Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Hydrology 307 (2005) 204–218
Journal of Hydrology
www.elsevier.com/locate/jhydrol

A comparative assessment of efficient uncertainty analysis techniques for environmental fate and transport models: application to the FACT model

Suhrid Balakrishnan[a,b], Amit Roy[b], Marianthi G. Ierapetritou[a], Gregory P. Flach[c], Panos G. Georgopoulos[b,*]

[a]*Department of Chemical and Biochemical Engineering, Rutgers University, Piscataway, NJ 08854, USA*
[b]*Environmental and Occupational Health Sciences Institute, UMDNJ—R.W. Johnson Medical School and Rutgers University, Piscataway, NJ 08854, USA*
[c]*Savannah River Technology Center, Savannah River Site, Aiken, SC 29808, USA*

Received 27 June 2003; revised 27 September 2004; accepted 1 October 2004

Abstract

This work presents a comparative assessment of efficient uncertainty modeling techniques, including Stochastic Response Surface Method (SRSM) and High Dimensional Model Representation (HDMR). This assessment considers improvement achieved with respect to conventional techniques of modeling uncertainty (Monte Carlo). Given that traditional methods for characterizing uncertainty are very computationally demanding, when they are applied in conjunction with complex environmental fate and transport models, this study aims to assess how accurately these efficient (and hence viable) techniques for uncertainty propagation can capture complex model output uncertainty. As a part of this effort, the efficacy of HDMR, which has primarily been used in the past as a model reduction tool, is also demonstrated for uncertainty analysis. The application chosen to highlight the accuracy of these new techniques is the steady state analysis of the groundwater flow in the Savannah River Site General Separations Area (GSA) using the subsurface Flow And Contaminant Transport (FACT) code. Uncertain inputs included three-dimensional hydraulic conductivity fields, and a two-dimensional recharge rate field. The output variables under consideration were the simulated stream baseflows and hydraulic head values. Results show that the uncertainty analysis outcomes obtained using SRSM and HDMR are practically indistinguishable from those obtained using the conventional Monte Carlo method, while requiring orders of magnitude fewer model simulations.

Keywords: Mathematical methods; Statistical analysis; Hydrology; Computer programs; Data processing

* Corresponding author. Address: Environmental and Occupational Health Sciences Institute, UMDNJ—R.W. Johnson Medical School and Rutgers University, Piscataway, NJ 08854, USA. Tel.: +1 732 445 0159.
E-mail addresses: suhrid@sol.rutgers.edu (S. Balakrishnan), amit.roy@bms.com (A. Roy), marianth@sol.rutgers.edu (M.G. Ierapetritou), gregory.flach@srs.gov (G.P. Flach), panosg@fidelio.rutgers.edu (P.G. Georgopoulos).

0022-1694/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.jhydrol.2004.10.010

1. Introduction

The presence of uncertainty often complicates the mechanistic modeling of physical systems. Uncertainty arises in such modeling efforts through various channels: natural or irreducible uncertainty, wherein the physical system being modeled itself is inherently uncertain (Brownian motion, etc.); model uncertainty, which is engendered through many correlated factors such as model structure and approximations used, extrapolations and model boundaries, and model resolution; parametric and data uncertainty, which include experimental and data measurement errors, imprecise device calibration biases, etc.

The purpose of systematic uncertainty analysis is to provide insight into the level of confidence in model estimates, identify key sources of uncertainty, and quantify the degree of confidence in the existing data and models. The first step in such an analysis requires the selection of an approach for the representation of uncertainty. Of the gamut of techniques available (set theory, interval mathematics, fuzzy set theory, etc.) probabilistic and statistical representation of uncertainty has gained extremely wide acceptance and is the approach adopted in this study.

Monte Carlo methods are the most widely used techniques for statistical/probabilistic uncertainty analysis, with diverse applications. Given input uncertainty distributions (frequency or probability density data) these methods involve repeated generation of pseudo-random instantiations (sampling) of inputs followed by application of the model to these instantiations to yield a set of model responses. These model outputs are then further analyzed statistically.

An established disadvantage of these traditional sampling based techniques is the large number of model simulations required to achieve acceptable levels of confidence about model output uncertainty characterizations. The large numbers of samples (and hence model simulations) required imply that the applicability of these methods is sometimes limited to relatively simple models. In the case of computationally intensive models, the time and resources required by these methods can easily prove to be prohibitively expensive.

