MS# 001810

Uncertainty Reduction and Characterization of Complex Environmental Fate and Transport Models: An Empirical Bayesian Framework Incorporating the Stochastic Response Surface Method

Suhrid Balakrishnan†,‡, Amit Roy‡, Marianthi G. Ierapetritou†,

Gregory P. Flach°, Panos G. Georgopoulos†,‡[1]

†Department of Chemical and Biochemical Engineering,

Rutgers University, Piscataway, NJ 08854, USA

‡Environmental and Occupational Health Sciences Institute, UMDNJ - R.W. Johnson Medical School and Rutgers University, Piscataway, NJ 08854, USA

°Savannah River Technology Center, Savannah River Site, Aiken, SC 29808, USA

WRR Section Heading: SUBSURFACE HYDROLOGY

[1] Author to whom all correspondence should be addressed: Phone: 732-445-0159; Email: panosg@fidelio.rutgers.edu

Abstract: In this work, a computationally efficient Bayesian framework for the reduction and characterization of parametric uncertainty in computationally demanding environmental 3-D numerical models has been developed. The framework is based on the combined application of the Stochastic Response Surface Method (SRSM, which generates accurate and computationally efficient statistically equivalent reduced models) and the Markov Chain Monte Carlo method. The application selected to demonstrate this framework involves steady state groundwater flow at the U.S. Department of Energy Savannah River Site General Separations Area, modeled using the Subsurface Flow And Contaminant Transport (FACT) code. Input parameter uncertainty, based initially on expert opinion, was found to decrease in all variables of the posterior distribution. The joint posterior distribution obtained was then further used for the final uncertainty analysis of the stream baseflows and well location hydraulic head values.

Index terms: 6309 - Decision making under uncertainty; 1832 - Groundwater transport; 1869 - Stochastic processes; 3210 - Modeling

Keywords: SRSM, Bayesian Inference, Metropolis Hastings, Markov Chain Monte Carlo (MCMC), Input Distributions, Uncertainty Analysis, Groundwater Flow

1 Introduction

In many environmental applications, complex computer programs are often used to model and analyze real phenomena. The systematic accounting of parametric uncertainty in such models is valuable, as this aids in the quantification of the degree of confidence in model observations and data. Uncertainty almost always exists in such systems in both the observed data and in the parameters themselves (such as recharge rate and conductivity fields in the case of models of groundwater flow) and its disregard may easily lead to ill-fitted or poorly calibrated

models. Additionally, for environmental problems, prior information on physically reasonable values of these parameters is also often available from expert analysis/opinion/experience, as is calibration information in the form of actual field measurements of the model predicted output quantities (observational data). This leads to an interesting and challenging parameter estimation problem, namely, the adjustment of the prior uncertain information in view of the comprehensive environmental model and the observed field data.

The Markov Chain Monte Carlo (MCMC) approach is a class of numerical Bayesian inference methods that address many of the problems associated with the estimation of model parameters for complex nonlinear systems. Bayesian inference is particularly suited to the estimation of parameters in mechanistic models because *a priori* information on physically reasonable values of these parameters can be conveniently incorporated into the parameter estimation process, in the form of informative prior distributions. MCMC methods generate probability distributions rather than point estimates for model parameters, and use a probabilistic framework to "merge" the information contained in observational data with the prior information about model parameters.

In the case of computationally intensive models, often encountered in environmental applications, the time and resources required by MCMC methods can prove to be prohibitively expensive when applied directly to the complex numerical codes themselves (which may have substantial individual simulation run times). The present work utilizes the Stochastic Response Surface Method (SRSM, Isukapalli 1999) in order to provide a statistically equivalent reduced model, which serves as a computationally efficient but sufficiently accurate surrogate for the original model in the numerical Bayesian inference step.

Through SRSM, the statistics of outputs of an uncertain system are approximated by series expansions of standard random variables containing coefficients that can be calculated from the results of a limited number of model simulations. The utility of the reduced model rests in the computationally fast evaluations of approximations of the outputs of the complex code (via polynomial evaluations); this provides a feasible and convenient route to apply the MCMC techniques indirectly to the complex model.

Once the Bayesian inference procedure is complete, and model parametric input uncertainty is estimated in a manner consistent with observed data (which in essence, is an "inverse modeling" task), the resulting joint posterior distribution of the inputs can be utilized to estimate uncertainty in the outputs of the full model. This is once again done accurately and efficiently via the use of the SRSM.

The simulation of saturated groundwater flow beneath the U.S. Department of Energy Savannah River Site General Separations Area (GSA) (Figure 1 and Figure 2), using the Subsurface Flow And Contaminant Transport (FACT) code (Hamm and Aleman 2000), requires model inputs such as three-dimensional conductivity fields and a two-dimensional recharge rate field, in which uncertainty is inherent. However, this uncertainty can be fairly well characterized through expert opinion, based on years of collected data and experience gained through traditional trial and error model calibration. FACT is a complex three-dimensional transport model employing the finite element method for the numerical solution of the continuity equation and Darcy's Law and is computationally quite demanding, averaging about half an hour per FACT code simulation (when run on a Sun Fire 280R with two 750MHz UltraSPARC III CPUs and 4GB memory). Quantities of interest in the uncertainty analysis tasks are model outputs such as hydraulic head values (for which actual data are available from field measurements) and

stream baseflow rates. The methods and results presented in this paper are new developments which build upon and extend those described in the article by Balakrishnan et al. (A Comparative Assessment of Efficient Uncertainty Techniques for Environmental Fate and Transport Models: Application to the FACT Model, submitted to *Journal of Hydrology*, 2003)

The following sections of this paper describe the uncertainty analysis procedure, the Bayesian inference task and its numerical solution via the MCMC technique, followed by the details of SRSM and the case study that employs these methods.

2 Framework for Empirical Bayesian Uncertainty Reduction and Analysis

The task of parametric uncertainty analysis, as used in this work, refers to deterministic computer codes (where the program run twice on the same inputs results in identical outputs being produced) being run over ranges, or distributions of input parameters, thus resulting in corresponding distributions of output quantities of interest. The objective here, is to make inference about the distributions of the outputs of the complex computer code, given a joint distribution of input parameters. It should be noted that the underlying assumption in this type of analysis, is that the complex code accurately or adequately predicts the outputs of interest over the entire range of input conditions. It is thus assumed that model or structural uncertainty (where the model itself is in question) is minimal in the cases examined. When such effects are strong, the proposed procedure may not be the most appropriate and the reader is referred to Draper (1995) and Kennedy and O'Hagan (2001) for discussion/description of alternative treatments. In order to proceed with parametric uncertainty analysis of this type, it is evident that an input distribution of parameters is necessary. While it is often the case that physical constraints combined with expert opinion are sufficient to specify these input distributions and

hence successfully carry out an uncertainty analysis, it is often more desirable to incorporate in a consistent manner any available observed data relating to the actual phenomenon being modeled, in order to obtain a more accurate estimate of the uncertainty in the input parameters - a task that the Bayesian framework is ideally suited to.

In this manner the Bayesian inference task may be viewed as updating/refining uncertainty "encapsulated" in the expert opinion based input distributions (e.g. Brand and Small 1995). Examples of previous work on these lines and some pertinent references include Kennedy and O'Hagan (2001), Currin et al. (1991), Brand and Small (1995), and O'Hagan et al. (1999). Further references that deal with similar inference tasks from an "inverse modeling"/model calibration perspective can be found in Hill (1998) and Porter (2000). While analytical results are tractable for idealized cases where significant simplications in input and model assumptions can be made (for example the linearization of flow equations), the present approach to this problem makes no restrictions on the complexity of the model or process itself and aims to have the input distributions specified in a fairly unrestrictive manner while making posterior (i.e. updated) Bayesian inference numerically via MCMC simulation. This work most notably differs from the previous (and more standard) approaches to Bayesian methods in hydrogeology (Christakos and Li 1998, Christakos et al. 1999; Kitanidis 1986, Kitanidis 1995, Kitanidis 1997; Woodbury and Ulrych 2000, etc.) by providing a common framework to enable the utilization of available complex computer codes (in our case study, a three-dimensional finite element code) in the Bayesian inference task, as well as the inverse modeling task.

2.1 Markov Chain Monte Carlo Simulation

MCMC simulation involves the generation of numerical approximations of a distribution using the probabilities of individual realizations of the distribution. In a Bayesian context, the distribution of interest is the joint posterior distribution of the model parameters $P(\boldsymbol{\theta}|\boldsymbol{d})$, which incorporates both prior information on the model parameters $\boldsymbol{\theta}$, as well as additional information in the form of observed data $\boldsymbol{d}$. The joint posterior distribution of the model is then given by Bayes rule as:

$$P(\boldsymbol{\theta} \mid \boldsymbol{d}) \;\propto\; P(\boldsymbol{d} \mid \boldsymbol{\theta}) P(\boldsymbol{\theta})$$

where $P(\boldsymbol{\theta})$ is the prior distribution and $P(\boldsymbol{d}|\boldsymbol{\theta})$ is the likelihood function for model parameters. In general, $P(\boldsymbol{\theta}|\boldsymbol{d})$ cannot be estimated directly, because it is usually not possible to sample from the likelihood function. However, it is possible to calculate the value of the likelihood function for a given realization of the model parameters, and MCMC exploits this property to generate samples from $P(\boldsymbol{\theta}|\boldsymbol{d})$ when the chain has converged. A sufficiently large number of these samples can be used as a good numerical approximation of $P(\boldsymbol{\theta}|\boldsymbol{d})$. Figure 3 shows a flowchart entailing the steps involved in a typical MCMC simulation.

