Computational Statistics and Data Analysis 55 (2011) 1638–1655

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Mechanism-based emulation of dynamic simulation models: Concept and application in hydrology

P. Reichert [a,e,*], G. White [b,e], M.J. Bayarri [c,e], E.B. Pitman [d,e]

[a] *Eawag, Swiss Federal Institute of Aquatic Science and Technology, Department of Systems Analysis, Integrated Assessment and Modelling, 8600 Dübendorf, Switzerland*

[b] *North Carolina State University, Department of Statistics, Raleigh, NC 27695, USA*

[c] *University of Valencia, Department of Statistics, 46010 Valencia, Spain*

[d] *University of Buffalo, Department of Mathematics, Buffalo, NY 14260, USA*

[e] *Statistical and Applied Mathematical Sciences Institute (SAMSI), Research Triangle Park, NC 27709, USA*

ARTICLE INFO

Article history:
Received 23 December 2008
Received in revised form 1 October 2010
Accepted 7 October 2010
Available online 26 October 2010

Keywords:
Dynamic model
Emulator
Optimization
Sensitivity analysis
Statistical inference

ABSTRACT

Many model-based investigation techniques, such as sensitivity analysis, optimization, and statistical inference, require a large number of model evaluations to be performed at different input and/or parameter values. This limits the application of these techniques to models that can be implemented in computationally efficient computer codes. Emulators, by providing efficient interpolation between outputs of deterministic simulation models, can considerably extend the field of applicability of such computationally demanding techniques. So far, the dominant techniques for developing emulators have been priors in the form of Gaussian stochastic processes (GASP) that were conditioned with a design data set of inputs and corresponding model outputs. In the context of dynamic models, this approach has two essential disadvantages: (i) these emulators do not consider our knowledge of the structure of the model, and (ii) they run into numerical difficulties if there are a large number of closely spaced input points as is often the case in the time dimension of dynamic models. To address both of these problems, a new concept of developing emulators for dynamic models is proposed. This concept is based on a prior that combines a simplified linear state space model of the temporal evolution of the dynamic model with Gaussian stochastic processes for the innovation terms as functions of model parameters and/or inputs. These innovation terms are intended to correct the error of the linear model at each output step. Conditioning this prior to the design data set is done by Kalman smoothing. This leads to an efficient emulator that, due to the consideration of our knowledge about dominant mechanisms built into the simulation model, can be expected to outperform purely statistical emulators at least in cases in which the design data set is small. The feasibility and potential difficulties of the proposed approach are demonstrated by the application to a simple hydrological model.

1. Introduction

For computationally demanding models the application of systems analytical techniques, such as optimization, sensitivity analysis, and statistical inference, may be hampered by the required number of model evaluations. A particularly important example is the numerical implementation of Bayesian inference by Markov chain Monte Carlo

* Corresponding author at: Eawag, Swiss Federal Institute of Aquatic Science and Technology, Department of Systems Analysis, Integrated Assessment and Modelling, 8600 Dübendorf, Switzerland.
E-mail address: reichert@eawag.ch (P. Reichert).

0167-9473/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2010.10.011

techniques (Gelman et al., 2004; Gamerman, 2006). This is a problem, because the combination of information from data with prior knowledge (to account for the poor identifiability of the parameters of many complex simulation models), and the need of addressing model deficiency, require the use of Bayesian techniques (Kennedy and O'Hagan, 2001; Bayarri et al., 2007b).