The motivation underlying the development of the Stochastic Response Surface Method (SRSM) was precisely to reduce the number of model simulations required for adequate estimation of uncertainty, as compared to conventional methods. This is accomplished by approximating both inputs and outputs of the uncertain system through series expansions of standard random variables; the series expansions of the outputs contain coefficients which can be calculated from the results of a limited number of model simulations. The net result is to create a statistically equivalent polynomial approximation to the model outputs.

Another tool developed in order to express input–output relations of complex, computationally burdensome models in terms of hierarchical correlated function expansions is the High Dimensional Model Representation (HDMR). Application of the HDMR methodology to a complex nonlinear model also provides an efficient means to obtain an accurate reduced model of the original system. The uncertainty analysis of the outputs of the computationally burdensome model can then be well approximated by a Monte Carlo analysis of the corresponding reduced model outputs, which is thus performed at a much lower computational cost without compromising accuracy (as shown in the analysis that follows).

The uncertainty analysis of the simulation of saturated groundwater flow beneath the US Department of Energy (USDOE) Savannah River Site General Separations Area (GSA) using the subsurface Flow And Contaminant Transport (FACT) code has many desirable features that make it a good case study for the comparison of various methods of efficient uncertainty analysis. On one hand, the model inputs, such as conductivity fields and recharge rate fields, are inherently uncertain but can be fairly well characterized, and on the other hand the model is complex, involving the use of the finite element method for numerical solution of the continuity equation and Darcy's Law, resulting in significant computational demand.

The objective of this work is to show that newly available uncertainty analysis techniques, like SRSM and HDMR, represent output uncertainties very well in complex fate and transport models, while being orders of magnitude more efficient computationally, as compared to traditional Mote Carlo techniques. In a separate recent publication (Balakrishnan et al., 2003) it was shown how SRSM can be used to facilitate

a Bayesian uncertainty reduction analysis; however, the focus of the present article is the computational challenge of the uncertainty analysis itself. That said, readers may find a comparison of the results of the final uncertainty analysis (based on the posterior distributions) in Balakrishnan et al. (2003) and of the results obtained in this manuscript, interesting in its own right (as the present approach would correspond to an uncertainty analysis based on the prior distributions in that article). The following sections of this article outline the model and case study, describe the methods used (Monte Carlo, SRSM and HDMR), and finally present the results of the comparative analysis and related conclusions.

2. Problem description and modeling

2.1. Model description: FACT applied to the GSA

The GSA of the USDOE Savannah River Site covers an 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 GSA model covers the above area and extends from the ground surface to the bottom of the Gordon Aquifer (Fig. 1).

