Pragmatic Approach to Calibrating Distributed Parameter Groundwater Models from Pumping Test Data Using Adaptive Delayed Acceptance MCMC

Tiangang Cui1; Nicholas Dudley Ward2; Simon Eveson3; and Timo Lähivaara4

Abstract: Calibration of distributed parameter groundwater models in the Bayesian framework using Markov-chain Monte Carlo (MCMC) random sampling is often hampered by the large number of simulations required to make reliable uncertainty estimates. In particular, naive application of the ubiquitous random walk metropolis Hastings (MH) algorithm can take an unsatisfactorily long time to draw samples from the posterior distribution and hence make the required uncertainty estimates. This note addresses the issue of obtaining feasible uncertainty estimates using accelerated MCMC. A pragmatic approach is investigated, based on the adaptive delayed acceptance MH algorithms of Cui et al. First, adoption of an appropriate prior model over the parameters indicates that the number of estimated parameters can be reduced from a thousand parameters to several tens without essential loss of information. Secondly, the algorithm is initialized by a least squares estimate and the covariance of the parameters approximated by the Hessian of the objective function, which is then taken to be an initial proposal distribution for the MH algorithm. The method is evaluated with a numerical simulation, in which the calibration time is reduced five fold compared with previous results of the authors. DOI: 10.1061/(ASCE)HE.1943-5584.0001267. © 2015 American Society of Civil Engineers.

Author keywords: Groundwater; Pumping tests; Parameter estimation; Bayesian inference; Uncertainty quantification; Markov-chain Monte Carlo.

Introduction

In Cui et al. (2014a) the authors considered the problem of estimating aquifer parameters for a groundwater model with spatially distributed parameters from pumping test data. The model was solved using the finite element method, where the number of free parameters was determined by the number of elements in the computational mesh. The associated inverse problem of characterizing the model parameters from data without any further hypothesis is ill-posed (does not have a solution) and is obviously of very high dimension. Given the very small number of observations usually available from a pump test (≤ 5), it does not make sense prima facie to solve a high dimensional problem. In contrast, some kind of regularization of the parameter space is required to ensure the existence of a solution. This is frequently carried out in a least squares setting using Tikhonov regularization. The main objection to working in a least squares framework is that the regularizing parameter has no definite physical interpretation. The other objection is that least squares does not provide a suitable platform for quantifying uncertainty. Although there are situations in which a reliable estimate of uncertainty can be given in a least squares context, e.g., when there is no model error, see Dudley Ward and Kaipio (2014), it is nevertheless very easy to construct examples in which a least squares estimate is orders of magnitude from the actual value (Dudley Ward and Kaipio 2014; Dudley Ward and Fox 2012). On these grounds, the Bayesian framework merits attention because it takes a more serious approach to modeling uncertainties and providing a platform for quantifying posterior uncertainty. In this setting, it is necessary to build two stochastic models. The first is the likelihood model which measures the probability that the model generates the observed data over all permitted parameter choices. The second ingredient is the prior model which plays the part of the regularizer in the least squares setting. In theory, the prior distribution encodes all prior knowledge about the unknowns—a priori to any measurements. However, in practice, in groundwater there is often not enough physical information that might determine the choice of prior so that the actual choice may depend as much on computational issues as much as physical intuition. In the case of a pumping test, it is reasonable to suppose that the data will only provide information on the model parameters in a neighborhood of the observation wells, whereas outside this neighborhood the estimated parameters will be dominated by the prior model, i.e., the data gives little or no information about the parameters.

Usually, a prior distribution for a spatially distributed parameter considers the spatial covariance of parameters over the computational domain, and hence some structure is imposed on the parameter space. A covariance function can be directly specified in terms of an explicit function (e.g., Gaussian or exponential prior) or as a smoothness prior based on a differential operator (Rubin 2003; Kaipio and Somersalo 2005). As in Cui et al. (2014a), we take the second approach because the choice of smoothness prior...
permits very significant reduction of the parameter space. Once
the likelihood and prior models have been specified the solution
to the inverse problem is coded in the posterior probability distri-
bution which by Bayes’ theorem is simply the product of the
likelihood and prior distributions. Because the emphasis is on un-
certainty, solutions to the inverse problem are now given as sum-
mary statistics over the posterior distribution.

The final task is to estimate the summary statistics, and this
is usually done by Monte Carlo integration in the form of
Markov-chain Monte Carlo (MCMC), in which typically the
metropolis Hastings algorithm is used. The hitch, as is very well
known, is that MCMC samplers can be very slow because it is
necessary to draw a very large number of samples from the pos-
terior distribution to make reliable estimates of uncertainty, and
this becomes even more challenging when the number of param-
eters is large as in the cases studied in this paper. It is shown in this
note that the number of estimated parameters can be reduced
significantly (from over 1,000 to several 10
s) by truncating the
Karhunen-Loéve (KL) expansion of the covariance of the prior

method is then used in which the initial parameter accep-
tance is determined by the coarse model, whereas the second stage
tolerance is determined by the fine model. Thus the first stage
acceptance is determined by the coarse model, whereas the second stage
acceptance is determined by the fine model. Thus the first stage
accepts as a low cost sieve of potential draws from the posterior
distribution. The use of surrogate models to reduce computational
burden induced by expensive forward models is common. Some
notable applications in hydrology include Marzouk and Najm
(2009), who applied the stochastic colocation method (Marzouk
and Xiu 2009). Also, Liebermann et al. (2010), Cui et al.
(2014e), and Boyce and Yeh (2014) use projection-based reduced
order models, whereas Cui et al. (2011), Mondal et al. (2010), and
Ketelson et al. (2013) use coarse scale models.

