Advances in Water Resources 51 (2013) 457–478

Contents lists available at SciVerse ScienceDirect

Advances in Water Resources

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

Hydrologic data assimilation using particle Markov chain Monte Carlo simulation: Theory, concepts and applications

Jasper A. Vrugt [a,b,*], Cajo J.F. ter Braak [c], Cees G.H. Diks [d], Gerrit Schoups [e]

[a] Department of Civil and Environmental Engineering, University of California, Irvine, USA
[b] Institute for Biodiversity and Ecosystem Dynamics, University of Amsterdam, Amsterdam, The Netherlands
[c] Biometris, Wageningen University and Research Centre, 6700 AC, Wageningen, The Netherlands
[d] Center for Nonlinear Dynamics in Economics and Finance, University of Amsterdam, The Netherlands
[e] Department of Water Management, Delft University of Technology, Delft, The Netherlands

ARTICLE INFO

Article history:
Available online 21 April 2012

Keywords:
Particle filtering
Bayes
Sequential Monte Carlo
Markov chain Monte Carlo
DREAM
Parameter and state estimation

ABSTRACT

During the past decades much progress has been made in the development of computer based methods for parameter and predictive uncertainty estimation of hydrologic models. The goal of this paper is two-fold. As part of this special anniversary issue we first shortly review the most important historical developments in hydrologic model calibration and uncertainty analysis that has led to current perspectives. Then, we introduce theory, concepts and simulation results of a novel data assimilation scheme for joint inference of model parameters and state variables. This Particle-DREAM method combines the strengths of sequential Monte Carlo sampling and Markov chain Monte Carlo simulation and is especially designed for treatment of forcing, parameter, model structural and calibration data error. Two different variants of Particle-DREAM are presented to satisfy assumptions regarding the temporal behavior of the model parameters. Simulation results using a 40-dimensional atmospheric "toy" model, the Lorenz attractor and a rainfall–runoff model show that Particle-DREAM, P-DREAM$_{(VP)}$ and P-DREAM$_{(IP)}$ require far fewer particles than current state-of-the-art filters to closely track the evolving target distribution of interest, and provide important insights into the information content of discharge data and non-stationarity of model parameters. Our development follows formal Bayes, yet Particle-DREAM and its variants readily accommodate hydrologic signatures, informal likelihood functions or other (in)sufficient statistics if those better represent the salient features of the calibration data and simulation model used.

Published by Elsevier Ltd.

1. Introduction and scope

Hydrologic models, no matter how sophisticated and spatially explicit, aggregate at some level of detail complex, spatially distributed vegetation and subsurface properties into much simpler homogeneous storages with transfer functions that describe the flow of water within and between these different compartments. These conceptual storages correspond to physically identifiable control volumes in real space, even though the boundaries of these control volumes are generally not known. A consequence of this aggregation process is that most of the parameters in these models cannot be inferred through direct observation in the field, but can only be meaningfully derived by calibration against an input–output record of the catchment response. In this process the parameters are adjusted in such a way that the model approximates as closely and consistently as possible the response of the catchment over some historical period of time. The parameters estimated in this manner represent effective conceptual representations of spatially and temporally heterogeneous watershed properties.

Fig. 1 provides a schematic overview of the resulting model calibration problem. In this plot, the symbol $\oplus$ represents the observation process that provides n measurements of forcing, $\mathbf{u}_{1:n} = \{u_t; t = 1,\ldots,n\}$ (observed input) and output $\tilde{\mathbf{y}}_{1:n} = \{\tilde{y}_t; t = 1,\ldots,n\}$ (observed response). These measurements may deviate significantly from their actual values due to measurement error and uncertainty. The square box represents the conceptual model with functional shape $f(\cdot)$ which is only an approximation of the underlying system (the curly box) it is trying to represent. The label *output* on the y-axis of the plot on the right hand side can represent any time series of data; in this paper we consider it to be the streamflow response and represent this with the n-dimensional vector $\mathbf{y}_{1:n} = \{y_t;\ t = 1,\ldots,n\}$ for the model output (simulated response) and $\tilde{\mathbf{y}}_{1:n} = \{\tilde{y}_t; t = 1,\ldots,n\}$ for the measurements (observed response).

The predictions of the model, $\mathbf{y}_{1:n}$ (indicated with the gray line) are behaviorally consistent with the observations, $\tilde{\mathbf{y}}_{1:n}$ (dotted line),

* Corresponding author at: Department of Civil and Environmental Engineering, University of California, Irvine, USA.
E-mail address: jasper@uci.edu (J.A. Vrugt).

0309-1708/$ - see front matter Published by Elsevier Ltd.
http://dx.doi.org/10.1016/j.advwatres.2012.04.002

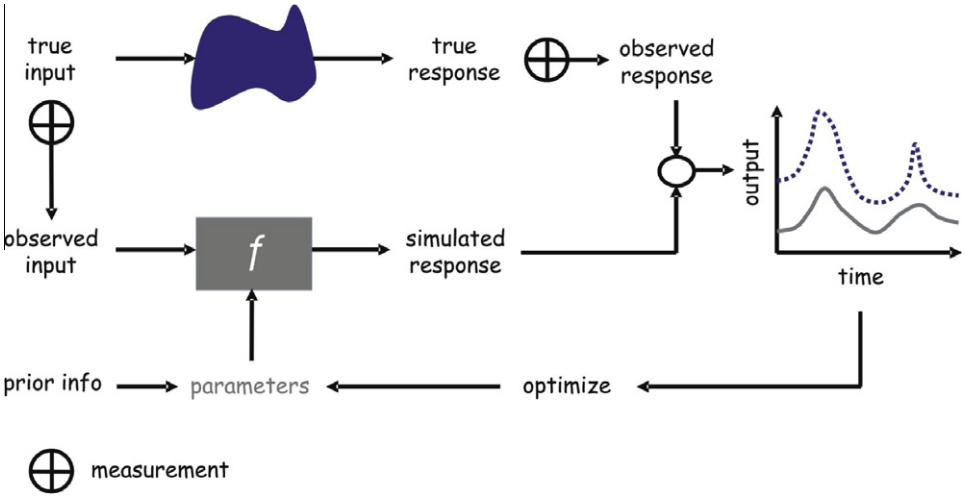

Fig. 1. Schematic overview of the model calibration problem: the model parameters are iteratively adjusted so that the predictions of the model, *f*, (represented with the solid line) approximate as closely and consistently as possible the observed response (indicated with the dotted line).