A number of algorithms exist that allow the construction of the Markov chain. For the purpose of this study, one of the most widely applied algorithms was employed, namely, the Metropolis-Hastings algorithm (Metropolis et al. 1953, Hastings 1970). When using a symmetric proposal distribution $(J_t(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(t-1)}) = J_t(\boldsymbol{\theta}^{(t-1)} \mid \boldsymbol{\theta}^*))$ the Metropolis-Hastings algorithm involves:

1. Obtaining $\boldsymbol{\theta}^{(0)}$, an initial realization of the parameter vector.
2. Obtaining $\boldsymbol{\theta}^{(t)}$, the t^{th} realization of the parameter vector by :

a. Sampling from a proposal distribution, $J_t(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t-1)})$, to obtain a candidate parameter vector $\boldsymbol{\theta}^*$,

b. Calculating an acceptance probability α as:

$$\alpha = \frac{P(\boldsymbol{\theta}^*|\boldsymbol{d})}{P(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{d})} = \frac{P(\boldsymbol{d}|\boldsymbol{\theta}^*)P(\boldsymbol{\theta}^*)}{P(\boldsymbol{d}|\boldsymbol{\theta}^{(t-1)})P(\boldsymbol{\theta}^{(t-1)})} \tag{1}$$

c. Accepting the candidate point as per

$$\boldsymbol{\theta}^{(t)} = \begin{cases} \boldsymbol{\theta}^* \text{ with probability } \min(\alpha, 1) \\ \boldsymbol{\theta}^{(t-1)} \text{otherwise.} \end{cases}$$

The Markov chain generated by the Metropolis-Hastings algorithm will eventually converge to the distribution used in the calculation of the acceptance criteria (which is the joint posterior distribution of model parameters, in this case) for any form of the proposal distribution given that the Markov chain is ergodic (Gilks et al. 1996).

A crucial aspect of MCMC simulation is the verification of convergence of the Markov chain to a stationary distribution. The present study, employs two widely used diagnostic convergence metrics, the Gelman and Rubin (1992) and Brooks and Gelman (1998) criteria. As a large number of iterations of the steps of the algorithm are usually required for convergence, the computational constraints imposed by complex models necessitate the use of a computationally efficient, yet accurate surrogate model of the process, for which we advocate the use of the SRSM.

2.2 The Stochastic Response Surface Method

The Stochastic Response Surface Method (SRSM, Isukapalli 1999) is an extension of the classical deterministic Response Surface Methods and the Deterministic Equivalent Modeling Method (Tatang 1995). The motivation underlying the development of the SRSM was to reduce the number of computationally demanding full model simulations required for adequate estimation of statistics of their output probability densities, as compared to conventional methods of uncertainty analysis (like Monte Carlo, Latin Hypercube sampling etc.). This is accomplished by approximating both the inputs and outputs of the uncertain system through series expansions of standard random variables; the series expansions of the outputs contain coefficients that can be calculated from the results of a limited number of full model simulations. Thus, successful application of the SRSM results in generating a computationally less demanding, statistically equivalent, reduced model of the chosen system outputs.

Evaluating an SRSM expansion consists of the following steps (Figure 4): (a) input uncertainties are expressed in terms of a set of standard random variables (*srv*s), (b) a functional form is assumed for selected outputs or output metrics, and (c) the parameters of the functional approximation are determined.

The *srv*s are selected from a set of independent, identically distributed (*iid*) normal random variables, $\{\xi_i\}_{i=1}^{n}$, where n is the number of independent inputs, and each ξ_i has zero mean and unit variance. When the input random variables are independent, the uncertainty in the i^{th} model input X_i, is expressed directly as a function of the i^{th} *srv*, ξ_i; i.e., a transformation of X_i to ξ_i is employed. Such transformations are useful in the standardized representation of the random inputs, each of which could have very different distribution properties. Table 1 presents

a list of transformations for some probability distributions commonly employed in transport-transformation modeling. In cases where the random inputs are correlated and the interdependence of the variables described by a correlation matrix, a transformation process can be applied to the inputs as outlined in Isukapalli (1999).

The next step involved in implementing SRSM is expressing the series expansion of normal random variables in terms of Hermite polynomials; the "polynomial chaos expansion" (Ghanem and Spanos 1991). When normal random variables are used as *srv*s, an output can be approximated by a polynomial chaos expansion on the set $\{\xi_i\}_{i=1}^{n}$, given by:

$$\begin{aligned} y &= a_o + \sum_{i_1=1}^{n} a_{i_1}\Gamma_1\left(\xi_{i_1}\right) + \sum_{i_1=1}^{n}\sum_{i_2=1}^{i_1} a_{i_1 i_2}\Gamma_2\left(\xi_{i_1},\xi_{i_2}\right) \\ &+ \sum_{i_1=1}^{n}\sum_{i_2=1}^{i_1}\sum_{i_3=1}^{i_2} a_{i_1 i_2 i_3}\Gamma_3\left(\xi_{i_1},\xi_{i_2},\xi_{i_3}\right) + \ldots, \end{aligned} \tag{2}$$

where y is an uncertain output of the model, the $a_{i_1}\ldots$s are deterministic constants to be evaluated, and the $\Gamma_p\left(\xi_{i_1},\ldots,\xi_{i_p}\right)$s are multi-dimensional Hermite polynomials of degree p, given by:

$$\Gamma_p\left(\xi_{i_1},\ldots,\xi_{i_p}\right) = (-1)^p e^{\frac{1}{2}\xi^T\xi}\frac{\partial^p}{\partial\xi_{i_1}\ldots\partial\xi_{i_p}}e^{-\frac{1}{2}\xi^T\xi}, \tag{3}$$

where $\boldsymbol{\xi}$ is the vector of *p iid* normal random variables $\{\xi_{i_k}\}_{k=1}^{p}$, that are used to represent input uncertainty.

It is known that the set of multi-dimensional Hermite polynomials form an orthogonal basis for the space of square-integrable probability distribution functions, and that the polynomial chaos expansion converges in the mean-square sense (Ghanem and Spanos 1991). In general, the accuracy of the approximation increases as the order of the polynomial chaos expansion increases and thus the order of the expansion can be selected to reflect accuracy needs and computational constraints.

For example, an uncertain model output y, can be expressed via first, second and third order SRSM polynomial approximations, y_1, y_2 and y_3 as follows:

$$y_1 = a_{0,1} + \sum_{i=1}^{n} a_{i,1}\xi_i \tag{4}$$

$$y_2 = a_{0,2} + \sum_{i=1}^{n} a_{i,2}\xi_i + \sum_{i=1}^{n} a_{ii,2}\left(\xi_i^2 - 1\right) + \sum_{i=1}^{n-1}\sum_{j>i}^{n} a_{ij,2}\xi_i\xi_j \tag{5}$$

$$y_3 = a_{0,3} + \sum_{i=1}^{n} a_{i,3}\xi_i + \sum_{i=1}^{n} a_{ii,3}\left(\xi_i^2 - 1\right) + \sum_{i=1}^{n} a_{iii,3}\left(\xi_i^3 - 3\xi_i\right)$$

$$+\sum_{i=1}^{n-1}\sum_{j>i}^{n} a_{ij,3}\xi_i\xi_j + \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ijj,3}\left(\xi_i\xi_i^2 - \xi_i\right) + \sum_{i=1}^{n-2}\sum_{j>i}^{n-1}\sum_{k>j}^{n} a_{ijk,3}\xi_i\xi_j\xi_k \qquad (6)$$

where n is the number of *srv*s used to represent the uncertainty in the model inputs, and $a_{i,m}$, $a_{ij,m}$, $a_{ijj,m}$, and $a_{ijk,m}$ are the coefficients to be estimated, where m represents the order of polynomial expansion.

The final step in the SRSM implementation is to determine these coefficients of the polynomial chaos expansion, which is accomplished through an extension of traditional collocation methods that is based on a combination of regression and an improved input collocation scheme called the Efficient Collocation Method (Isukapalli 1999). This improved sampling scheme selects points based on a modification of the standard orthogonal collocation method of Tatang (Tatang 1995, Villadsen and Michelsen 1978). The points are selected so that each standard normal random variable ξ_i takes the values of either zero or one of the roots of the higher order Hermite-polynomial. Borrowing from Gaussian quadrature, this scheme attempts to increase the order of behavior a polynomial (of fixed order) can capture. Details of the ECM and other aspects of SRSM can be found in Isukapalli (1999), Isukapalli et al. (2000), and Isukapalli and Georgopoulos (1999).

After the set of sample input points is generated using the ECM and suitable transformations and corresponding outputs obtained by running the model at these points, regression is employed to obtain robust estimates of the coefficients. The model outputs at the selected sample points are equated with the estimates from the series approximation, resulting in

a set of linear equations with more equations than unknowns. This system of equations is then solved using singular value decomposition.

The final results obtained upon successful application of the above methodology, are efficient and statistically equivalent SRSM polynomial expansions that accurately represent the full model outputs given a set of input distributions.

2.3 Methodology

The present approach to the problem of parametric uncertainty reduction and analysis of complex models, in summary, consists of (I) updating/reducing the uncertainty in the expert opinion based input parameter distributions (priors) given additional information, and (II) using these updated (posterior) input parameter distributions, which may equivalently be specified in the form of the joint distribution of these parameters, for the task of uncertainty analysis of the output quantities of interest.

For complex models, where it is desired to make minimal restrictions on the input parameters and the model process, itself, step (I) in the above procedure is done numerically via MCMC. As it is often unfeasible to apply MCMC techniques to convergence on the complex models directly, we suggest the use of SRSM expansions to approximate the outputs of the complex model in this step (i.e. to create a statistically equivalent surrogate model). The prior distributions can be specified based on expert opinion (informative) or can be vaguely defined/flat (non-informative) depending on the case specifics. SRSM expansions can handle both these types of distributions as input variables. SRSM expansions are then fit to the outputs of the complex model as shown in the previous subsection and these expansions used in the numerical Bayesian uncertainty updating.

In other words, what we propose for step (I) is:

1. An SRSM approximate model of the complex model outputs, of appropriate order, be fit following the exposition in the previous subsection and the development in equations 2-6, based on the prior distributions of the parameters. This results in functions of the form $y(\boldsymbol{\theta})$, that give direct input output relations (Note that the *srv*s ξ_i are related to corresponding θ_i's through the functions used to fit the θ_i's, e.g. via Table 1).
2. Numerical Bayesian inference then be carried out via MCMC based on the statistically equivalent SRSM approximate model constructed in the previous step. Note that likelihood values $P(\boldsymbol{d}|\boldsymbol{\theta})$, for the probability model under consideration, are now evaluated based on the SRSM polynomial outputs, $y(\boldsymbol{\theta})$.