In this note these three enhancements are put together and tested
by two numerical simulations.

\section*{Methods}

\section*{Forward Problem}

This sections contains a brief description of the forward problem
and its numerical solution. Denoting the computational domain by
$\Omega$ with boundary $\partial \Omega$ the following forward model is considered:

\begin{equation}
S(x, y) \frac{\partial s(x, y, t)}{\partial t} = \nabla \cdot \left[ T(x, y) \nabla s(x, y, t) \right] + Q \delta(x) \delta(y) \Omega \tag{1}
\end{equation}

\begin{equation}
s(x, y, t) = 0 \quad \text{on} \quad \partial \Omega \tag{2}
\end{equation}

\begin{equation}
s(x, y, 0) = 0 \tag{3}
\end{equation}

where $s = s(x, y, t)$ is drawdown; $T = T(x, y)$ and $S = S(x, y)$ are spatially distributed transmissivities and storativities; $\delta(\cdot)$ is
Dirac delta function; and $Q$ is constant pumping rate over the duration
of pumping.

In this paper, the preceding initial value problem is solved using
the finite element method (FEM) (Zienkiewicz and Taylor 2005). The software package DistMesh (Persson and Strang 2004) was
used to generate a triangular mesh $\Omega_j$ where $j$ is an index over
the elements. It is necessary to take a large computational domain
to avoid boundary affects and to generate the mesh with some care.

In the present study, the computational domain was taken to have a
50 km radius, whereas the nearest observation well was only 50 m
from the pumping well. This indicates that the mesh needs to be
sufficiently resolved near the pumping well, otherwise the solution
will be very inaccurate. To generate the mesh, an inner mesh was

generated on a disc centered at the pumped well with a radius of
1 km. Conditioned on nodes on the inner boundary an outer mesh
was then generated on the outer annulus. The unknown parameters
$(T, S)$ in the inverse problem are specified by setting

\begin{equation}
T(x, y) = T_j \tag{4}
\end{equation}

\begin{equation}
S(x, y) = S_j \tag{5}
\end{equation}

in element $j$. Thus $T$ and $S$ are assumed to be piecewise constant
functions. Therefore, the parameter vector $\theta$ for the inverse
problem is

\begin{equation}
\theta = \begin{pmatrix} T \\ S \end{pmatrix} \in \mathbb{R}^n, \quad n = 2n_e \tag{6}
\end{equation}

where $n_e$ is number of elements. For the FEM forward solver we
write $\tilde{s} = s(\theta)$ to denote the forward map between the discrete
parameter space and the noise-free drawdown.

\section*{From Least Squares to Bayes: Intuition}

The general problem of fitting a model to data is most conveniently
formulated in the Bayesian framework. In the least squares setting
the aim is to minimize the potential

\begin{equation}
V(\theta) = \frac{1}{2} \sum_i [d_i - s_i(\theta)]^2 \tag{7}
\end{equation}

where $\theta = \text{vector of the model parameters}$; and $s_i(\theta) = \text{model prediction at data point } d_i$ and parameter values $\theta$. Usually, for
distributed parameters problems, it is not possible to minimize $V(\theta)$ over the space of all possible parameters. It is, therefore, necessary to
regularize the problem and set limits on the ranges of permitted

\begin{equation}
\end{equation}
parameters. Another potential \( \tilde{V}(\theta) \) is introduced and \( V(\theta) \) is replaced by

\[
V(\theta) = \frac{1}{2} \sum_i [d_i - s_i(\theta)]^2 + \tilde{V}(\theta) \tag{8}
\]

Under appropriate conditions on \( \tilde{V}(\theta) \), \( V(\theta) \) given by Eq. (8) can be minimized and at least a local minimum can be found.

In the case that normal measurement errors may be assumed, the function \( e^{-V(\theta)} \), where \( V(\theta) \) is defined by Eq. (7) and normalized by the variance \( \sigma^2 \) of the measurement error, may be interpreted as a probability distribution, the likelihood, over the parameter space. Similarly, \( e^{-V(\theta)} \) may be constructed as a probability distribution over the parameter space and, in the Bayesian setting, is called the prior distribution of model parameters. The product of these distributions defines another probability distribution and is known as the posterior distribution of parameters relative to measurements. Thus parameters may be estimated as the posterior distribution of parameters relative to measurements.

It is clear that in the case the regularizing potential is carefully constructed, so that \( e^{-V(\theta)} \) may be interpreted as a probability distribution, the general calibration problem falls into a broader probability framework in which the least squares solution to Eq. (8) is simply one of an infinite range of permitted solutions.

A more formal development is given in the next section.

**Inverse Problem: Bayesian Framework**

Summary details of the Bayesian framework for inverse problems are given in this section. Further details can be found in Dudley Ward and Fox (2012), Cui and Dudley Ward (2012), Cui et al. (2014a), Kaipio and Somersalo (2005), and Tarantola (2005).