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

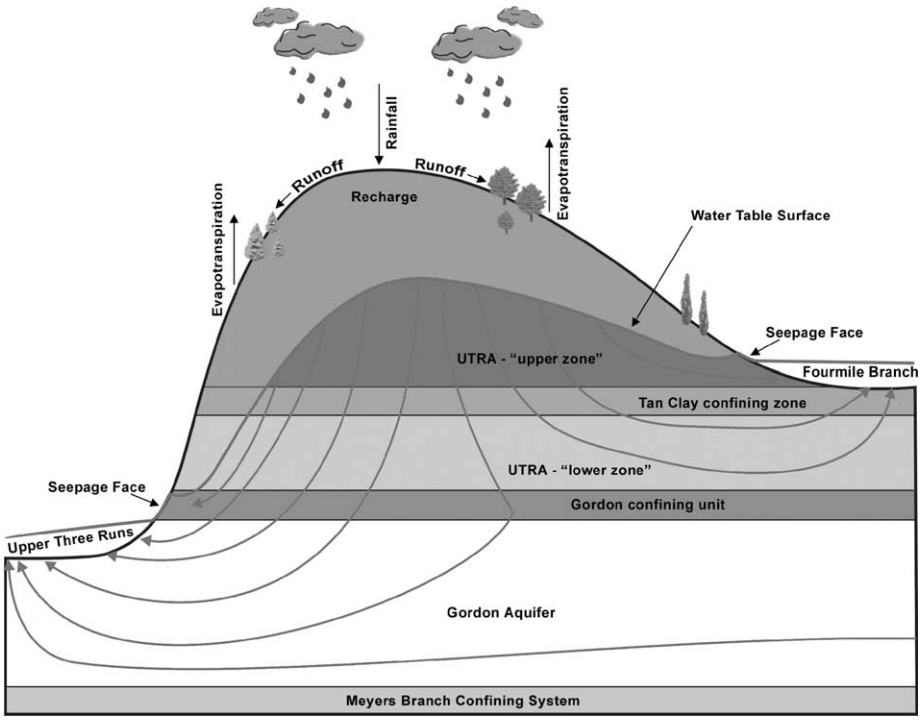

Fig. 1. The simplified conceptual model of the general separations area showing the various aquifer units, recharge/discharge areas and mechanisms.

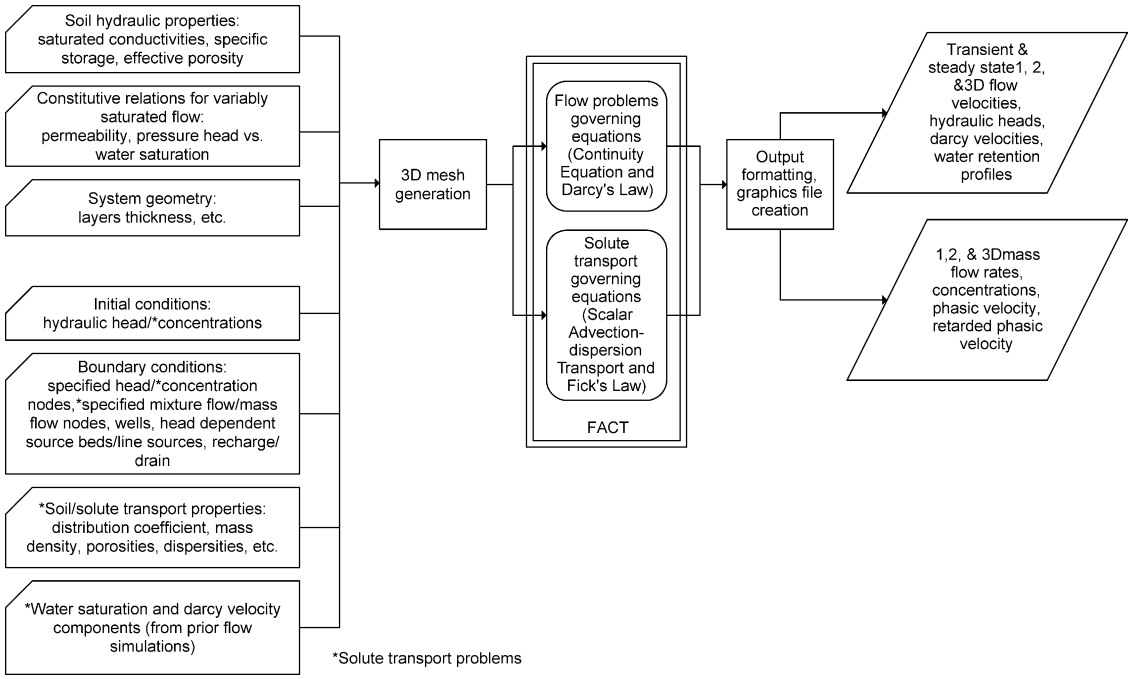

Fig. 2. Flowchart showing the FACT code capabilities and describing the problems it is suited to handle.

