Environmental Modelling & Software 66 (2015) 98–109

Contents lists available at ScienceDirect

Environmental Modelling & Software

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

Software framework for inverse modeling and uncertainty characterization

Carlos A. Osorio-Murillo [a,*], Matthew W. Over [b], Heather Savoy [b], Daniel P. Ames [a], Yoram Rubin [b]

[a] Brigham Young University, Department of Civil and Environmental Engineering, Provo, UT, USA
[b] University of California, Berkeley, Department of Civil and Environmental Engineering, Berkeley, CA, USA

ARTICLE INFO

Article history:
Received 5 September 2014
Received in revised form
16 December 2014
Accepted 3 January 2015
Available online 21 January 2015

Keywords:
Modeling frameworks
Method of Anchored Distributions
Inverse modeling
Spatial random fields
Model integration

ABSTRACT

Estimation of spatial random fields (SRFs) is required for predicting groundwater flow, subsurface contaminant movement, and other areas of environmental and earth sciences modeling. This paper presents an inverse modeling framework called MAD# for characterizing SRFs, which is an implementation of the Bayesian inverse modeling technique Method of Anchored Distributions (MAD). MAD# allows modelers to "wrap" simulation models using an extensible driver architecture that exposes model parameters to the inversion engine. MAD# is implemented in an open source software package with the goal of lowering the barrier to using inverse modeling in education, research, and resource management. MAD# includes an intentionally simple user interface for simulation configuration, external software integration, spatial domain and model output visualization, and evaluation of model convergence. Four test cases are presented demonstrating the novel functionality of this framework to apply inversion in order to calibrate the model parameters characterizing a groundwater aquifer.

Software availability

MAD# is made available through collaboration with the Consortium of Universities for the Advancement of Hydrologic Science (CUAHSI) Hydrologic Data Center. MAD# source code and documentation can be accessed at the MAD code repository website http://mad.codeplex.com. MAD# and its source code are released under the New Berkeley Software Distribution (BSD) License which allows for liberal reuse of the software and code.

1. Introduction

1.1. Overview

Spatial phenomena variability is typically evaluated through analytical and numerical models that describe the general properties of spatial random fields (SRFs). These models employ parameters and observations to define spatial variability. The characteristics – and hence variability – of an SRF can be discerned by the relationship between model parameters, direct, and indirect information. A number of hydrogeological studies have been conducted using SRF analysis (Delhomme, 1979; Carrera and Neuman, 1986; Dagan, 1987; Bates and Townley, 1988; Bellin and Rubin, 1996; Yeh et al., 2002; Kanso et al., 2003; Gallagher and Doherty, 2007; Farmani et al., 2008). This paper introduces an open source inverse modeling framework, called MAD# (pronounced "mad sharp"), focused on the characterization of SRFs using the Method of Anchored Distributions (MAD), a Bayesian inverse modeling technique (Rubin et al., 2010).

The process of estimating model parameters from the inversion of governing equation(s) and observations is called inverse modeling. For over fifteen years, researchers have advocated for the development of flexible and easy-to-use inverse modeling tools, with the understanding that the shortage of such tools hinders the development of comprehensive and credible uncertainty quantification tools (Poeter and Hill, 1997, 1999; Rubin, 2004; Dagan, 2011). Carrera et al. (2005) identified five features that are needed for broad adoption of inverse modeling tools in hydrogeology: 1) incorporating geological data, 2) improving the flexibility of the code and procedures to handle any and all relevant data types, 3) a

Acronyms: FM, Forward Model; FMD, Forward Model Driver; GIS, Geographic Information System; MAD, Method of Anchored Distributions; MAD#, Method of Anchored Distributions C# Program; pdf, Probability Density Function; RFG, Random Field Generator; RFGD, Random Field Generator Driver; SRF, Spatial Random Field.

* Corresponding author.
E-mail address: carosocali@gmail.com (C.A. Osorio-Murillo).

http://dx.doi.org/10.1016/j.envsoft.2015.01.002
1364-8152/

complete quantification of uncertainty, 4) reducing the difficulty of code operation, and 5) coupling inverse modeling techniques with a geographic information system (GIS) platform.

A number of existing simulation model software tools include model parameter estimation and uncertainty characterization as embedded functions within the program. For example, WEAP (Yates et al., 2005) and PMWIN (Chiang and Kinzelbach, 2001) both are applications that use forward models (FMs) and model parameter estimation software applications like PEST (Doherty, 1994). These and related software tools have aided adoption of uncertainty characterization and inverse modeling to some degree. However, we recognize a need for additional tools that provide a more general set of capabilities and that address the issues raised by Carrera et al. (2005).

MAD has been shown by Rubin et al. (2010), Murakami et al. (2011), and Chen et al. (2012) to be a flexible stochastic inverse modeling technique that addresses the first three challenges posed by Carrera et al. (2005). Specifically, MAD can account for geology (Challenge #1) via the representation of geological features through SRFs modeled using structural parameters; handles multiple relevant data types (Challenge #2) through use of direct measurements and measurements that are indirectly related to the variable modeled; and accommodates uncertainty (Challenge #3) by explicitly incorporating observation uncertainties and quantifying uncertainty of geostatistical structural parameters and a new concept called "anchors".