The likelihood function is the probability distribution of observations for any given set of true parameters, denoted by \( \pi(d|\theta) \), where \( d \) denotes measurement data, and \( \theta \) is the parameter space given by Eq. (6). In this paper additive Gaussian measurement error \( e \) is assumed, so that the observation model is given by

\[
d = s(\theta) + e \tag{9}
\]

where \( d = \) vector of measured drawdowns; and \( e = \) independent and identically distributed sample from a normal distribution with mean 0 and standard deviation \( \sigma \). Denoting the probability distribution by \( \pi_x \) the likelihood model is given by

\[
\pi(d|\theta) = \pi_x[d - s(\theta)] \propto \exp \left\{ -\frac{1}{2\sigma^2} \sum_i [d_i - s_i(\theta)]^2 \right\} \tag{10}
\]

where \( s_i(\theta) = \) predicted drawdown at measurement \( d_i \). Thus the likelihood measures the offset of the model from data. In broad terms the likelihood quantifies the chance that the model, with specific parameters, generates the observations.

The prior model is the stochastic model for all unknowns a priori to any measurements and is denoted by \( \pi(\theta) \). Given the likelihood and prior models, all information about the unknowns once the measurements have been obtained is coded in the posterior distribution \( \pi(\theta|d) \) and is given by Bayes’ formula

\[
\pi(\theta|d) \propto \pi(d|\theta)\pi(\theta) \tag{11}
\]

Solutions to the inverse problem are given as summary statistics over the posterior distribution. Thus parameters may be estimated by averaging over the range of possibilities given by the posterior distribution, or by taking the most likely values [the maximum a posteriori (MAP) estimate]. However, the primary interest is uncertainty quantification and the posterior standard deviations over parameters provide measures of uncertainty.

**Prior Model**

In this paper, the second order smoothness prior of Calvetti et al. (2006) is used (Kaipio and Somersalo 2005). As discussed in the introduction, many choices of prior are feasible, and it is usually not possible to make a choice entirely on physical grounds. In Cui et al. (2014a), it was shown that a second order smoothness prior leads to good reconstructions while permitting significant model reduction without essential loss of information.

In Cui et al. (2014a), the logarithms of transmissivity and storativity were modelled as spatially smooth independent Gaussian Markov random fields (GMRF), with means \( \mu_T \) and \( \mu_S \) and covariances \( \Gamma_T \) and \( \Gamma_S \), respectively, (Rue and Held 2005). Thus sampling was carried out on a computational parameter (the logarithm of the physical parameters), and mapped back to the physical parameters. Because the storativity is bounded by one, it was necessary to make this constraint to ensure sampling over physically sensible parameters. However, it is somewhat more convenient to use a different function of the storativity. In this paper, we map the storativity to the computational space with the inverse error function.

It is further assumed that \( T \) and \( S \) are statistically independent so that \( f(\theta) \sim \pi_G(\theta) = N(\mu_\theta, \Gamma_\theta) \), where

\[
\mu_\theta = \begin{pmatrix} \mu_T \\
\mu_S \end{pmatrix}, \quad \Gamma_\theta = \begin{pmatrix} \Gamma_T & 0 \\
0 & \Gamma_S \end{pmatrix}, \quad \Gamma_\theta^{-1} = L_0^T L_0 \tag{12}
\]

Thus \( N(\mu_\theta, \Gamma_\theta) \) is a multivariate normal distribution with mean \( \mu_\theta \) and covariance \( \Gamma_\theta \).

A smoothness prior model for \( \theta = (T, S) \) can be constructed by discretizing the Laplacian and forming a second order difference matrix \( D \) between adjacent pixels in the image basis [Eqs. (4) and (5)]. Setting

\[
L = \begin{pmatrix} \zeta_T D & 0 \\
0 & \zeta_S D \end{pmatrix} \tag{13}
\]

where \( \zeta_T > 0 \) and \( \zeta_S > 0 \) are scalar parameters, a second order smoothness prior is defined by

\[
\pi_G(\theta) \propto \exp \left\{ -\frac{1}{2} \| L \theta \|^2 \right\} = \exp \left\{ -\frac{1}{2} \theta^T B \theta \right\} \tag{14}
\]

where \( B = L^T L \).

As discussed in Cui et al. (2014a), this generates an improper prior with the undesirable property of infinite prior variance for each variable. The Calvetti, Kaipio, and Somersalo construction consists of specifying \( k \) pixels with an initial prior distribution. Conditioned on these pixels a distribution is constructed based on the discrete Laplacian discussed previously. The product of these distributions defines a proper smoothness prior over the parameters. The exact construction of this proper prior is quite involved, and the reader is referred to Kaipio and Somersalo (2005), pages 79–89, for a full construction. However the essential idea is as follows. Suppose the pixels are reordered so that any choice of parameters \( \theta \) may be written as \( \theta = [\theta', \theta''] \) where \( \theta' \) are the fixed \( k \) values, \( \theta'' \) the unspecified values. The \( k \) specified values have an assumed probability distribution \( \pi_0(\theta') \), say \( \theta' \sim N(\mu_0, \sigma_0^2 I) \). A new proper density is constructed by setting
The distances between the specified pixels define correlation lengths, which can be adjusted to take into account structural prior information.