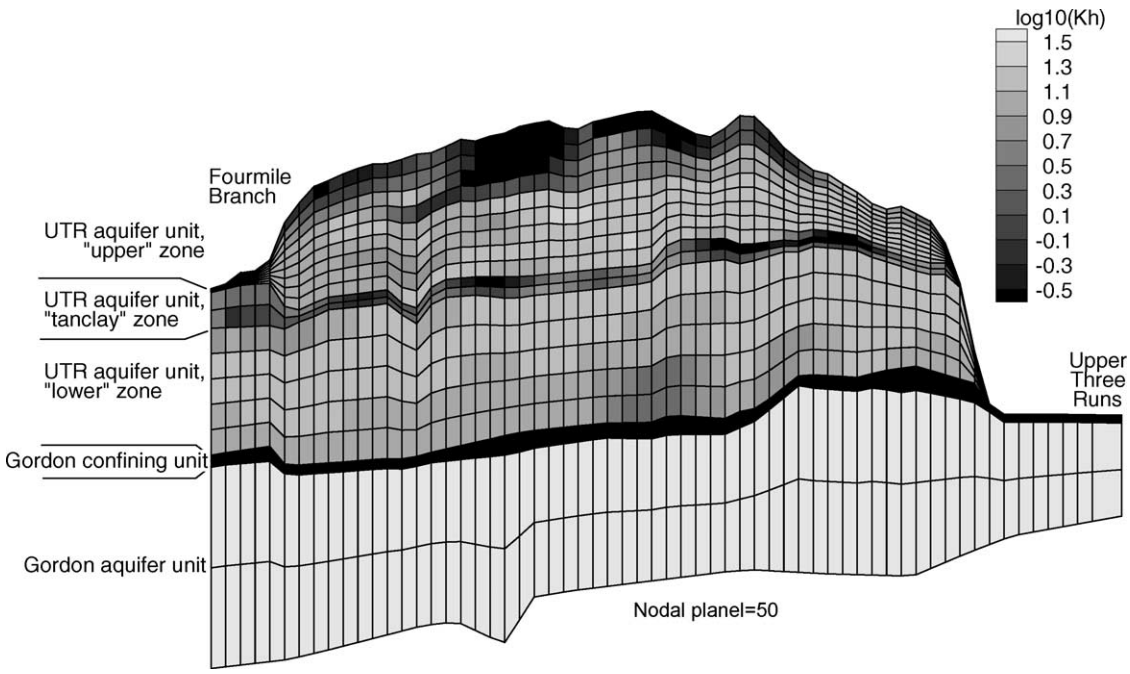

Fig. 3. A typical model cross-sectional view showing hydrostratigraphy and $\log_{10}(K_h)$ field of the GSA simulated by FACT.

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 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) (Fig. 2). 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 and also offers great flexibility in creating grids for complex flow domains.

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 solute transport equation is approximated using an upstream-weighted residual finite element method designed to overcome (or alleviate) numerical oscillations. Transport mechanisms considered include: advection, hydrodynamic dispersion, linear equilibrium adsorption, mobile/immobile first-order mass transfer, first-order degradation and radioactive decay effects.

The area resolution of the model is approximately 18.6 m^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 70–375 depending on the surface. The Upper Aquifer Zone (UAZ) of the UTR Aquifer unit is represented with nine finite-elements in the vertical direction. The vadose zone is included in the model. The Lower Aquifer Zone (LAZ) contains five finite-elements while the Tan Clay Confining Zone (TCCZ) separating the aquifer zones is modeled with two vertical elements. The Gordon confining and Aquifer units each contain two elements, for a total of 20 vertical elements from ground surface to the bottom of the Gordon Aquifer. The three-dimensional mesh size is therefore $108 \times 77 \times 20 = 166{,}320$ elements or $109 \times 78 \times 21 = 178{,}542$ nodes. A typical model cross-sectional view can be seen in Fig. 3.

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.

2.2. Uncertainty analysis: problem formulation

The three-dimensional conductivity fields and the two-dimensional recharge rate field are the uncertain inputs to the saturated flow model for the case study considered in this work. Each zone is heterogeneous and through preliminary analysis, five key uncertain variables for the GSA simulated by the FACT model were identified, namely: the horizontal conductivity field (K_v) values for the Gordon Confining Unit (GCU) and the Tan Clay Confining Zone (TCCZ), the vertical conductivity field (K_h) values for the Lower UTR Aquifer Zone (LAZ) and the Upper UTR Aquifer Zone (UAZ) and the recharge rate (RECH). The problem was simplified for this study 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 assumption retains spatial

variability although in a restricted setting. Specifically, it ensures that relative spatial variations dictated by characterization and subsequent model calibration are preserved, while the variable values of the entire field are perturbed in the uncertainty analysis (Flach and Harris, 2000).

As a further simplification, these global multiplier variables were treated as independent random variables with distributions determined using the best available engineering estimates taking into account field observations. Note that while these are (necessary) simplifications made to render the problem analytically tractable, in our opinion and experience (taking into account the FACT model, GSA field observations and results of the analysis) although they might be slightly simplistic, they are entirely reasonable.

The details of these input distributions are shown in Table 1 (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 $N(\mu,\sigma)$).