Once the Bayesian inference step is complete and the joint posterior distribution of the input parameters has been obtained (i.e. the MCMC simulation deemed converged), this joint posterior distribution can then be used as an updated input to the statistically equivalent SRSM model (the same expansions utilized in the Bayesian inference step) and step (II) carried out to obtain the final uncertainty analysis of the outputs of the system.

The following sections introduce the case study, which is a steady state groundwater flow model of the U.S. Department of Energy Savannah River Site General Separations Area using the FACT code and highlights the various aspects of application of this methodology.

3 Case Study Description: The GSA Steady State Groundwater Flow as Simulated by the FACT Code

The General Separations Area (GSA) of the U.S. Department of Energy Savannah River Site model covers a 40 km^2 area bounded by Fourmile Branch on the south, Upper Three Runs on the north, F-area on the west, and McQueen Branch on the east (Flach and Harris 2000). The FACT numerical model extends from the ground surface to the bottom of the Gordon aquifer (Figure 1 and Figure 2), which is overlain by the Upper Three Runs aquifer.

Groundwater from the Upper Three Runs (UTR) aquifer unit is assumed to discharge equally from each side of Upper Three Runs, Fourmile Branch and McQueen Branch. Therefore, these streams provide natural, no-flow boundary conditions for most of the UTR aquifer unit. On the west side of the unit, hydraulic head values from a contour map of measured water elevations are prescribed. The Gordon aquifer is assumed to discharge equally from both sides of Upper Three Runs and so a no-flow boundary condition is specified over the north face of the model. Lacking natural boundary conditions, hydraulic heads are specified over the west, south and east faces of the model within the Gordon aquifer. Areas of groundwater recharge and discharge, consistent with computed hydraulic head at ground surface, are computed as part of the model solution using a combined recharge/drain boundary condition, applied over the entire top surface of the model. Groundwater discharges to surface water in regions where the computed head is above ground elevation. The Subsurface Flow And Contaminant Transport (FACT) code is a transient three-dimensional, finite element code designed to simulate isothermal groundwater flow, moisture movement, and solute transport in variably saturated and fully saturated subsurface porous media (Hamm and Aleman 2000). The code is designed specifically to handle complex multi-layer and/or heterogeneous aquifer systems in an efficient

manner and accommodates a wide range of boundary conditions. The code uses simple rectangular (plane or brick) elements that can be deformed in the vertical dimension to accommodate stratigraphic variations.

The groundwater flow equation is approximated using the Bubnov-Galerkin finite element method in conjunction with an efficient symmetric PCG (Preconditioned Conjugate Gradient, ICCG) matrix solver.

The areal resolution of the model is 200 ft^2 except in peripheral areas. There are 108 elements along the east-west axis, and 77 elements along the north-south axis. The vertical resolution varies depending on hydrogeologic unit and terrain/hydrostratigraphic surface variations. Each hydrostratigraphic surface is defined by numerous picks ranging in number from approximately 70 to 375 depending on the surface. The Upper Aquifer Zone (UAZ) of the UTR aquifer unit is represented with 9 finite-element layers in the vertical direction. The vadose zone is included in the model, although the physical processes therein are not modeled in detail. The Lower Aquifer Zone (LAZ) contains 5 finite-element layers while the Tan Clay confining Zone (TCZ) separating the aquifer zones is modeled with 2 vertical elements. The Gordon confining and aquifer units each contain 2 elements, for a total of 20 vertical elements from ground surface to the bottom of the Gordon aquifer. The 3D mesh size is therefore 108x77x20=166,320 elements or 109x78x21=178,542 nodes.

Hydraulic conductivity values in the model are based directly on a large characterization database comprised of approximately 85 pumping and 481 slug test data points, 258 laboratory permeability measurements, and nearly 37,500 lithology data records. The conductivity field is heterogeneous within hydrogeologic units and reflects variations present in the characterization

data. The initial conductivity values are further refined through model calibration to observed heads in wells.

Prior groundwater budget studies provide an estimate into the average natural recharge over the entire model domain. Various man-made features (e.g. basins) provide additional recharge in localized areas.

The estimated discharge rates to Upper Three Runs, Fourmile Branch, McQueen Branch, and Crouch Branch and predicted seepage faces within the model domain are consistent with field observations. Simulated hydraulic heads, vertically-averaged over the entire thickness of the upper UTR, lower UTR, and Gordon aquifer zones, agree well with potentiometric maps based on measured heads. Simulated flow directions vertically-averaged over the entire thickness of the aquifer zones further agree with conceptual models of groundwater flow.

4 Application of the Bayesian Framework to the GSA as Simulated by the FACT Code

4.1 Numerical Bayesian Uncertainty Updating

The three-dimensional conductivity fields and the two-dimensional recharge rate field are the dominant uncertain inputs to the saturated flow model for the analysis considered in this work. This is consistent with the earlier uncertainty analysis of the area (Balakrishnan et al., A Comparative Assessment of Efficient Uncertainty Techniques for Environmental Fate and Transport Models: Application to the FACT Model, submitted to *Journal of Hydrology,* 2003). Each zone is heterogeneous and five key uncertain variables for the GSA simulated by the FACT model were taken to be the vertical hydraulic conductivity field (K_v) values for the Gordon Confining Unit (GCU) and the Tan Clay Confining Zone (TCCZ), the horizontal hydraulic

conductivity field (K_h) values for the Lower UTR Aquifer Zone (LAZ) and the Upper UTR Aquifer Zone (UAZ) and the groundwater recharge rate (RECH).

As in the previous uncertainty analysis, the problem was simplified, by individually assigning a global multiplier to each of the three-dimensional conductivity field variables as well as one to the two-dimensional recharge rate field. The global multiplier ensures that relative spatial variations dictated by characterization and subsequent model calibration are preserved, while the mean value of the field is perturbed (Flach and Harris 2000). The main output variables monitored were the simulated hydraulic head values in the various aquifers (which are direct model outputs) and the stream baseflows in the main discharge regions (Fourmile, Crouch and McQueen Branches and the Upper Three Runs) which are model post processed results. The uncertainty updating task for the GSA as simulated by the FACT code essentially consisted of fusing (via numerical Bayesian inference) the information contained in the measured values of hydraulic head data with the model simulated values and the prior uncertain distributions, in order to obtain updated model and data consistent input parameter distributions. Hydraulic head field measurements were available for a total of 667 well locations which were divided into three groups according to their corresponding aquifers. The three groups have 79, 173 and 415 wells located in the upper UTR, lower UTR and the Gordon Aquifers respectively (a total of 667 wells). The vector of these measured hydraulic head values $d_{k,l}$, corresponds to the observed data $\boldsymbol{d}$, as specified in our development. In this and further expressions, the subscript k will be used to index the set of observed well locations ($k \in \{1,2,\ldots,667\}$). The subscript l, is uniquely determined by k, and is used simply to specify which of the three aquifers the k^{th} well is located in ($l \in \{1,2,3\}$).

The expert opinion based prior distributions derived from measurements and traditional model calibration for global multiplier variables were the same as those used in the previous uncertainty analysis study, namely independent random variables with the distributions shown in Table 2 (Note: the Log_{10}normal(μ,σ) distribution refers to a random variable whose Log (base 10) of the distribution results in a Normal distribution $N(\mu,\sigma)$).

These priors were used as the input distributions for the application of the SRSM. Collocation points were generated based on these input distributions, the full model was run at these specified points and results were collated in order to obtain the coefficients of the SRSM polynomials for each of the outputs under consideration (all 667 well hydraulic head values and stream baseflow rates). Both second and third order SRSM expansions were fit to the full model outputs requiring only 51 and 191 simulations, respectively. The convergence of the resulting output distributions obtained by SRSM polynomials of second and third order revealed that second order SRSM polynomials were sufficient for the representation of this system. Note that we employed the heuristic method advised in Isukapalli (1999) for this decision, i.e. that the order of SRSM expansion was determined as the lower of two successive orders (second and third order in our case) of SRSM expansions whose output probability distribution functions, as estimated by the SRSM polynomials, were deemed visually convergent. Figure 5 shows the resultant second and third order SRSM polynomial based probability distributions obtained at a particular well location (for its hydraulic head value) and a stream baseflow (for the discharge rate). In order for the reader to gauge the accuracy of the expansions, also shown on the plots are the probability distributions estimated via a Monte Carlo sample (of 1000 realizations) from the prior distribution. Thus, a second order SRSM approximate model for the GSA as simulated by FACT was obtained, i.e. expressions of the form $y_{2,k}(\boldsymbol{\theta})$ (based on equation 5). MCMC

simulation using the Metropolis-Hastings algorithm was subsequently implemented on the SRSM reduced model as follows; the probability model utilized was:

$$d_{k,l}(\boldsymbol{\theta}) = y_{2,k}(\boldsymbol{\theta}) + \varepsilon_l \tag{7}$$

i.e., the observed well hydraulic head value $d_{k,l}(\boldsymbol{\theta})$ is modeled as the sum of $y_{2,k}(\boldsymbol{\theta})$, the SRSM approximated hydraulic head value and $\varepsilon_l \sim N(0, \sigma_l)$, a normally distributed random error with zero mean and variance σ_l. Three independent error models were utilized for each of the three output groups corresponding to each aquifer ($l \in \{1,2,3\}$), as we expected the magnitudes of random error to be similar only within each aquifer. This often used Gaussian probability model lends itself to easy interpretation, in that we assumed that the outputs the SRSM model provided were consistent and unbiased at any set of inputs, but that each actual observation itself was likely to be correct only to an approximation due to independent normally distributed random errors (which can be interpreted as due to effects such as measurement error, noise etc.).