Model Reduction

Model reduction of the parameter space is a crucial ingredient in the numerical recipe for model calibration proposed in this paper. In simple terms, the computational parameter \( \phi \) is expanded as a sum of weighted eigenvectors of the covariance matrix corresponding to the eigenvalues written in decreasing order. The expansion is then truncated to include only the \( p \)'th highest modes. In this way only the greater part of the total variability of \( \phi \) is considered. This is sometimes referred to as a proper orthogonal decomposition. The expansion falls under the umbrella of the Karhunen-Loève theory, and is often referred to as a KL expansion.

Specifically, the KL expansion of the computational parameters \( \phi \) is given by

\[
\phi = \sum_{i=1}^{2n_e} \sqrt{\lambda_i} w_i \mathbf{e}_i
\]  

where \( \lambda_i, i = 1, \ldots, 2n_e \) are the eigenvalues listed in decreasing order; \( \mathbf{e}_i \) = corresponding eigenvectors; and \( w_i \) = independent and identically distributed random samples from a normal distribution with mean zero and standard deviation one. The eigenvectors have \( n_e \) terms, the number of elements in the mesh, and the total variance is given by the sum of the eigenvalues. The KL expansion [Eq. (16)] amounts to a decorrelation of the random variable \( \phi \) because the random variables \( w_i \) are independent and a fortiori uncorrelated.

The practical relevance of a KL expansion is that a truncated expansion of \( p \) terms explains the proportion

\[
\sum_{i=1}^{p} \lambda_i
\]

of the total variance of \( \phi \). As will be seen in the numerical section, there is no firm rule on how much of the total variance is required to obtain sufficiently accurate reconstructions. Because the data available in a pump test are extremely limited, experience suggests that 80–90% will give very satisfactory reconstructions. This indicates that the parameter space is reduced from over a thousand to several tens of parameters.

Samples of \( \phi \) are then computed from the truncated expansion

\[
\phi \approx \sum_{i=1}^{p} \sqrt{\lambda_i} w_i \mathbf{e}_i
\]  

MAP Estimate

The Gauss-Newton method (Nocedal and Wright 2005) is used to solve the optimization problem, in which the gradient and Gauss-Newton approximation of the Hessian are computed by adjoint models. In this setup, the Hessian is not explicitly calculated and only the matrix product vector (matvec) with the Hessian is evaluated. Interested readers can find more details in Petra and Stadler (2011) and the references therein. In a wide range of problems (Petra and Stadler 2011; Petra et al. 2012; Bui-Thanh et al. 2012), it has been shown that the number of evaluations of the objective function, the gradient, and the matvec with Hessian do not depend on the parameter dimension (that is, the grid resolution). In fact, it depends on the intrinsic dimension of the inverse problems (Spantini et al. 2014; Cui et al. 2014b). Thus, only a fixed number of evaluations of the objective function, the gradient and the matvec with Hessian are required to obtain the MAP, regardless of the grid resolution.

Compared with the cost of MCMC sampling, the cost of computing the MAP is negligible. In the case studies that follow the optimization was run several times with multiple initial points. The optimization results suggest that the posterior is unimodal in both cases.

Markov Chain Monte Carlo

This section contains an account of the adaptive delayed acceptance metropolis Hastings algorithm used in this paper. For the convenience of the reader a brief account of the generic metropolis Hastings algorithm is given (Metropolis et al. 1953; Hastings 1970). The aim of any MCMC random sampler is to draw samples from a target distribution which, in this paper, is the posterior distribution \( \pi(\theta|d) \) to generate a histogram, or estimate summary statistics. In its simplest form the random walk Metropolis Hastings algorithm takes the following form:

Step 0—Initialize: Begin with an arbitrary choice \( \theta = (\theta_1, \ldots, \theta_n) \).

Step 1—Proposal: Propose a random perturbation \( \theta' \) by drawing \( n \) independent samples \( \Delta \theta = (\Delta \theta_1, \ldots, \Delta \theta_n) \) from a normal distribution \( N(0, \tau) \), and set \( \theta' = \theta + \Delta \theta \).

Step 2—Accept/reject: Define the acceptance ratio \( \alpha = \min \{1, [\pi(\theta')/\pi(\theta)]\} \), and accept or reject \( \theta' \) with probability \( \alpha \).

To carry out the last step draw a uniform random number \( U \) between 0 and 1 and set

\[
\Theta_{n+1} = \begin{cases} 
\theta' & \text{if } U \leq \alpha \\
\Theta_n & \text{otherwise}
\end{cases}
\]  

Thus a sequence \( \Theta_n \) of samples is constructed iteratively in which the \( n + 1 \)th sample is either updated or remains the same as the previous sample. Because it is only guaranteed that in the long term the samples approximate samples from the target, the initial samples, the so-called burn-in, needs to be discarded (Dudley Ward and Fox 2012). Although the magnitude of the step can be tuned by adjusting \( \tau \) to change the acceptance rate, the random walk MH algorithm does not take into account underlying properties of the target distribution (e.g., the covariance), so can take an inconveniently long time to draw a sufficient number of samples [an acceptance rate of 20–30% is usually considered acceptable (Roberts and Rosenthal 2009)]. For a general introduction to MCMC see Liu (2001).

Haario et al. (2001) made a significant advance in MH algorithms. Here the idea is to estimate the empirical covariance matrix with an initial MH run, in which the random walk can be correlated with an initial estimate of the covariance matrix. After \( t_0 \) samples are drawn, the covariance matrix is replaced by the empirical covariance matrix \( \Sigma \) and the random perturbation \( \Delta \theta \) drawn from a multinomial normal distribution with zero mean and covariance matrix \( \Sigma \). For low dimensional examples considered, for example, in Dudley Ward and Fox (2012) and Cui and Dudley Ward (2012) the adaptive scheme of Haario et al. makes a very significant improvement on the sampling. In this case, only a very approximate estimate of the posterior covariance is required.