but demonstrate a significant bias towards lower streamflow values. A better compliance between model and data can be achieved by tuning the model parameters. If we write the dynamic nonlinear watershed model in a state-space formulation, the model calibration problem considered herein (Fig. 1) can be expressed as follows:

$$\mathbf{x}_{t+1} = f(\mathbf{x}_t, \theta, \mathbf{u}_t) + \omega_t, \tag{1}$$

where $f(\cdot)$ is the (nonlinear) model operator expressing the watershed transition in response to forcing data $\mathbf{u}_t$ (rainfall, and potential evapotranspiration), model parameters, θ and state variables, $\mathbf{x}_t$. The variable ω_t represents errors in the model formulation, but is completely neglected in classical model calibration studies. In the remainder of this paper, we assume that the parameter space is bounded, $\theta \in \Theta \in \mathbb{R}^d$ and that the state space, Ω is of dimension $\mathbf{x}_t \in \Omega \in \mathbb{R}^\lambda$. Examples of system states within the context of watershed modeling include but are not limited to (spatially distributed) measurements of soil moisture content, pressure head, and groundwater table depth.

The measurement operator, $h(\cdot)$, defines the observation process and projects the model states, $\mathbf{x}_{t+1}$ to the model output, $\tilde{y}_{t+1}$ (observed response):

$$\tilde{y}_{t+1} = h(\mathbf{x}_{t+1}, \phi) + v_{t+1}, \tag{2}$$

where v_{t+1} denotes the measurement error and ϕ stores any additional measurement parameters. This equation is quite popular and used in many textbooks and publications, but assumes that the system state at $t+1$ contains all necessary information to accurately predict the quantity of interest, y_{t+1}. On the contrary, if the calibration data consists of some time-averaged variable, then knowledge of the system (watershed) prior to $t+1$ is required, and thus an alternative formulation of Eq. (2) is warranted.

To establish whether $f(\cdot)$ provides an accurate description of the underlying system it is intended to represent, it is common practice to compare the simulated system behavior $\mathbf{y}_{1:n}(\theta)$ with respective observations $\tilde{\mathbf{y}}_{1:n}$. The difference between both is encapsulated in a residual vector $\mathbf{E}_{1:n}(\theta)$:

$$\mathbf{E}_{1:n}(\theta) = G[\tilde{\mathbf{y}}_{1:n}] - G[\mathbf{y}_{1:n}(\theta)] = \{e_1(\theta), \ldots, e_n(\theta)\}, \tag{3}$$

where $G[\,\cdot\,]$ allows for linear and nonlinear transformations of the model predictions and observational data. Examples include square root, Box-Cox [11], and normal quantile transformations [87,61], and the use of flow duration curves [116,10], spectral analysis [82,89] and wavelet spectral analysis [113,98,26]. Unfortunately, it is not particularly easy to work with the n-dimensional vector of error residuals, $\mathbf{E}_{1:n}$ directly and find the preferred parameter values. Instead, it is much easier to summarize the vector of residuals into a single measure of length, F, also called objective function. The goal of model calibration then simply becomes finding those values of θ that minimizes (maximizes, if appropriate) this criterion.

During the past 4 decades much progress has been made in the fitting of hydrologic models to data. That research has primarily focused on six different issues: (1) the development of specialized objective functions that appropriately represent and summarize the residual errors between model predictions and observations [53,106,66,5,57,99], (2) the search for efficient optimization algorithms that can reliably solve the watershed model calibration problem [133,31,12,75,136,102,119,77,62,123,112,125,88,130], (3) the determination of the appropriate quantity and most informative kind of data [67,107,44,135,124], (4) the selection and development of efficient and accurate numerical solvers for the partially structured differential and algebraic equation systems of hydrologic models [56,59,60,100], (5) the representation of uncertainty [7,9,65,94,111,134,34,45,69,117,118,132,16,35,80,85,86,8,51,57,58,64,76,103,122,1,36,37,126,127,20,81,109,97,105,128,129-,63,93,24], and (6) the development of inference techniques for refining the structural equations of hydrologic models [138,121,14,139].

Despite the progress made, increasing concern is surfacing in the hydrologic literature that the "classical" approach to model calibration introduced by *Carl Friedrich Gauss* (1794) has some serious deficiencies that necessitate the development of a more powerful paradigm. One of these deficiencies is that model structural and forcing (input) data errors are assumed to be either "negligibly small" or to be somehow "absorbed" into the error residuals, $\mathbf{E}_{1:n}(\theta)$. The residuals are then expected to behave statistically similar as the calibration data measurement error. Yet, many contributions to the hydrologic literature have demonstrated that this assumption is unrealistic. This is evidenced by error residuals that typically exhibit considerable variation in bias, variance, and correlation structures at different parts of the model response. Another deficiency is that the use of a single performance metric, F, no matter how carefully chosen, is inadequate to extract all information from the available calibration data. The use of such "insufficient statistic" promotes equifinality, and makes it unnecessarily difficult to find the preferred parameter values.