The priors for the uncertain variables (global multiplier values) were chosen as described previously, with each aquifer error model variance σ_l, set as a parameter to be estimated by the Bayesian inference process (non-informative/flat priors were used for these quantities). Thus, the final vector of parameters for Bayesian inference $\boldsymbol{\theta}$, consisted of the five uncertain global multiplier variables and three unknown probability model variances σ_l. The proposal distributions $J_t(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t-1)})$ were each chosen to be Gaussian with the evolving MCMC parameter realization values as their mean values and variance set at 20% variance of their corresponding prior (enabling the direct application of the symmetric proposal distribution acceptance probability given in equation 1). The Gaussian probability model in equation 7, then simply determines the likelihood value $P(\mathbf{d}|\boldsymbol{\theta})$, via the expression:

$$P(\boldsymbol{d} \mid \boldsymbol{\theta}) = \prod_{k=1}^{667} \varphi\left(d_{k,l} \mid y_{2,k}(\boldsymbol{\theta}), \sigma_l\right)$$

where φ is the Gaussian density function (which can be easily evaluated given any realization of the parameters, $\boldsymbol{\theta}$, and the measured hydraulic head values, $\boldsymbol{d}$).

Sets of highly overdispersed values were chosen as initial points of the MCMC simulation and 4 chains were successfully run for 10,000 iterations. After confirming convergence, the final joint posterior distribution of the input parameters $P(\boldsymbol{\theta}|\boldsymbol{d})$, was obtained after discarding the initial burn-in (the first half of the converged chain) for the chain with the least variance in the most variables.

4.2 Final Uncertainty Analysis

Subsequent to obtaining the converged joint posterior distribution of the input parameters, two further stages of analysis were carried out in order to obtain the final output variable distributions (the stream baseflow rates and the well location hydraulic head values).

First, in order to validate the accuracy of the reduced model for the purposes of the final uncertainty analysis, the full model (GSA as simulated by the FACT code) was run on the largest practically feasible subset of the converged joint posterior distribution of the inputs (which consists of 5,000 realizations of the MCMC simulation). Current computational limitations placed this constraint to about 1,000 full model simulations, and hence, the joint posterior realizations of the chosen chain (5,000 realizations) were sampled at every fifth location for a final sample size of 1,000 realizations.

The second stage of analysis was then carried out wherein the output distributions for the converged chain were computed at every location of the chosen converged chain (at each of the

5,000 realizations using the SRSM second order approximation to the full model). Note that this allowed for direct comparison of the accuracy of the SRSM reduced model of the GSA as simulated by FACT, as both the full model and reduced model output values were available at the same input parameter values.

Upon successful validation of accuracy of the reduced model, the complete joint posterior distribution of the inputs obtained via the MCMC simulation was fully utilized for the uncertainty analysis task by the computation of SRSM approximated values of the ouputs in order to obtain the final output distributions of the system variables.

5 Results and Discussion

The four MCMC simulations performed with the SRSM second order reduced model of the system, converged to practically identical joint posterior distributions. Convergence characteristics of the chains were confirmed via two well-known MCMC multiple chain convergence criteria, Gelman and Rubin (1992) and Brooks and Gelman (1998). The maximum value of the Gelman and Rubin's $\sqrt{R}$ metric was found to be 1.007 over all the parameters. The maximum Potential Shrink Rate Factor (PRSF) value (used by the Brooks and Gelman criteria) over all the parameters was found to be approximately 1.01 (in both cases, values close to 1 are deemed convergent).

Subsequent analysis of the joint posterior distribution was based on the final 5,000 realizations (discarding the initial half of the chain as burn-in) of the converged chain whose variance in the most input parameters was found to be minimal, namely chain 4. It serves to note though, that the choice of the chain has an insignificant effect on the remaining analysis and

results, based on the convergence of all the chains to a practically identical joint distribution of parameters.

The uncertainty in the posteriors, when compared to the priors, as evaluated by Shannon's $E\left[In\left(p\right)\right]$ metric (Shannon and Weaver 1964), was found to decrease for all the variables, as indicated by the positive value of information gain, which is defined as $IG = -E\left[In\left(p^0\right)\right] - \left(-E\left[In\left(p*\right)\right]\right)$ where p^0 and $p*$ are the prior and posterior probability density functions, obtained in the case of each uncertain variable (Figure 6). The computed values of *IG* were 2.18, 1.62, 3.37, 0.8 and 6.93 for the GCU K_v, LAZ K_h, TCZ K_v, UAZ K_h and recharge rate global multiplier distributions, respectively. Thus an uncertainty reduction has indeed been achieved by the Bayesian "fusion" of the field observations and the model predictions as approximated by the SRSM expansions.

This fact was further corroborated by running the full model with the global multiplier values set at the mean posterior parameter values and comparing the results of such a simulation to those obtained with the global multiplier values set at the mean prior parameter values (which was the nominal point of the expert opinion based analysis before the numerical Bayesian inference procedure). Results showed that the Root Mean Square (RMS) error for the hydraulic head values for all three groups had indeed become lower upon utilization of the selected posterior based parameter values (an overall drop from 4.65 to 4.56 for all the 667 well head locations). Corresponding reductions were also observed in the maximum deviations within the group and average deviations from the target observed well location hydraulic head values.

Additionally, information about the possible correlations between the global multiplier model input parameters was also estimated from the posterior parameter joint distribution (based

on the responses monitored, namely, the well location hydraulic head values). Figure 7 shows the pairwise scatter plot matrix showing the joint posterior input distrubtion correlations obtained for chain 4.

Finally, the uncertainty analysis of the outputs of the system was carried out in the two stage manner as stated in the previous subsection. For the 1,000 realizations subset of the final converged joint input posterior distribution, the full model and SRSM reduced model predict very similar output values for the quantities of interest, examples of which can be seen in Figure 8 and Figure 9. Plots such as these and the low root mean squared error for the output quantities (representative values over the 1,000 realizations for the hydraulic head values of wells BGO 8A, BG 93 and BG 26 in groups 1, 2 and 3 are 9.8×10^{-6}, 2.4×10^{-4} and 2.2×10^{-3} and those for the stream baseflows in the Fourmile Branch and Upper Three Runs are 5.5×10^{-4} and 6.38×10^{-4}) highlight the accuracy of the reduced model on the posterior input parameter realizations and permit the computation of the output distributions based on the complete joint posterior distribution as approximated by the second order SRSM reduced model of the system. A set of final output distributions, are shown in Figure 10, showing representative distributions of the computed hydraulic head values for a well from each aquifer/group and the stream baseflow rates for the Fourmile Branch and the Upper Three Runs.

Thus, the application of an efficient Bayesian framework of analysis has been demonstrated; this framework can aid practitioners in dealing with the task of the uncertainty analysis of complex computer codes in the presence of additional observational information. The framework was successfully applied to the steady state groundwater flow of the U.S. Department of Energy Savannah River Site GSA as simulated by the FACT model. Uncertainty in the posterior distributions generated by the numerical Bayesian inference procedure applied to

the statistically equivalent (SRSM) reduced model of the system was found to decrease and the final uncertainty analysis of the system was successfully completed. The work lends to easy extension in the form of various other algorithms for the MCMC procedure itself and for the future incorporation of model/structural uncertainty into this Bayesian framework.

6 Acknowledgements

We are indebted to the personnel at the Savannah River Technology Center and SRS who not only provided valuable data and the model but also deep insight into key areas of the problem and its formulation. The authors would also like to express their gratitude to Linda Everett for aid in preparation of figures and the general typesetting of this manuscript. This work has been funded in part by the U.S. Environmental Protection Agency under Cooperative Agreement # EPAR-827033 to the Environmental and Occupational Health Sciences Institute; and by a grant to the Institute for Responsible Management, Consortium for Risk Evaluation with Stakeholder Participation from the U.S. Department of Energy, Instrument DE-FG2600NT 40938. The viewpoints expressed in this work are solely the responsibility of the authors and do not necessarily reflect the views of the U.S. Department of Energy, the US Environmental Protection Agency, or their contractors.

7 References

Brand, K. P., and M. J. Small. Updating Uncertainty in an Intergrated Risk Assessment: Conceptual Framework and Methods. *Risk Analysis* **15**:719-731. 1995.

Brooks, S., and A. Gelman. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics* **7**:434-455. 1998.

Christakos, G., D. T. Hristopulos, and M. L. Serre. BME studies of stochastic differential equation representing physical laws - Part I. *in* Proceedings of the 5th Annual Conference of the International Association for Mathematical Geology, Trodheim, Norway. 1999.

Christakos, G., and X. Li. Bayesian maximum entropy analysis and mapping: A farewell to kriging estimators? *Mathematical Geology* **30**:435-462. 1998.

Currin, C., T. Mitchell, D. Morris, and D. Ylvisaker. Bayesian Prediction of Deterministic Functions, With Applications to the Design and Analysis of Computer Experiments. *Journal of the American Statistical Association* **86**:953-963. 1991.

Draper, D. Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society Series B* **57**:45-97. 1995.

Flach, G. P., and M. K. Harris. Integrated Hydrogeological Modeling of the General Separations Area. WSRC-TR-96-00399, Savannah River Technology Center, Aiken, South Carolina. 2000.

Gelman, A., and D. Rubin. Inference from Iterative Simulation using Multiple Sequences. *Statistical Science* **7**:457-511. 1992.

Ghanem, R. G., and P. D. Spanos. Stochastic Finite Elements: A Spectral Approach. Springer-Verlag, New York. 1991.

Gilks, W., S. Richardson, and D. Spiegelhalter. Markov Chain Monte Carlo in Practice. Chapman and Hall, London. 1996.

Hamm, L. L., and S. E. Aleman. FACT Subsurface Flow and Contaminant Transport Documentation and User's Guide. WSRC-TR-99-00282, Savannah River Technology Center, Aiken, South Carolina. 2000.

Hastings, W. K. Monte Carlo sampling methods using Markov Chains and their applications. *Biometrika* **57**:97-109. 1970.

Hill, M. C. Methods and guidelines for effective model calibration. Report 98-4005, U.S. Geological Survey Water Resources Investigations, Reston, Virginia. 1998.

Isukapalli, S. S. Uncertainty Analysis of Transport-Transformation Models. Ph.D. Thesis. Rutgers University, Piscataway, NJ. 1999.

Isukapalli, S. S., and P. G. Georgopoulos. Computational methods for efficient sensitivity and uncertainty analysis of models for environmental and biological systems. CCL/EDMAS-03, Piscataway, New Jersey. 1999.

Isukapalli, S. S., A. Roy, and P. G. Georgopoulos. Efficient sensitivity/uncertainty analysis using the combined Stochastic Response Surface Method and Automated Differentiation: Application to environmental and biological systems. *Risk Analysis* **20**:591-602. 2000.