Cui et al. (2011) extended the adaptive MH algorithm to deal with a distributed parameter problem for a geothermal model, in which improvements in computational speed are vital because of the cost of computing the forward map. Here the idea is to use an inaccurate and accurate (high-fidelity) model, in which
the former is significantly faster to compute. Acceptance of a proposal is based on a two stage MH algorithm, in which the inaccurate model provides an initial sieve of candidate parameter values. If a proposal is rejected using the inaccurate model another candidate proposal is generated, although if it is accepted, it passes onto the next stage, and the accurate model evaluated and a modified acceptance ratio computed. It is then accepted or rejected depending on the MH criteria discussed previously (step 3). However, there is a subtlety because there is usually an offset between the posterior distribution using the inaccurate model and the actual posterior distribution. This can be corrected by considering the discrepancy between the accurate and inaccurate model, the so-called approximation error (Kaipio and Somersalo 2005) and estimating the offset statistics. Details are given in the paragraphs that follow.

To define the adaptive delayed acceptance algorithm, the posterior distribution is first reformulated in following form:

\[ \pi(\theta | d) \propto \exp\left\{-\frac{1}{2} \sigma^2 \right\} \pi(\theta) \]  

(20)

where \( \Sigma_e \) = covariance matrix of the observation noise. In the cases considered in this note \( \Sigma_e \) is an \( m \times m \) diagonal matrix \( \text{diag}(\sigma^2, \ldots, \sigma^2) \), whereas \( m \) is the number of observations. In the adaptive delayed acceptance (ADA) algorithm, the high fidelity forward model \( s(\theta) \) is coupled with a low fidelity, significantly faster reduced order model \( s^*(\theta) \) together to reduce the computational cost of the MCMC, while maintaining sampling accuracy. The construction of fast reduced order models—including coarsening the grid (Christen and Fox 2005; Cui et al. 2011) of the high fidelity model or interpolating the high fidelity model at some selected points (Marzouk and Xiu 2009; Laloy et al. 2013)—usually sacrifices numerical accuracy. To capture the error introduced by the reduced order model, the ADA algorithm adapts the approximation error method of (Kaipio and Somersalo 2005), in which the parameter-data relation is modelled by a reduced order model of the form of

\[ d = s^*(\theta) + e + b \]  

(21)

where \( b = \) random variable that captures the statistics of the error between the reduced order model and the high fidelity model, \( s(\theta) - s^*(\theta) \). Here the model error \( b \) is assumed to be normal, \( b \sim N(\mu_b, \Sigma_b) \), and independent of the parameter \( \theta \) to eliminate the difficulties associated with fitting a high dimensional response surface. ADA first constructs the approximate posterior:

\[ \pi_n(\theta | d) \propto \exp\left\{-\frac{1}{2} \left( \sigma_n^2 \right) + \mu_{b,n} - d \right\} \pi(\theta) \times (\Sigma_{b,n} + \Sigma_e)^{-1} \left( s^*(\theta) + \mu_{b,n} - d \right) \pi(x) \]  

(22)

using some initial estimates of the mean \( \mu_{b,n} \) and covariance \( \Sigma_{b,n} \) of the model error, where \( n = 1 \). Here the subscript \( n \) represents the adaptive construction of the approximate posterior. Then ADA employs this approximate posterior [Eq. (22)] to speed up the MCMC sampling and evaluates the high fidelity model occasionally to ensure the resulting samples are drawn from the full posterior [Eq. (20)]. Such high fidelity model evaluations are also used to adaptively improve the estimation of the statistics of the model error \( b \).

At step \( n \) of ADA, suppose that the current state \( \Theta_n = \theta \). Then the next state \( \Theta_{n+1} \) is determined in the following way:

Step 1—Proposal: Generate a proposal \( \theta' \) from some proposal distribution \( q(\theta', \theta) \).

Step 2—First stage accept/reject: With probability

\[ \alpha(\theta', \theta) = \min \left\{ 1, \frac{\pi_n(\theta') | d_q(\theta', \theta) |}{\pi_n(\theta) | d_q(\theta, \theta') |} \right\} \]  

(23)

accept candidate \( \theta' \) by setting \( \Theta_{n+1} = \theta' \) and proceed to Step 3. Otherwise set \( \Theta_{n+1} = \theta \) and proceed to Step 4.

Step 3—Second stage accept/reject: With probability

\[ \beta(\theta, \theta') = \min \left\{ 1, \frac{\pi_n(\theta | d) \pi_n(\theta') | d_q(\theta', \theta) |}{\pi_n(\theta | d) \pi_n(\theta | d_q(\theta, \theta') |)} \right\} \]  

(24)

accept candidate \( \theta' \) by setting \( \Theta_{n+1} = \theta' \) and proceed to Step 4. Otherwise set \( \Theta_{n+1} = \theta \) and proceed to Step 4.