There are two obvious strategies for dealing with this problem: improving the efficiency of the evaluation of the model (including simplifying the model if possible) and improving the efficiency of the computationally demanding technique to be applied. As an additional option, computationally demanding model runs can be replaced by the evaluation of an efficient emulator of the model. An emulator is a statistical approximation of a deterministic model (O'Hagan, 2006). It provides a probability distribution expressing the combination of our prior knowledge of the simulation results with the information gained from runs of the full model at carefully chosen design parameter values. As there is no "measurement error" when emulating a deterministic simulation model, there is no uncertainty in emulation results corresponding to design parameter values. Note that we can also use the emulator as a deterministic interpolator by taking the mean of its probability distribution. As the evaluation of the emulator is (designed to be) fast, replacing the model by the emulator makes computationally expensive systems analytical techniques available even for computationally demanding models. For this reason, emulators of deterministic models have gained considerable attention in the literature over the recent years (Sacks et al., 1989; Currin et al., 1991; O'Hagan, 1992; Craig et al., 2001; Kennedy and O'Hagan, 2001; Oakley and O'Hagan, 2002; Santner et al., 2003; O'Hagan, 2006; Bayarri et al., 2007b,a). Most of these emulators are based on the formulation of a prior of the model output or of residual terms of a linear combination of adequately chosen basis functions in the form of a Gaussian stochastic process (GASP).

The general statistical emulators which were in the focus of development and application in the past can, in principle, be (and have been) applied to dynamic models as well. However, they have two significant deficiencies: (i) they do not consider our knowledge of the structure of the model, and (ii) they run into numerical difficulties if there are a large number of closely spaced input points as is often the case in the time dimension of dynamic models. The first point can be assumed to be a disadvantage for situations in which the design set is small — more specifically, when the design set has to be small due to computational limitations. It can then be expected that consideration of the mechanisms built into the model could significantly improve emulator performance. For these reasons, there is a need for the development of mechanism-based emulators of dynamic models. This is the topic of this paper.

The output of dynamic simulation models consists of a sequence of outputs at different points in time. Often, the output at the next point in time depends only on the previous output, the input and the model parameters. This is in particular true for state space models, if we follow the behavior of all states calculated by the model (however, it is typically not the case for the observations). Four approaches have been proposed for developing emulators of dynamic simulation models:

1. The Gaussian process methodology can be applied to dynamic models by treating the time dimension as an additional model input (see the references given above). However, the dense coverage of the time dimension of the output can lead to numerical problems of emulation with Gaussian processes.
2. The observations can be described by a truncated series expansions with carefully selected basis functions. The Gaussian process methodology can then be applied to the coefficients of this expansion instead of the observations directly. This idea was successfully applied using Fourier series, principal component approximations (Higdon et al., 2007), and wavelet decompositions (Bayarri et al., 2007a) of the observations.
3. If the model output at a given point in time depends only on the output at the previous point in time, input and model parameters, an emulator of the dynamic simulation model can be constructed by emulating the transfer function of output from one point in time to the next (Bhattacharya, 2007; Conti et al., 2009). This significantly reduces the dimensionality of the problem but requires emulation of the complete state space and explicit emulation as a function of driving forces in input in addition to model parameters.
4. Finally, for a dynamic model, a dynamic stochastic model with innovation terms in the form of Gaussian stochastic processes can be used as a prior. This strategy has been proposed based on a general autoregressive model (Liu, 2007; Liu and West, 2009). In this approach, the innovations have to correct for the bias of the autoregressive model and the effects resulting from driving forces represented by input.

All of the methodologies for constructing emulators described above are statistical descriptions of the simulation model that rely exclusively on its input–output behavior (with the possible partial exception of a model-specific choice of basis functions for constructing the mean of the Gaussian stochastic process). This statistical approach has the advantage that the methodologies are general and applicable to any simulation model. On the other hand, not incorporating available knowledge regarding the mechanisms described by the simulation model may be a significant disadvantage of these approaches. It can be expected that consideration of the mechanistic structure of the simulation model could lead to significant improvement of the performance of the emulator; this is especially true in situations in which the design data set is small.

In this paper we propose an alternative procedure for constructing emulators, one that uses knowledge about the mechanisms modeled in the simulation. This approach is similar to approach 4 above. It is based on the formulation of a simplified, linear state space model that describes the approximate dynamics of the model as a function of input and model parameters. As in the statistical approach mentioned above (Liu, 2007; Liu and West, 2009), this emulator construction combines this knowledge together with a Gaussian stochastic process for the innovations as a function of model parameters