Kennedy, M., and A. O'Hagan. Bayesian Calibration of Complex Computer Models. *Journal of the Royal Statistical Society Series B* **63**:425-464. 2001.

Kitanidis, P. K. Parameter Uncertainty in Estimation of Spatial Functions: Bayesian Analysis. *Water Resources Research* **22**:499-507. 1986.

Kitanidis, P. K. Recent advances in geostatistical inference on hydrogeological variables. Pages 1103-1109 in *U.S. National Report to International Union of Geodesy and Geophysics 1991-1994, Reviews of Geophysics, Supplement*. American Geophysical Union 1995.

Kitanidis, P. K. Comment on "A reassessment of the groundwater inverse problem" by D. McLaughlin and L. R. Townley. *Water Resources Research* **33**:2199-2202. 1997.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *Journal of Chemical Physics* **21**:1087-1091. 1953.

O'Hagan, A., M. Kennedy, and J. Oakley. Uncertainty Analysis and other Inference Tools for Complex Computer Codes. in *Bayesian Statistics 6*. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors. Oxford University Press, Oxford, England. 1999.

Porter, D. W., B. P. Gibbs, W. F. Jones, P. S. Huyakorn, L. L. Hamm, and G. P. Flach. Data Fusion Methods for groundwater systems. *Journal of Contaminant Hydrology* **42**:303-335. 2000.

Shannon, C. E., and W. Weaver. The Mathematical Theory of Communication. The University of Illinois Press, Urbana, IL. 1964.

Tatang, M. A. Direct Incorporation of Uncertainty in Chemical and Environmental Engineering Systems. Ph.D. Thesis. Massachusetts Institute of Technology, Cambridge, MA. 1995.

Villadsen, J., and M. L. Michelsen. Solution of Differential Equation Models by Polynomial Approximation. Prentice-Hall, Englewood Cliffs, New Jersey. 1978.

Woodbury, A. D., and T. J. Ulrych. A full-Bayesian approach to the groundwater inverse problem for steady state flow. *Water Resources Research* **36**:2081-2093. 2000.

Figure Captions

Figure 1. The simplified conceptual model of the General Separations Area showing the various aquifer units, recharge/discharge areas and mechanisms (figure adapted from Flach and Harris 2000)

Figure 2. An isometric view of the General Separations Area model highlighting the model finite element mesh

Figure 3. Flowchart outlining the typical steps involved in MCMC simulation

Figure 4. Flowchart outlining the steps involved in the SRSM Procedure

Figure 5. The convergence of the distributions for establishing the adequate order of SRSM expansion. Shown are representative probability distribution functions for (a) a well from the LAZ aquifer (Well BGO 8A) and (b) a stream baseflow (the Fourmile Branch). Each plot shows the Monte Carlo (MC) and SRSM approximated distributions (both second and third order), which are visibly convergent

Figure 6. Comparison of the expert opinion based prior distributions and the post-MCMC analysis based updated posterior marginal distributions for the uncertain variables in the system

Figure 7. The pairwise scatter plot matrix of chain 4 highlighting the joint posterior model input distribution correlations. The diagonal shows the marginal distributions of the variables

Figure 8. Comparison of the full model (GSA as simulated by the FACT code) predictions vs. the SRSM reduced model predictions of the hydraulic head values of Well BG 26 (in the Gordon Aquifer) at the 1000 (every fifth sampled) input points of the joint posterior distribution

Figure 9. Comparison of the full model (GSA as simulated by the FACT code) predictions vs. the SRSM reduced model predictions of the stream baseflow discharge rate of the Fourmile Branch at the 1000 (every fifth sampled) input points of the joint posterior distribution

Figure10. Final marginal probability distributions obtained after the uncertainty analysis based on the SRSM approximated outputs with the joint posterior distribution (obtained after the MCMC simulation) as input. Shown are the marginal distributions for a representative well from each aquifer (top row left to right) and the stream baseflow rates in the Fourmile Branch and the Upper Three Runs

Tables

Table 1. Representation of common univariate distributions as functionals of normal random variables

Distribution Type	**Transformation**
Uniform (*a, b*)	$a+(b-a)\left(\frac{1}{2}+\frac{1}{2}(\xi/\sqrt{2})\right)$
Normal (μ, σ)	$\mu+\sigma\xi$
Lognormal (μ, σ)	$\exp(\mu+\sigma\xi)$
Gamma (a, b)	$ab\left(\xi\sqrt{\frac{1}{9a}}+1-\frac{1}{9a}\right)^3$
Exponential (λ)	$-\frac{1}{\lambda}\log\left(\frac{1}{2}+\frac{1}{2}(\xi/\sqrt{2})\right)$
Weibull (a)	$y^{\frac{1}{a}}$
Extreme Value	$-\log(z)$

Table 2. Distributions of the uncertain inputs (global multipliers for the conductivity fields and recharge rate field) for the uncertainty analysis study

Variable	**Units**	**Distribution Type**	**Parameters (μ, σ)**
GCU K_v	ft/d	Log_{10}normal	-5, 0.75
LAZ K_h	ft/d	Log_{10}normal	0.9685, 0.15
TCCZ K_v	ft/d	Log_{10}normal	-2.2, 0.5
UAZ K_h	ft/d	Log_{10}normal	0.9823, 0.15
RECH	in/yr	Normal	18, 3.1

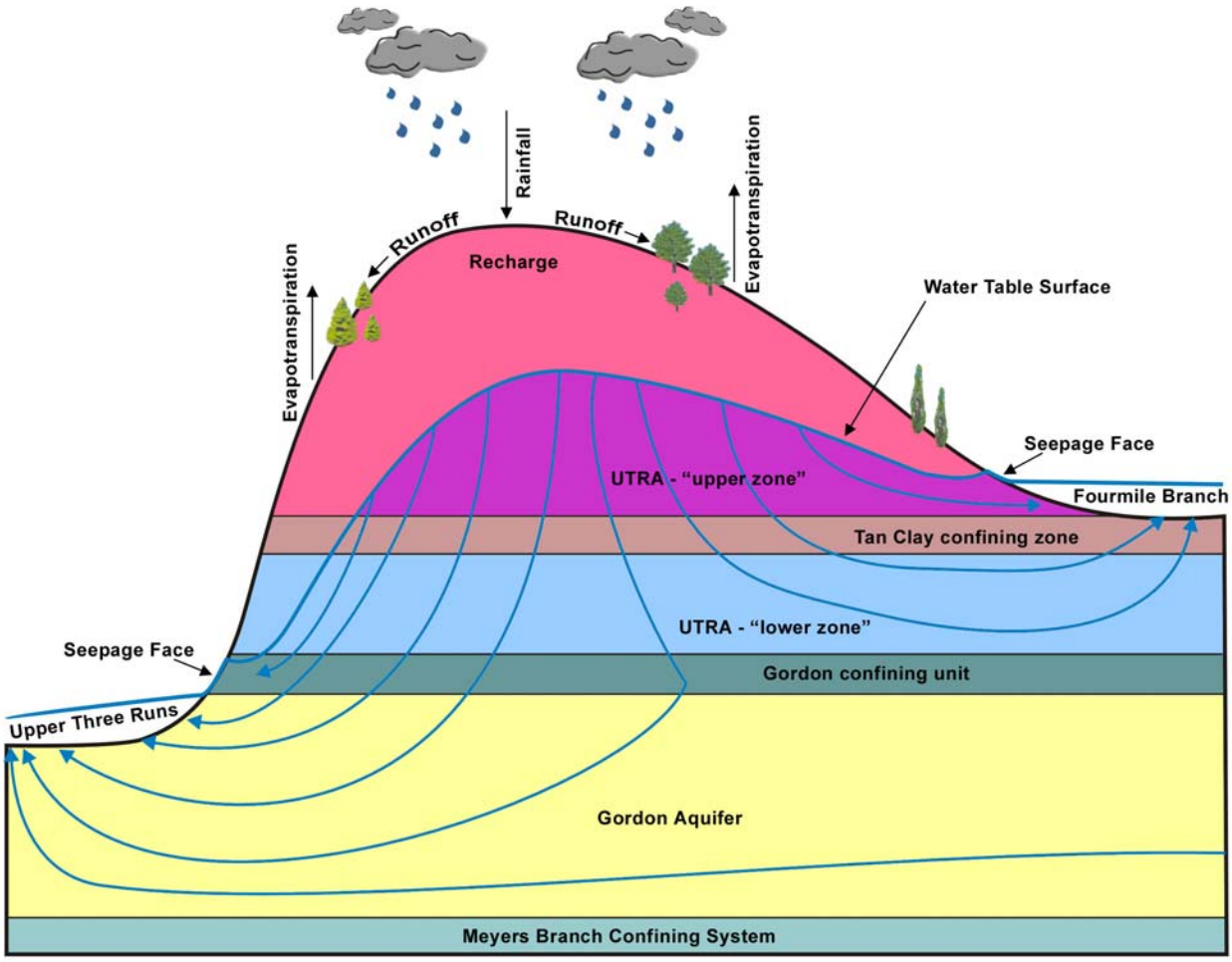

Figure 1. The simplified conceptual model of the General Separations Area showing the various aquifer units, recharge/discharge areas and mechanisms (figure adapted from Flach and Harris 2000)