Step 4—Update approximation error statistics: Given the model error \( b_{n+1} = s(\Theta_{n+1}) - s^*(\Theta_{n+1}) \), update the approximate posterior \( \pi_n(\theta | d) \) to \( \pi_{n+1}(\theta | d) \) by updating the mean and covariance

\[ \mu_{b,n+1} = \frac{1}{n+1} (n \mu_{b,n} + b_{n+1}) \]  

(25)

\[ \Sigma_{b,n+1} = \frac{1}{n} \left( (n-1) \Sigma_{b,n} + b_{n+1} b_{n+1}^\top - n \mu_{b,n+1} \mu_{b,n+1}^\top \right) \]  

(26)

In step 4 of the previous algorithm, the high fidelity model evaluations \( s(\Theta_{n+1}) \) and \( s^*(\Theta_{n+1}) \) are precomputed in steps 2 and 3, and thus the cost of updating the approximate posterior is negligible compared with the cost of reduced order model and high fidelity model.

As discussed in the introduction, it is sufficient in practice to initialize the sampler with a MAP-estimate, i.e., the most probable parameter values, which can be obtained through optimization using, for example, the Gauss-Newton method (Nocedal and Wright 2005). Because the Hessian of the objective function at this point approximates the covariance of the posterior (Martin et al. 2012), this can be taken to initialize the estimate of the covariance of the posterior.

### Numerical Experiments

#### Case Study 1

For comparative purposes a numerical experiment was carried out using the same aquifer model as in Cui et al. (2014a). Here, the parameters exhibited a sharp shift in effective diffusion \( T/S \) across the lateral axis. The underlying true parameters were assumed to be \( T_{\text{max}} = 1 \times 10^4 \text{ m}^2/\text{day} \) for \( X < 200 \text{ m} \) and \( T_{\text{min}} = 2 \times 10^3 \text{ m}^2/\text{day} \) for \( X > 500 \text{ m} \). \( T \) decreases linearly between these values over the zone \( 200 < X < 500 \). Storativity varied similarly between \( S_{\text{min}} = 2 \times 10^{-4} \) and \( S_{\text{max}} = 1 \times 10^{-1} \). The model is shown in Fig. 1.

To avoid boundary effects in the numerical solution a circular computational domain of radius 50 km with 772 elements was generated in DistMesh. The interior of the mesh is shown in Fig. 2. The prior mean and standard deviations of the transmissivity were set to be \( 10^3 \text{ m}^2/\text{day} \) and \( 2.3 \times 10^3 \text{ m}^2/\text{day} \), respectively. For the storativity, they were set to be \( 1 \times 10^{-2} \) and \( 2 \times 10^{-2} \), respectively. The prior probability of a storativity of the order of \( 1 \times 10^{-2} \) is highly unlikely because the mass of the prior is centered around \( 1 \times 10^{-1} \).

A pumping test for the well marked in Fig. 1 pumped at \( 3,000 \text{ m}^3/\text{day} \) over five days and monitored for a further 10 days. Synthetic pumping test data were generated for eight wells, four
close labeled O1-O4, and four wells further away labeled P1-P4 in Fig. 1. The true data corrupted with Gaussian white noise with a standard deviation \( \sigma = 0.01 \) m are shown in Fig. 3 for the near wells and in Fig. 4 for the far wells.

The accurate model using quadratic basis elements took 0.4 s to run on a Mac Powerbook, whereas the approximate model using linear basis elements took 0.08 s to run, a five-fold decrease in the computational time.

The MCMC was initialized with the MAP estimate, which took approximately 2 min to compute. Five thousand samples were then drawn using the modified adaptive procedure discussed in the previous section. The total computational time was around 90 min.

In the first experiment model reduction of the prior distribution was carried out to include 90% of the total variability. This resulted in 26 unknown parameters. Fig. 5 shows the estimated predicted mean drawdown and recovery together with the measurement data for the four close observation wells O1-O4. The \( \pm 3 \) standard deviation uncertainty envelopes were omitted because they are almost indistinguishable from the mean predicted responses. It is evident that the calibration gives a very good reconstruction of the data. Figs. 6 and 7 show the parameter reconstructions of the transmissivities and storativities along a longitudinal transept of the domain. The plots on the left of each figure show detail over \( \pm 10 \) km, whereas the plots on the right show detail on the inner \( \pm 1 \) km. It is evident that the reconstruction is good over a neighborhood slightly larger than the region between the observation wells. Outside this interval the behavior of the prior distribution dominates the posterior distribution and the reconstruction first oscillates and then stabilizes to values fairly close to prior mean. As the reconstructions stabilize the uncertainty envelopes widen because the data gives no information about the model parameters outside the neighborhood.

It is interesting to consider the effect of further measurements. Fig. 8 shows the parameter reconstructions over the inner \( \pm 1 \) km using the data from O1-O4 and P1-P4, i.e., both near and far field
observation wells simultaneously. The data fits are similar to the fits in Fig. 5 and are omitted. The general picture is one of uncertainty reduction in which the region of effective reconstruction is now increased to include the outer observation wells.

Finally, the plot in Fig. 9 shows the effect of using the four observation wells O1-O4 together with a KL truncation which includes 99% of the variability. This now results in a parameter estimation problem with 142 variables. It is clear that the extra information included in the prior does very little to improve the parameter reconstructions.

