2014
Available online

Keywords:
Markov Chain Monte Carlo
UCODE_2014
Bayesian uncertainty analysis

ABSTRACT

This work develops a new functionality in UCODE_2014 to evaluate Bayesian credible intervals using the Markov Chain Monte Carlo (MCMC) method. The MCMC capability in UCODE_2014 is based on the FORTRAN version of the differential evolution adaptive Metropolis (DREAM) algorithm of Vrugt et al. (2009), which estimates the posterior probability density function of model parameters in high-dimensional and multimodal sampling problems. The UCODE MCMC capability provides eleven prior probability distributions and three ways to initialize the sampling process. It evaluates parametric and predictive uncertainties and it has parallel computing capability based on multiple chains to accelerate the sampling process. This paper tests and demonstrates the MCMC capability using a 10-dimensional multimodal mathematical function, a 100-dimensional Gaussian function, and a groundwater reactive transport model. The use of the MCMC capability is made straightforward and flexible by adopting the JUPITER API protocol. With the new MCMC capability, UCODE_2014 can be used to calculate three types of uncertainty intervals, which all can account for prior information: (1) linear confidence intervals which require linearity and Gaussian error assumptions and typically 10s–100s of highly parallelizable model runs after optimization, (2) nonlinear confidence intervals which require a smooth objective function surface and Gaussian observation error assumptions and typically 100s–1,000s of partially parallelizable model runs after optimization, and (3) MCMC Bayesian credible intervals which require few assumptions and commonly 10,000s–100,000s or more partially parallelizable model runs. Ready access allows users to select methods best suited to their work, and to compare methods in many circumstances.

Published by Elsevier Ltd.

Software availability

Name of software: The Markov Chain Monte Carlo capability in UCODE_2014
Description: The Markov Chain Monte Carlo capability developed in UCODE_2014 to generate parameter samples and evaluate parametric and predictive Bayesian uncertainties
Developer: Dan Lu (lud1@ornl.gov), Mary Hill (mchill@usgs.gov), and Eileen Poeter (epoeter@mines.edu)
Programming language: Fortran
Availability: Download from website http://igwmc.mines.edu/freeware/ucode/

1. Introduction

Quantifying uncertainty in evaluations and predictions of how anthropogenic and/or natural events affect the environment is an important step of any mathematically based modeling effort. The new version of UCODE, UCODE_2014, provides a set of uncertainty quantification methods that range from computationally frugal regression methods (as few as 10s–100s of model runs after optimization) with often significant restrictive assumptions, to computationally demanding Bayesian methods (commonly 10,000s–100,000s of model runs) with few restrictive assumptions. All methods are able to account for prior information. Having

* Corresponding author.
E-mail address: lud1@ornl.gov (D. Lu).

http://dx.doi.org/10.1016/j.envsoft.2014.06.002
1364-8152/Published by Elsevier Ltd.

this range of methods readily accessible to users as provided by UCODE_2014 is important to the following goals:

(1) investigative studies in which the different uncertainty intervals types are compared and guidance is provided about circumstances for which the more computationally expensive Bayesian credible intervals are likely to be important and when the computationally cheaper regression confidence intervals are potentially useful (for example, see Lu et al. (2012)),
(2) progressive calculation of intervals so that computationally frugal regression confidence intervals can be used routinely earlier in a study while the more expensive Bayesian credible intervals can be calculated occasionally and often later in the study,
(3) calculation of only Bayesian credible intervals (as needed for models with very irregular objective function surfaces and often with multiple local minima) and,
(4) calculation of only computationally frugal regression confidence intervals (as needed to enable use of computationally demanding models and evaluation using multiple alternative models, and valid if linearity or smoothness, and Gaussian assumptions are not violated too much).

A program supporting such flexible strategies is needed because of limitations in the existing programs developed for uncertainty analysis in the environmental community. For example, the DAKOTA optimization and uncertainty software (Adams et al., 2013) previously evaluated the Bayesian credible intervals using the DRAM algorithm (Haario et al., 2006) which can be less efficient and unreliable for complex and multimodal problems than the DREAM algorithm used in UCODE_2014 (Vrugt et al., 2009). DREAM is now in the process of being implemented in DAKOTA. UCODE_2005 (Poeter et al., 2005) and PEST (Doherty, 2005) (which are both inverse modeling codes that can be used with any process models with ASCII-based inputs and outputs) provide uncertainty analysis with linear and nonlinear regression confidence intervals. Null space Monte Carlo (NSMC), another uncertainty analysis method encapsulated in PEST, provides predictive probability distributions in a computationally efficient way (Keating et al., 2010), but can display erratic performance (Laloy and Vrugt, 2012). iTOUGH2 (Finsterle and Zhang, 2011a,b) evaluates predictive uncertainty using linear uncertainty propagation and simple Monte Carlo analysis based on the distributions of uncertain parameters. Although they both use a Monte Carlo method, neither NSMC nor iTOUGH2 analyzes the predictive uncertainty in a rigorous Bayesian way by evaluating the posterior distributions. MICA (Doherty, 2003) and DREAM (Vrugt et al., 2008, 2009) (which are both MCMC codes that can be used to generate parameter samples from their posterior probability distribution) calculate only Bayesian credible intervals. None of the listed programs calculate both regression confidence intervals and Bayesian credible intervals efficiently.