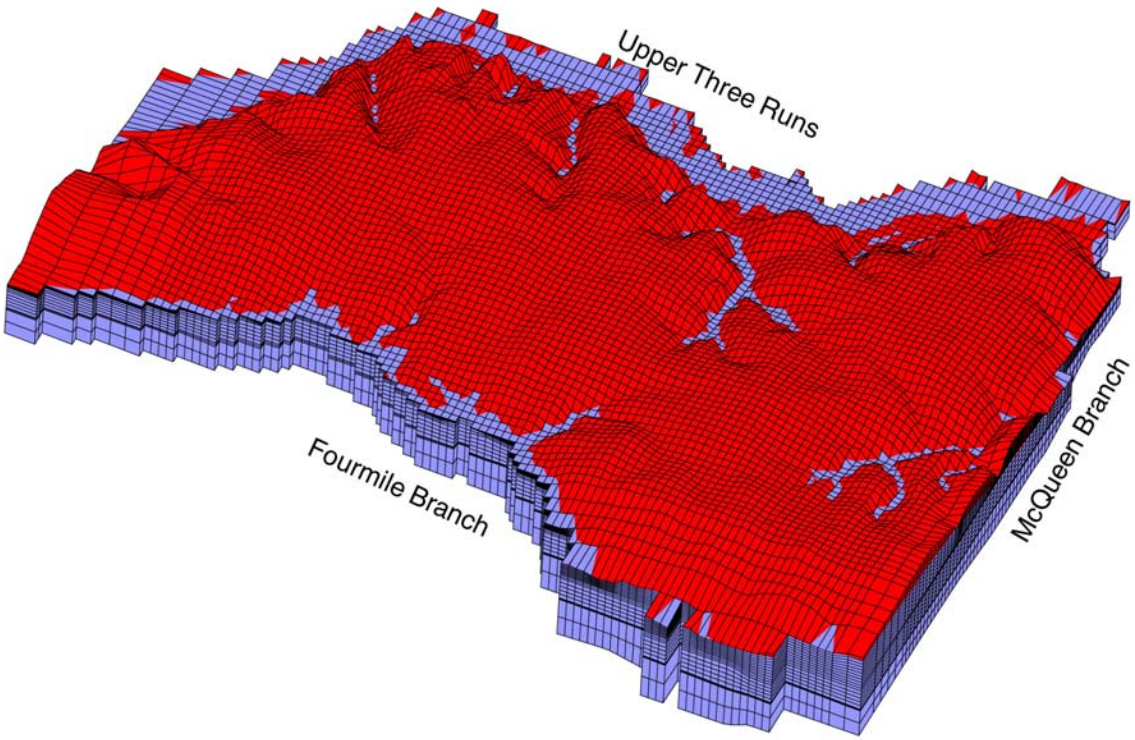

Figure 2. An isometric view of the General Separations Area model highlighting the model finite element mesh

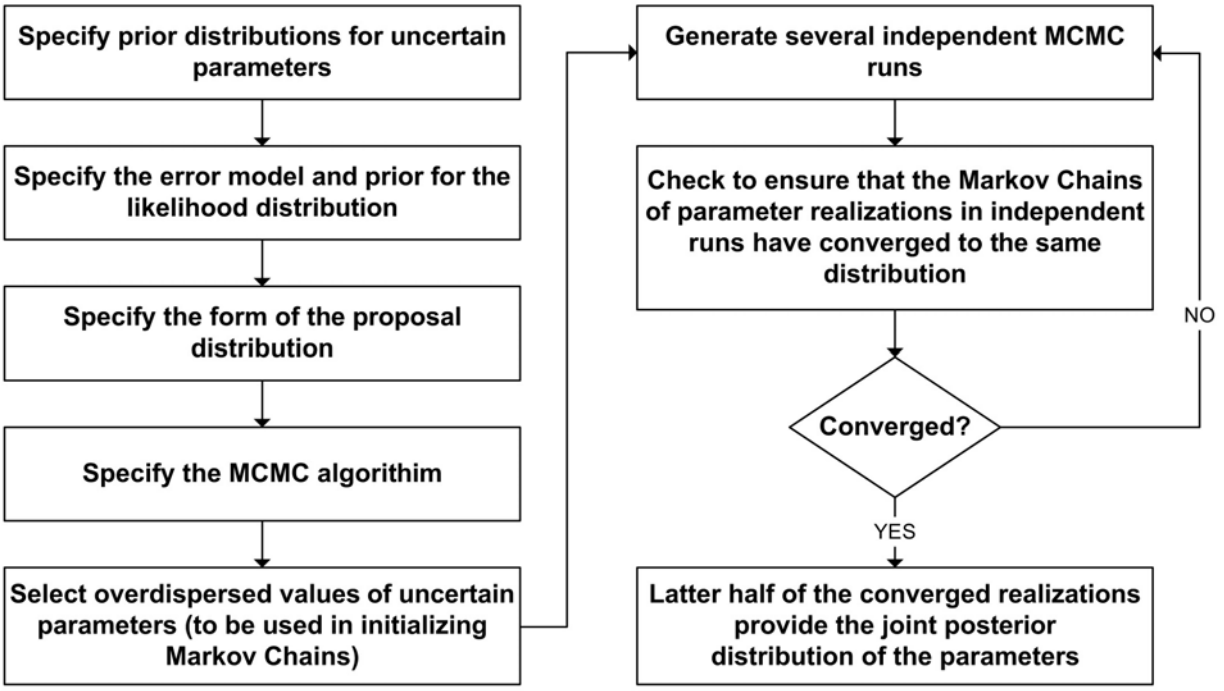

Figure 3. Flowchart outlining the typical steps involved in MCMC simulation

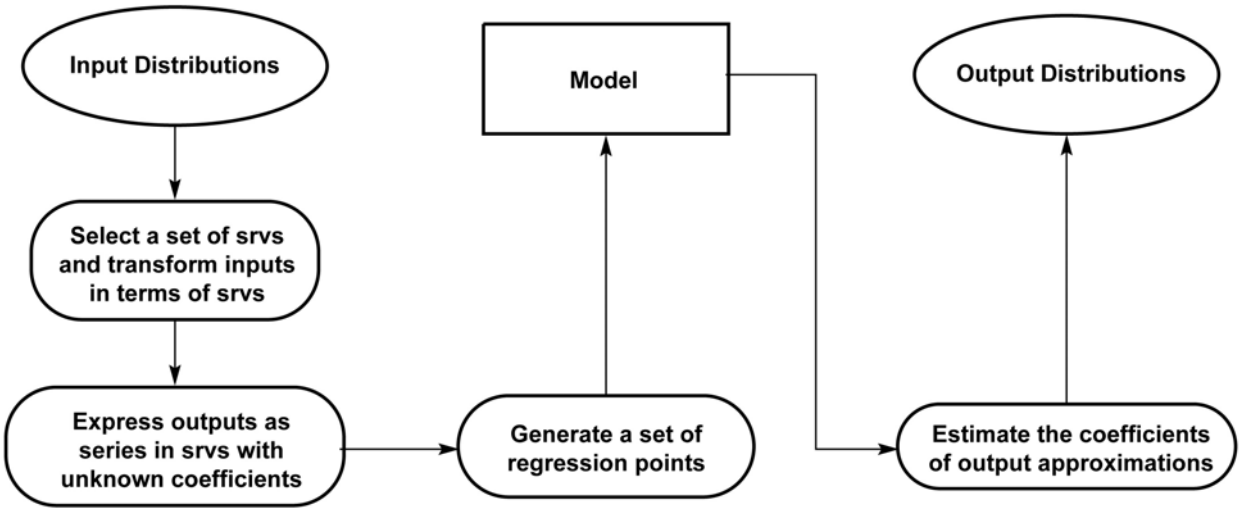

Figure 4. Flowchart outlining the steps involved in the SRSM Procedure

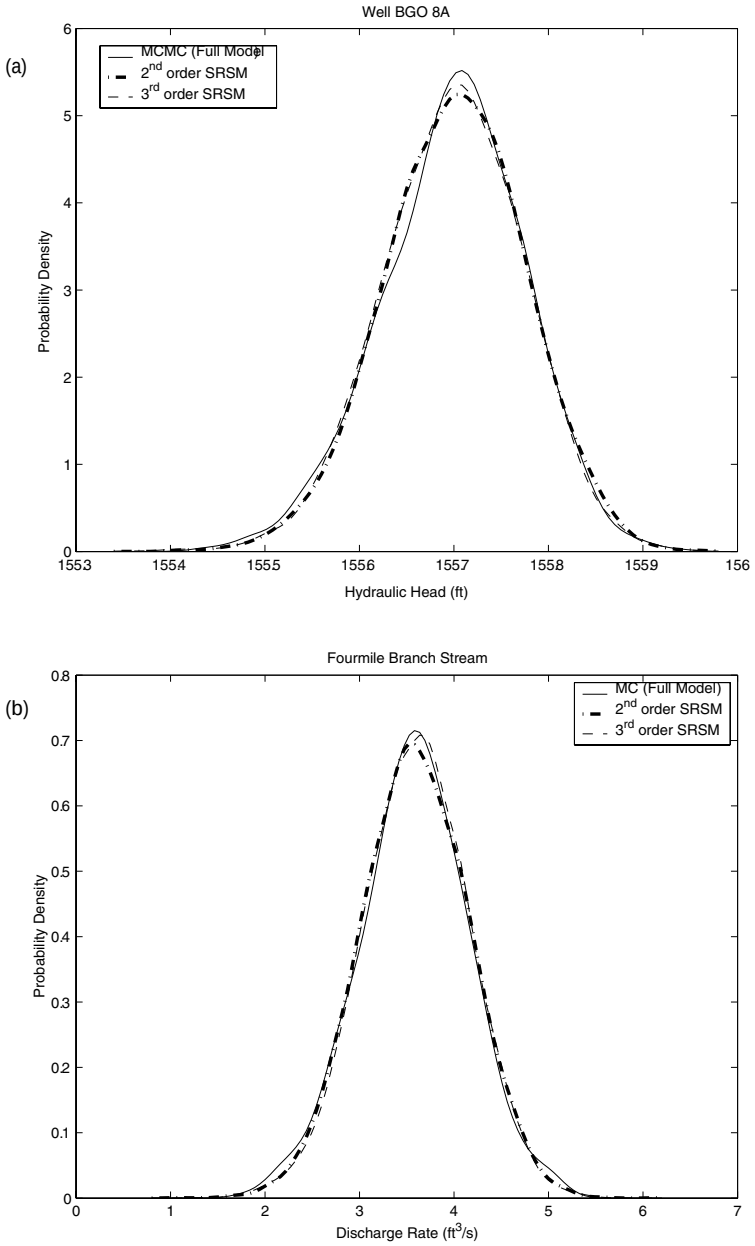

Figure 5. The convergence of the distributions for establishing the adequate order of SRSM expansion. Shown are representative probability distribution functions for (a) a well from the LAZ aquifer (Well BGO 8A) and (b) a stream baseflow (the Fourmile Branch). Each plot shows the Monte Carlo (MC) and SRSM approximated distributions (both second and third order), which are visibly convergent