Case Study 2

In this section, another case study is carried out in which pumping test data are generated from a heterogeneous confined aquifer. This case study is even more minimal because measurements are made in only three observation wells. To introduce a degree of model uncertainty the physical model parameters are generated from a GMRF with an exponential covariance function of the form

\[
\exp\left(-\frac{|x_1 - x_2|}{\sigma_x} - \frac{|y_1 - y_2|}{\sigma_y}\right)
\]

(27)

Fig. 4. Pumping test data from the distant observation wells P1-P4

Fig. 5. Mean estimated drawdown and recovery in observation wells O1-O4 (continuous lines) using pump test data from O1-O4 for case study 1
Fig. 6. Estimates of transmissivity through $Y = 0$ using pump test data from wells O1-O4; detail between $x = -1000$ and $1000$ m is shown in the right plot; the solid lines show the reconstructed values, the dots $\pm 3$ standard deviations, and the dashes the actual values.

Fig. 7. Estimates of storativity through $Y = 0$, using pump test data from wells O1-O4; detail between $x = -1000$ and $1000$ m is shown in right plot; the solid lines show the reconstructed values, the dots $\pm 3$ standard deviations, and the dashes the actual values.

Fig. 8. Estimates of transmissivity and storativity through $Y = 0$, using pump test data from wells O1-O4 and P1-P4; the solid lines show the reconstructed values, the dots $\pm 3$ standard deviations, and the dashes the actual values.

Fig. 9. Estimates of transmissivity and storativity through $Y = 0$, using pump test data from wells O1-O4 and prior with 99% variability.
Thus, a different GMRF is used to construct the data from the one used to carry out the inversion. The $x$ and $y$ correlation lengths are set to be $\sigma_x = 6 \times 10^2$ m and $\sigma_y = 1.2 \times 10^3$ m, so that the aquifer is assumed to be anisotropic. Details of the actual transmissivity and storativity field are shown in the upper plots in Fig. 12. The mean of the transmissivity field is $9.84 \times 10^3$ m$^2$/day with a standard deviation of $1.38 \times 10^3$. The mean of the storativity field is $1.14 \times 10^{-4}$ with a standard deviation of $7.7 \times 10^{-5}$.

The central well in Fig. 10 was pumped at 3,000 m$^2$/day over 3 days and monitored for a further two. Synthetic measurements were generated for the three observation wells marked in Fig. 10 and shown by the dots in Fig. 11. To avoid boundary effects in the numerical solution a circular computational domain of radius 10 km with 387 elements was generated using DistMesh.

To set the smoothness prior for the inversion no knowledge of the anisotropy was assumed, and 36 equally spaced control points were fixed over the central 4,000 m by 4,000 m of the computational domain. The prior mean and standard deviation of the transmissivity were set to be $2 \times 10^4$ m$^2$/day and $2.376 \times 10^2$ m$^2$/day, respectively. For the storativity they were set to be $1 \times 10^{-3}$ and $8 \times 10^{-4}$, respectively. The actual mean values are relatively unlikely with respect to the prior.

The following results assumed a KL truncations set at capturing 90% of the variability, corresponding to 22 terms, and hence an estimation problem for 44 parameters. As before, the MCMC was initialized with a MAP estimate. The total computation took 58 min on a desktop powered by an Intel i7-3770 K 3.50 GHz.

Cross sections through the parameter reconstructions are shown in the four lower plots in Fig. 12. It will be observed that the mean estimates are quite close to the MAP estimates through each cross section. The reconstruction of the transmissivity field is slightly better for the $y$ cross section because the underlying data have a longer correlation length (and hence less variability) in this direction. The broad picture is consistent with case study one, namely, in a neighborhood of the measuring points the data permit a fair reconstruction of the data, with reasonably small bounds on the uncertainty. As in the first case study, the uncertainty increases away from the measurement points.

**Discussion**

As discussed in the introduction, the main impediment to the adoption of the Bayesian framework, in particular the use of MCMC algorithms, to calibrating groundwater models is the computational burden. This paper considered the relatively simple situation of calibrating a model with spatially distributed parameters to pumping test data from a small number of test measurements. Compared with the results in Cui et al. (2014a) the augmented metropolis Hastings algorithm described in this paper achieves a significant improvement in the computational time taken to calibrate the model and quantify the parameter uncertainty: from 8–9 h to 90 min for case 1.
Using a MAP estimate to initialize the MCMC is a pragmatic way of ensuring that the sampling begins in a region in which the posterior probability distribution has nontrivial mass. This obviously dispenses with the need for long burn-ins. Although this may be objectionable from a purist Bayesian perspective, the authors carried out a large number of model calibrations with arbitrary initial parameter values for the MAP estimate and very long MCMC runs (>50,000 samples) and found that the proposed numerical recipe does not bias the results. That is, there are not other regions of the posterior distribution that were not explored. In contrast, taking arbitrary initializations for the ADA algorithm (i.e., not using a MAP estimate) can result in very long burn-in runs and is therefore impracticable. The authors’ experience suggests that unless the initial point is quite a small perturbation of the MAP estimate, it can prove challenging to escape even burn-in. The same remark can be made for not having a sufficiently good initial estimate of the posterior covariance. Doing the obvious and taking independent random walk MH does not prove efficacious because one tends to end up with very long runs of rejections.