1.2. Research goals

We have endeavored to address Carrera's Challenges #4 and #5 by implementing and testing MAD in an extensible, user-friendly software framework. Specific goals for the developed framework include:

1) It should be capable of generically accommodating FMs that relate target variables with observations.
2) It also should be flexible in supporting the use of other user-specified software packages for random field generators (RFGs).
3) It should be able to characterize the uncertainty associated with SRFs.
4) It should be well documented and transparent with independently verifiable results.

The remainder of this paper presents our approach to meeting the research goals noted above in the form of an open source inverse modeling software framework called MAD#. This new inverse modeling application builds upon a prototype architecture (Osorio et al., 2012), in which MAD was implemented as a HydroDesktop (Ames et al., 2012) plugin using an embedded steady-state head solver written in R statistical software. MAD# is a standalone desktop application and includes an architecture for adding custom random field generator drivers (RFGDs) and forward model drivers (FMDs) for incorporating new models. We present an architectural overview of MAD# and descriptions of drivers currently implemented. We also present a demonstration of MAD# in two synthetic pumping experiments using a MODFLOW (Harbaugh, 1996) project created in the PMWIN MODFLOW interface (Chiang and Kinzelbach, 2001).

The work presented here is related to an active area of research and development within the broader context of integrated environmental modeling in that our software framework is indeed a method for "integrating" different modeling software packages into a single cohesive environment. This approach is related to the approach supported by as OpenMI (Castronova et al., 2013; Knapen et al., 2013). Ridler et al. (2014) follow a strikingly similar approach to developing a data assimilation framework using OpenMI and an open data assimilation library, using the C# programming language. Another model integration framework that is rapidly growing in adoption is the Community Surface Dynamics Modeling System (CSDMS) which uses a "wrapper-style" common modeling interface approach which is similar in nature and purpose to the forward model driver approach we present herein (Overeem et al., 2013; Castronova and Goodall, 2010).

It is worth noting that the OpenMI and CSDMS approaches both presume the existence of software packages that implement particular hydrologic or environmental numerical models. In other words, these models exist as software packages that require specific input and output files – not simply as conceptual mathematical models. A somewhat different approach has been taken by integrated modeling efforts such as The Object Modeling System (OMS) which is suited to integrating small functions or codes that represent individual physical processes rather integrating large software packages (David et al., 2013).

In comparison to the OpenMI and CSDMS integrated modeling systems, our MAD# approach is relatively simple. Rather than facilitating the transmission of inputs and outputs between various numerical model packages, our approach focuses on a tightly managed system of a single model package connected through a single wrapper (or "driver") directly to our inversion software. This is much more manageable than the alternative of linking multiple models to multiple models, and helps us avoid the challenges and issues of the so-called "Integronsters" (Voinov and Shugart, 2013).

2. Methods

2.1. MAD theoretical background

Although a complete description of MAD is outside the scope of this paper, a brief introduction to the method is presented here. MAD is a Bayesian inverse modeling technique focused on characterizing SRFs by using Bayes' theorem and the following concepts intended to address the challenges stated in the previous section:

- Geostatistical models are used to capture large-scale trends and reproduce patterns of spatial variability in terms of SRFs.
- Data classification – MAD classifies data (measurements) in a general format that is not limited (or specific to) any particular discipline or application. MAD categorizes data as:
  - Type A data, $z_a = y(x_i) + \varepsilon_a$, $i = 1, \ldots, N$, which could include direct measurements (including measurement error ε) of the target variables (e.g. hydraulic conductivity) at location x_i, $i = 1, \ldots, N$, or other types of measurements (e.g. transmissivity) at x_i that could be directly related to the target variable at x_i,
  - Type B data, $z_b = M(x_i) + \varepsilon_b$, $j = 1, \ldots, M$, which include all measurements (including measurement error ε) that do not conform with Type-A, but are related to the target variable via a forward model, M (e.g. pressure head)
- Localization through anchored distributions (or "anchors"). An anchor is a statistical distribution of a target variable at a given location. Anchors can be employed for multiple target variables and/or locations. Anchors intend to capture local effects in the field of the target variables by conditioning realizations on fields.

MAD defines a target variable as a SRF, which is represented by a vector of geostatistical structural parameters (θ) capturing the global tendency, and anchors (ϑ) for quantifying local variations of the parameter field. MAD relies on the following proportionality (Rubin et al., 2010).

$$p(\theta, \vartheta | z_a, z_b) \propto p(\theta, \vartheta | z_a)\, p(z_b | \theta, \vartheta, z_a) \tag{1}$$

Where p indicates a probability density function (pdf) and $p(\theta, \vartheta | z_a)$ is the joint prior distribution of the structural parameters and anchors conditional on Type-A data vector z_a, and $p(z_b | \theta, \vartheta, z_a)$ is the likelihood of observing the Type-B data vector z_b given the structural parameters, anchors and Type-A data. Finally, $p(\theta, \vartheta | z_a, z_b)$ is the joint posterior distribution of the structural parameters and anchors conditional on both Type-A and Type-B data.

2.2. MAD methodological approach

MAD is applied in three stages: 1) Strategy, 2) Implementation, and 3) Assessment. These three stages are described in Fig. 1 and are discussed in depth in the following three subsections.