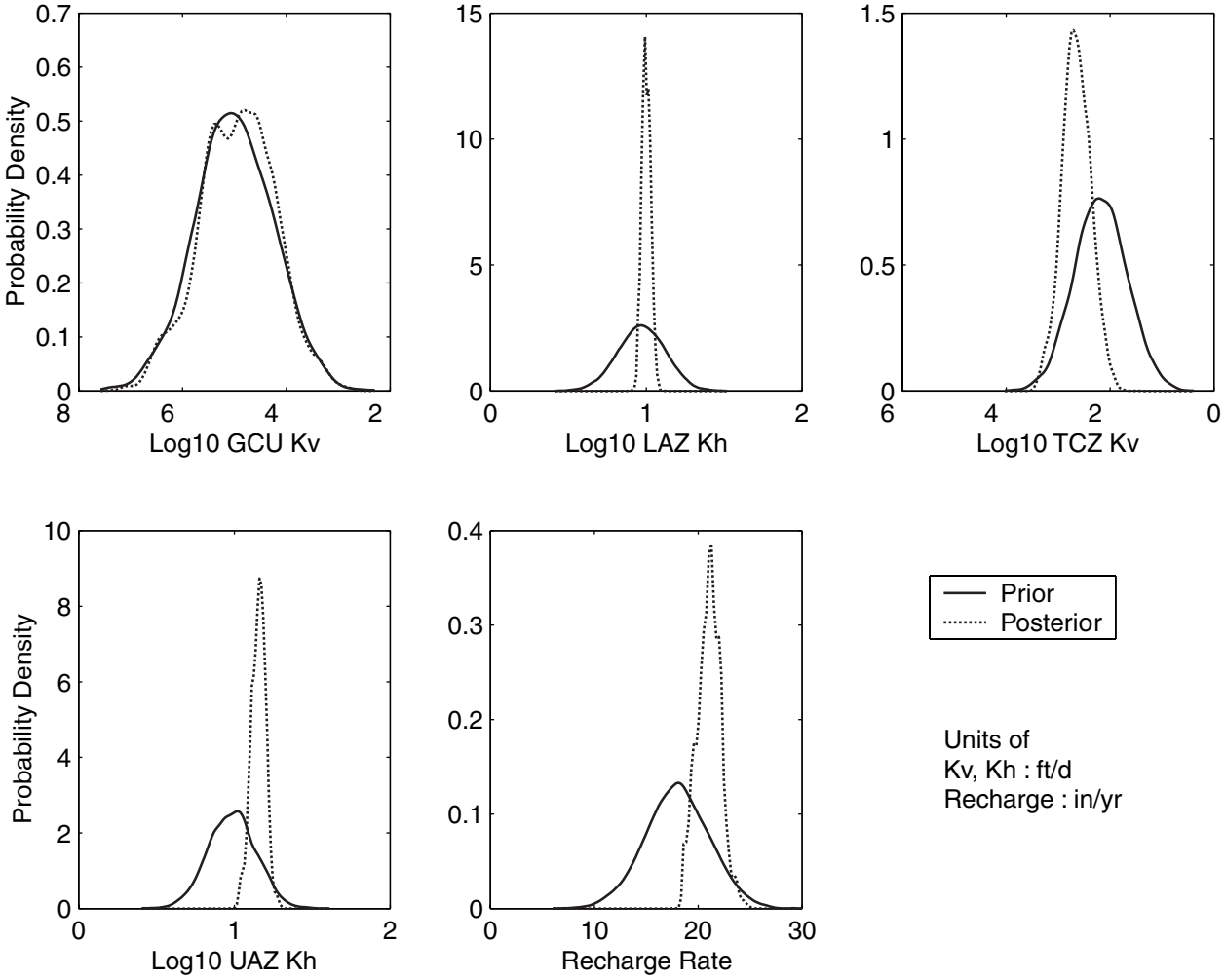

Figure 6. Comparison of the expert opinion based prior distributions and the post-MCMC analysis based updated posterior marginal distributions for the uncertain variables in the system

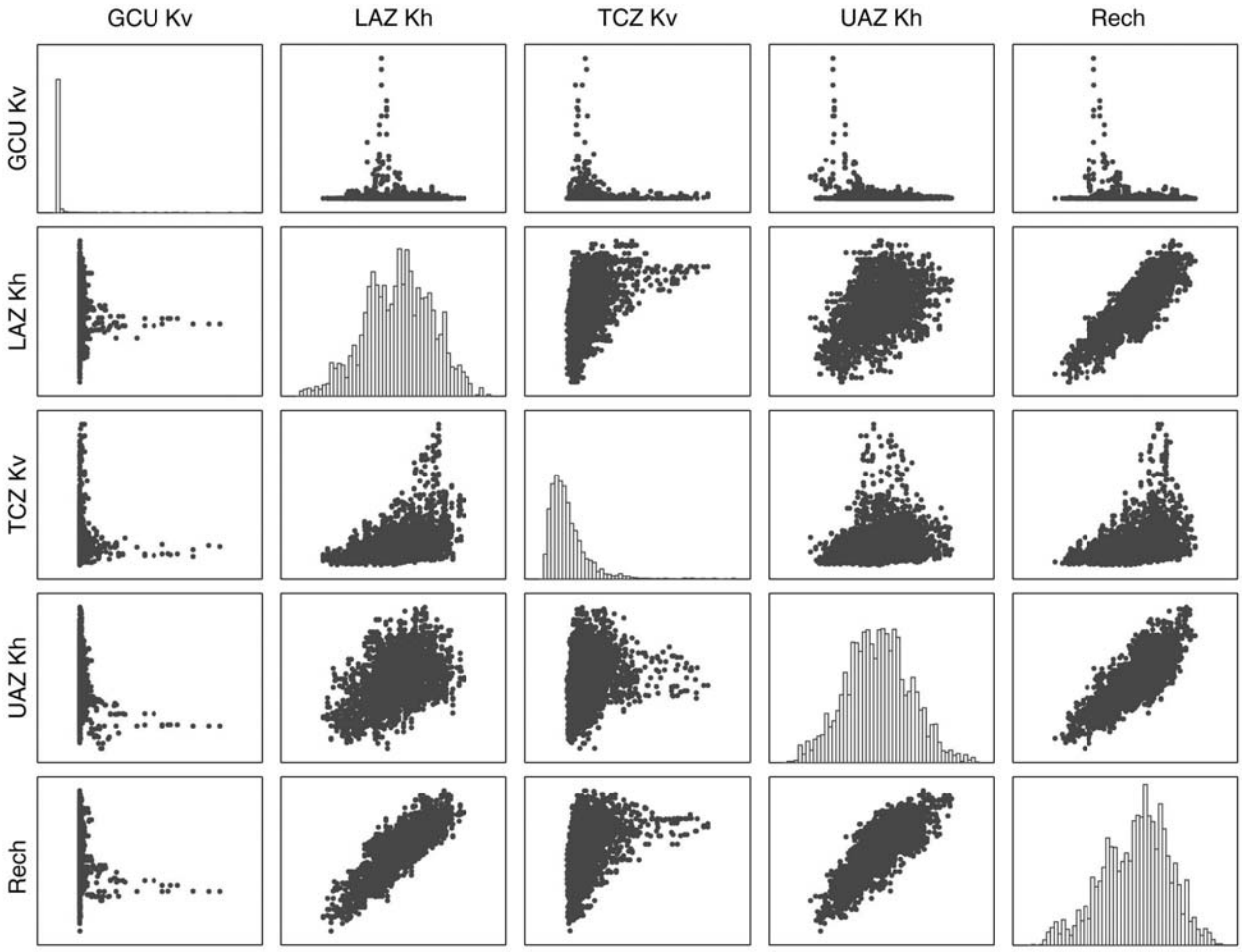

Figure 7. The pairwise scatter plot matrix of chain 4 highlighting the joint posterior model input distribution correlations. The diagonal shows the marginal distributions of the variables