The results reported on in this note assumed a model reduction of the prior parameter space in which 90% of the variability was accounted for in the reduced prior. As shown, this allowed very acceptable reconstructions of the model parameters. It is also possible to make further model reductions to around 70% of the variability of the variance which would lead to an estimation problem for six parameters and obtain almost the same reconstructions of the predictions of the drawdowns. However, more artifacts are produced in the parameter reconstructions. The Gibbs effects (oscillations) in the reconstructions are greater and the limiting values away from the observation wells are less well behaved. In contrast, inclusion of essentially the whole range of variability does very little to change to reconstruction.

It is unreasonable to expect the test data to give information about the aquifer parameters in a neighborhood that extends much beyond the measurement wells. Outside this region the data give negligible information outside this region, and the reconstruction is constrained by the prior distribution. However, the uncertainty is reflected in the much greater uncertainty envelope.

Unsurprisingly, more observation wells further constrain the parameter reconstruction. However, the picture is consistent: reliable reconstructions are obtained in a neighborhood of the more distant wells. In engineering practice, it is not often the case that more than a few observation wells are available because installing piezometers is usually impracticable because of budgetary constraints.

On purely engineering grounds the proposed recipe would appear to have merit. The most important task is to achieve reliable predictions, using calibrated models with physically sensible

Fig. 12. Transmissivity and storativity fields for case study 2 are shown in the upper two plots; the middle two plots show parameter reconstructions through y = 0, whereas the bottom two plots show parameter reconstructions through x = 0; the solid lines are reconstructed values, the crosses are MAP-estimates, whereas the dots are ±3 standard deviations and the broken lines are the actual values.
parameters. Furthermore, in aquifers which are highly heterogeneous, it does not make sense to fit analytic models, as is frequently done in pumping test analysis.

The techniques proposed in this note also have application to the calibration of large scale (i.e., catchment size) groundwater models. Although the code written to generate the results in this paper was written in MATLAB, rather than a faster low-level programming language, it was optimized to exploit sparse matrices and fast numerical routines. This may not be available with many off-the-shelf packages. However, the importance of reducing the computational expense of a model simulation cannot be underestimated, particularly when the aim is to explore uncertainty. In the experience of the authors, large scale catchment models can be over-solved because of the desire to capture relatively small scale features. For example, refining a mesh to resolve a river/stream network can add very significantly to the computational cost of a single forward run. The degree of resolution needs to take into account the availability and reliability of data. An approach that parallels the one taken in this paper is to build coarse and fine models and carry out the majority of calculations with the coarse model. Using the approximation error method, the statistics of the model discrepancy can be estimated with relatively few evaluations of the fine (expensive) model. These statistics are used to correct the posterior distribution using the approximation model. Usually the uncertainty increases modestly, but the corrected posterior has the desirable feature that it captures (predicts) what actually happens. This approach has been shown to be effective in Lähivaara et al. (2014) and Lähivaara et al. (2015), in which the authors used seismic tomography and full wave inversion to identify aquifer features.

The focus of this note has been pumping tests and the practical engineering problem of estimating both the quantity of groundwater and the rate at which it can be discharged for a given well using very few field measurements. The calibration of integrated catchment scale models poses further problems because the model and data uncertainties provide less constraint on the model parameters and may permit significant multimodal posterior distributions. Thus, the challenge of estimating the posterior uncertainty in these situations may be much greater.

Finally a remark on the deterministic approach to model calibration and uncertainty analysis. In this setting ill-posed inverse problems are solved by regularization methods such as Tikhonov regularization and truncated SVD. Along this line, the null space problem is solved by regularization methods such as Tikhonov regularization. Thus, the challenge of estimating the posterior uncertainty in these situations may be much greater.

Conclusions

This note considered a pragmatic approach to the problem of calibrating a distributed parameter groundwater model from pumping test data in the Bayesian framework using an adaptive delayed acceptance metropolis Hastings algorithm with several enhancements. First the smoothness prior of Calvetti, Kaipio, and Somersalo permitted the number of estimated parameters to be reduced from around a thousand to several tens. The MH algorithm was then initialized with a MAP estimate obtained by minimizing a regularized potential, whereas the Hessian was used as an initial estimate of the posterior covariance. Initial parameter acceptance was determined by a coarse (surrogate) model using linear basis functions, whereas the second stage acceptance was determined by an accurate (or high fidelity)—but relatively expensive—model using quadratic basis functions. Compared with previous work of the authors, the proposed numerical recipe resulted in a five fold saving in computational time.

Acknowledgments

Nick Dudley Ward was partially supported by the New Zealand Ministry of Science and Innovation and by the Väisälä-foundation. Timo Lähivaara has been supported by the Academy of Finland (projects 250215 and 257372).

Notation

The following symbols are used in this paper:

$N(\mu, \Gamma) =$ multivariate random variable with mean $\mu$ and covariance matrix;

$S = S(x, y) =$ spatially distributed storage coefficient;

$s = s(x, y, t) =$ drawdown;

$T = T(x, y) =$ spatially distributed transmissivity;

$t =$ time;

$Q =$ pumping rate;

$x^T =$ denotes the transpose of $x$;

$||x||_2 = <x, x>$, where $<\cdot, \cdot>$ is the usual inner product on $\mathbb{R}^n$; and

$(x, y) =$ spatial coordinates.

References


© ASCE

06015011-11 J. Hydrol. Eng.
“Likelihood-informed dimension reduction for nonlinear inverse problems.” Inverse Prob., 30(11), 114015.
MATLAB [Computer software]. Natick, MA, MathWorks.