MICA and DREAM are the two most widely used programs in the environmental community for Bayesian uncertainty analysis, and both codes are available at no charge from the developers. MICA is developed based on the Metropolis–Hastings algorithm. It is easy and straightforward to use with any process model that uses ASCII-based inputs and outputs. MICA input file and template and instruction files are similar or equivalent to those of PEST; the template and instruction files can also be used with UCODE_2014. MICA provides a wide range of probability density functions for the MCMC parameter prior distribution, and can evaluate parametric uncertainties for any or all model parameters and derived parameters. MICA cannot perform parallel MCMC computations. MICA works well for estimating unimodal posterior distributions, but for multimodal problems it cannot sample the target posterior distribution efficiently with a single proposal distribution (Gallagher and Doherty, 2007; Lu et al., 2012). This problem is resolved by DREAM (Vrugt et al., 2008, 2009) which is described in Section 2 of this paper.

This work integrates the DREAM algorithm into UCODE_2014, which is documented by Poeter et al. (2005, 2014). Inspired by the structure of MICA, the UCODE_2014 MCMC capability is user-friendly and can be easily used without in-depth knowledge of MCMC. The MCMC capability generates parameter samples and produces Bayesian predictive uncertainty by calculating model predictions from the generated parameter samples after a burn-in period (i.e., the parameter samples after chain convergence). The MCMC simulation in UCODE_2014 has parallel computing capability where the process model runs for different chains are accomplished on different processors for simultaneous execution. This greatly accelerates the MCMC sampling process.

With the new MCMC capability, UCODE_2014 can be used to calculate three types of uncertainty intervals: linear and nonlinear confidence intervals and Bayesian credible intervals. Confidence intervals are based on regression theories and credible intervals are based on Bayesian theories. While both can include the effect of prior information, confidence and credible intervals are conceptually different, and their differences and similarities are discussed in statistical literature including Jaynes (1976), Bates and Watts (1988), and Box and Tiao (1992). A recent discussion and literature review in the context of environmental modeling is presented by Lu et al. (2012). Given a nonlinear model and multi-Gaussian distributed observation errors, theory suggests that nonlinear confidence and credible intervals can be numerically identical if model nonlinearity is “small enough” and there are no local minima. They present a groundwater flow problem which indicates that even linear intervals can provide useful evaluations of uncertainty given common levels of nonlinearity. However, many environmental problems are so nonlinear that Bayesian methods with less restrictive assumptions are needed, and the ability to calculate both regression confidence intervals and Bayesian credible intervals is important (Vrugt and Bouten, 2002; Gallagher and Doherty, 2007; Liu et al., 2010; Shi et al., 2012, 2014).

The computational cost of calculating the confidence and credible intervals can be considerably different. Calculating the linear and nonlinear confidence intervals typically requires 10s–1,000s of model runs after a calibrated model is achieved. Model calibration includes identifying both the best fit parameter values and other aspects of model development. For a given model, one MCMC simulation can determine both the best fit parameter values and credible intervals that require neither smoothness nor Gaussian error assumptions. MCMC credible intervals can sometimes be obtained using 1,000s of model runs, but commonly require 10,000s, and even millions of model runs. For all methods, the number of model runs required tends to increase with problem dimensionality, though with linear confidence intervals the rate of increase is plus two runs for each additional parameter, more for nonlinear confidence intervals, and much more for MCMC credible intervals. Increasing nonlinearity leads to more model runs for nonlinear confidence intervals and MCMC credible intervals.

This paper introduces the MCMC capability implemented in UCODE_2014 and presents extensive tests. First, the MCMC method is briefly described in Section 2 with emphasis on the UCODE_2014 implementation. In Section 3, the features of the capability are discussed in detail. In Section 4, a 10-dimensional multimodal mathematical function and a 100-dimensional Gaussian function are used to test the MCMC capability in complex sampling problems, and a groundwater reactive transport model is presented to