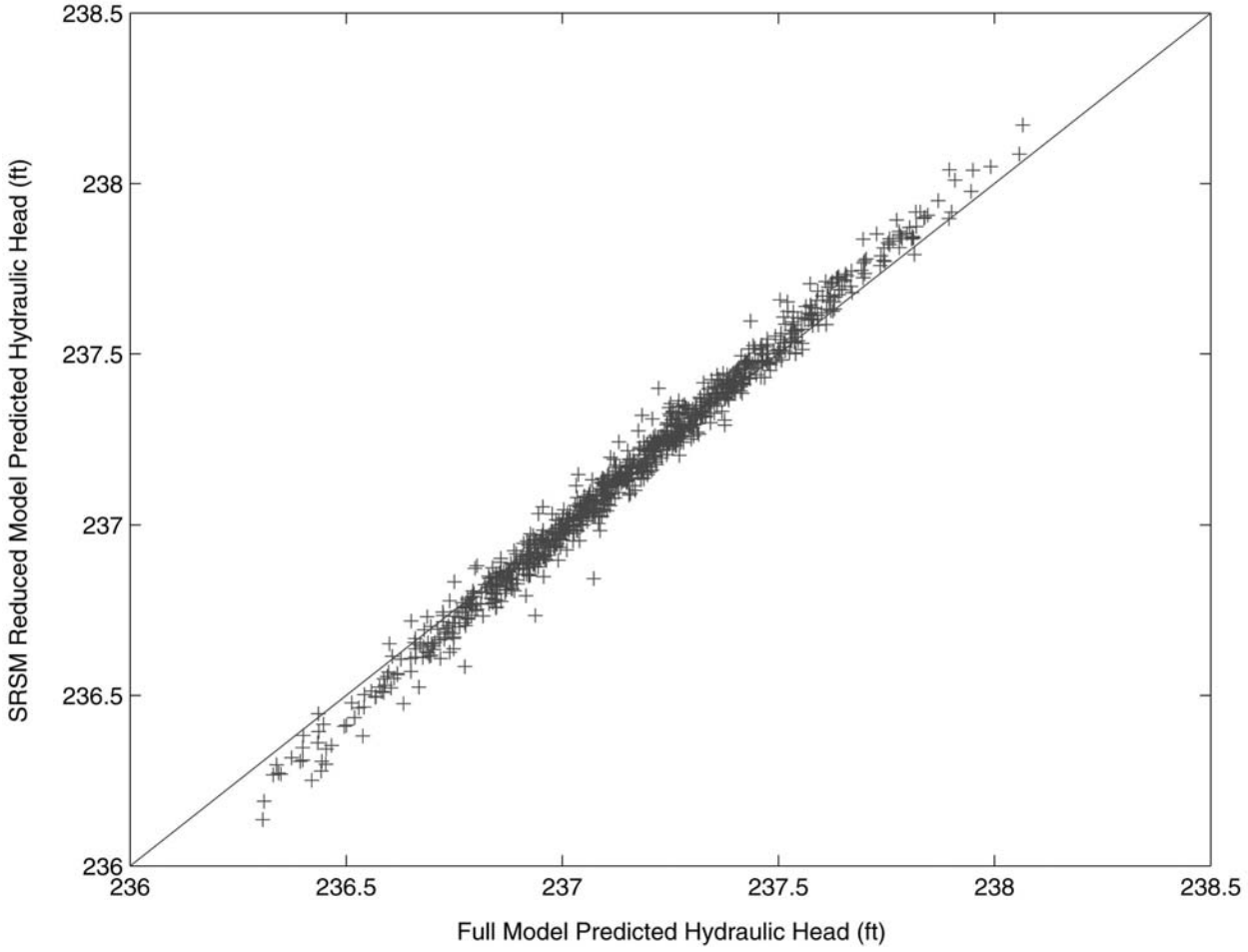

Figure 8. Comparison of the full model (GSA as simulated by the FACT code) predictions vs. the SRSM reduced model predictions of the hydraulic head values of Well BG 26 (in the Gordon Aquifer) at the 1000 (every fifth sampled) input points of the joint posterior distribution

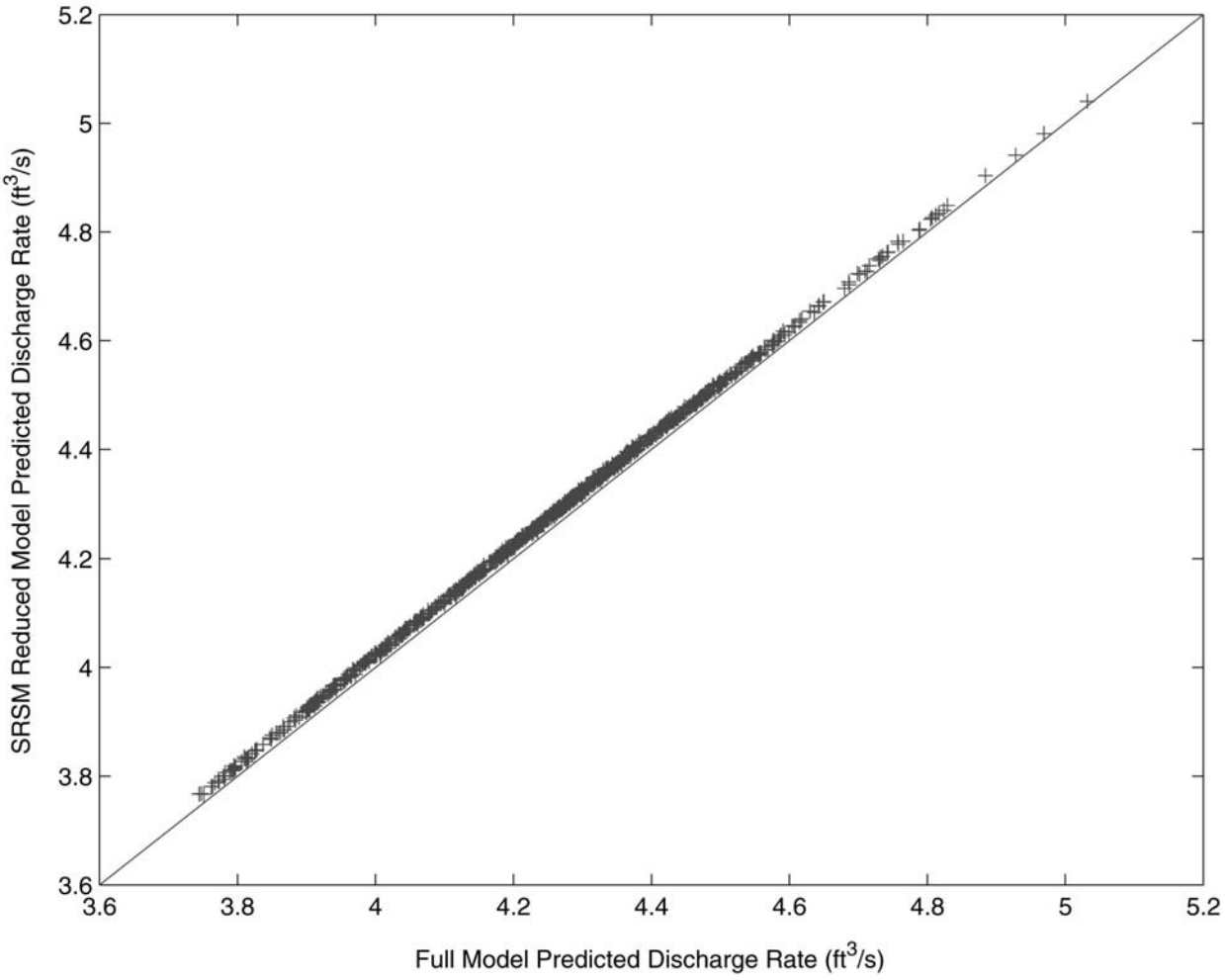

Figure 9. Comparison of the full model (GSA as simulated by the FACT code) predictions vs. the SRSM reduced model predictions of the stream baseflow discharge rate of the Fourmile Branch at the 1000 (every fifth sampled) input points of the joint posterior distribution

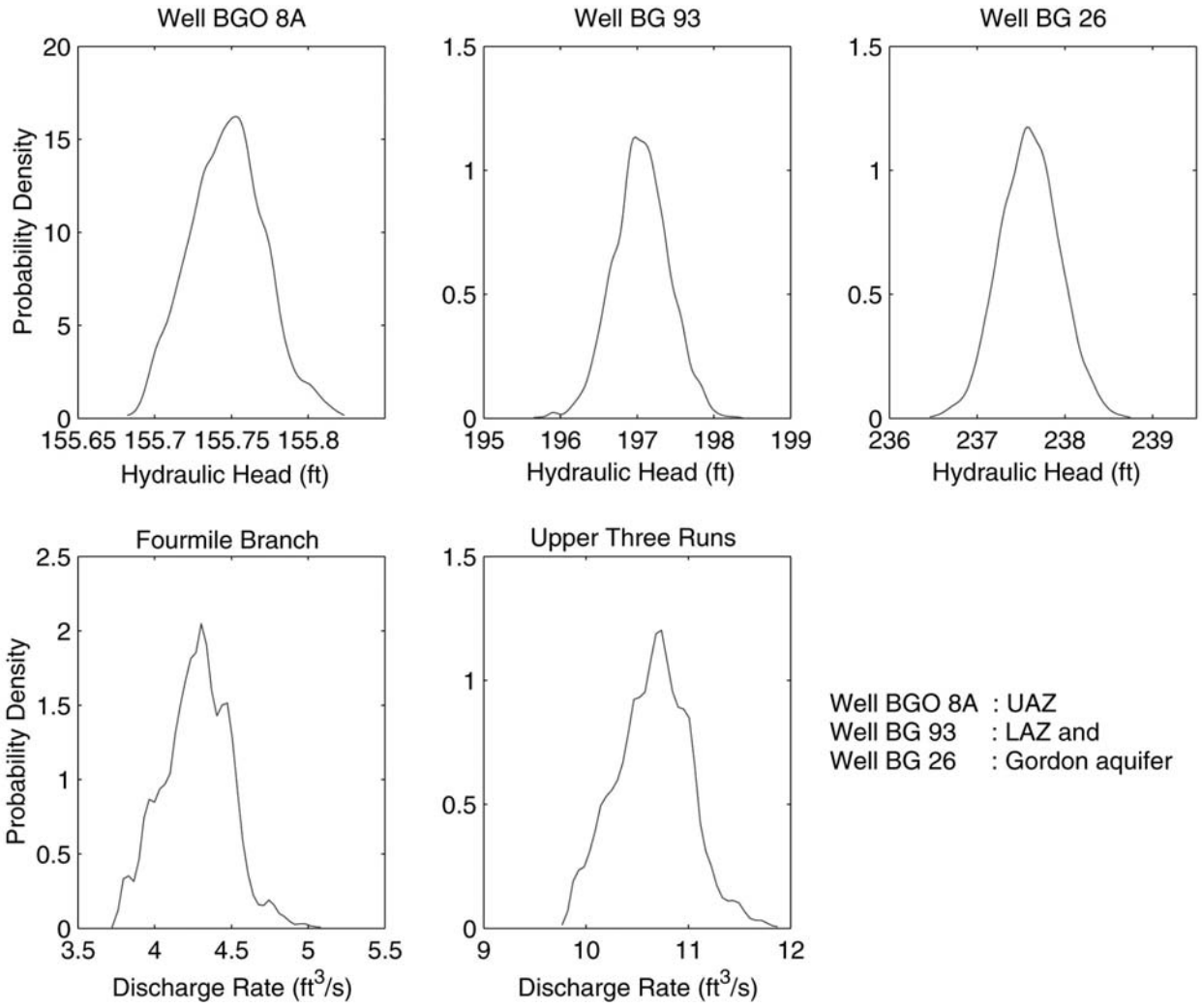

Figure 10. Final marginal probability distributions obtained after the uncertainty analysis based on the SRSM approximated outputs with the joint posterior distribution (obtained after the MCMC simulation) as input. Shown are the marginal distributions for a representative well from each aquifer (top row left to right) and the stream baseflow rates in the Fourmile Branch and the Upper Three Runs