The main output variables considered were the simulated hydraulic head values in the various aquifers (direct model outputs) as well as the stream baseflows in the three main discharge regions (Crouch and McQueen Branches and the Upper Three Runs) which are model post processed results. The hydraulic head values used for this study were obtained from the code at specific locations (locations where wells were drilled and field measurements taken for model calibration uses) and could be grouped according to their corresponding aquifer. The three groups have 79, 173 and 415 wells located in the upper UTR, lower UTR and Gordon Aquifer zones, respectively (a total of 667 wells). The problem addressed in this work was the determination of the empirical distribution functions (and typical probability density functions) of the outputs, in the face of the parametric uncertainty as outlined above. It serves to note that it was not only important to have estimates of the output distributions but also to have some indication of how good such estimates were.

Table 1
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	m/d	$\log_{10}$ Normal	−5.516, 0.2286
LAZ K_h	m/d	$\log_{10}$ Normal	0.4525, 0.0046
TCCZ K_v	m/d	$\log_{10}$ Normal	−2.716, 0.1524
UAZ K_h	m/d	$\log_{10}$ Normal	0.4663, 0.046
RECH	cm/year	Normal	45.72, 7.87

3. Methods used for uncertainty analysis

3.1. The Monte Carlo method

The Monte Carlo method, as applied to uncertainty analysis for empirical distribution/probability density determination, requires that for each input parameter that has associated uncertainty or variability, a probability distribution (or frequency distribution) be provided. The method then involves the repeated generation of independent pseudo-random values of the uncertain input variables (drawn from the known distribution and within the range of any imposed bounds) followed by the application of the model using these values to generate a set of model responses or outputs (for example, the well hydraulic head values). These responses are then analyzed statistically to yield the empirical distribution function/probability distribution of the model outputs.

One of the main advantages of this method is its ease of application. Another very useful feature of applying a Monte Carlo analysis is that the estimates on the responses obtained can be bounded within chosen confidence limits. If $X_1, X_2,\ldots, X_n$ are independent observations (output realizations or responses for a single variable) each having the same distribution function $U(x)=Pr(X_i<x)$ (the population distribution required to be estimated), and $F_n(x)$ is the empirical distribution function (or cumulative step-function) defined as the proportion of the $X_1, X_2,\ldots, X_n$ which are less than x (i.e. $F_n(x)=k/n$ where k is the number of observations less than or equal to x), then by the Strong Law of large numbers

$$F_n(x) \rightarrow U(x)$$

with probability 1 for each x. Also, even if $F_n(x)$ is unknown, if it is continuous, one can then bound the maximum deviation $d=\max|F_n(x)-U(x)|$ given

specific confidence limits using the Kolmogorov–Smirnov criterion (Kolmogorov, 1941; D'Agostino and Stephens, 1986). Alternatively, with known confidence limits, one can use the Kolmogorov–Smirnov criterion to determine the Monte Carlo sample size required to keep $F_n(x)$ within a specified deviation of the population distribution $U(x)$ (Massey, 1951). Note that this is not possible for slightly more efficient sampling based schemes like Latin Hypercube methods due to the dependency of the samples generated. For example, the estimation of the cumulative sampling (for the entire curve) within 0.02 percentage points with 99% confidence limits would require a sample size of approximately 6,650. While sample sizes of this magnitude may present no problem for computationally tractable models, it is fairly evident how analysis by these techniques readily becomes infeasible for expensive models if it is not desired to compromise on accuracy.

3.2. The Stochastic Response Surface Method

SRSM (Isukapalli et al., 1998, 2000; Isukapalli, 1999; Isukapalli and Georgopoulos, 1999) is an extension of the classical deterministic Response Surface Method (RSM) and the Deterministic Equivalent Modeling Method (DEMM) (Tatang, 1995). The motivation underlying the use of the SRSM is to reduce the number of model simulations required for adequate estimation of uncertainty, as compared to conventional methods. This is accomplished by approximating both 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 model simulations. Evaluating an SRSM expansion consists of the following steps (Fig. 4): (1) input uncertainties are expressed in terms of a set of standard random variables (*srv*s), (2) a functional form is assumed for selected outputs or output metrics, and (3) 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 ith model input X_i, is expressed directly as a function of the ith *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 2 presents a list of transformations for some probability distributions commonly employed in transport-transformation modeling.

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

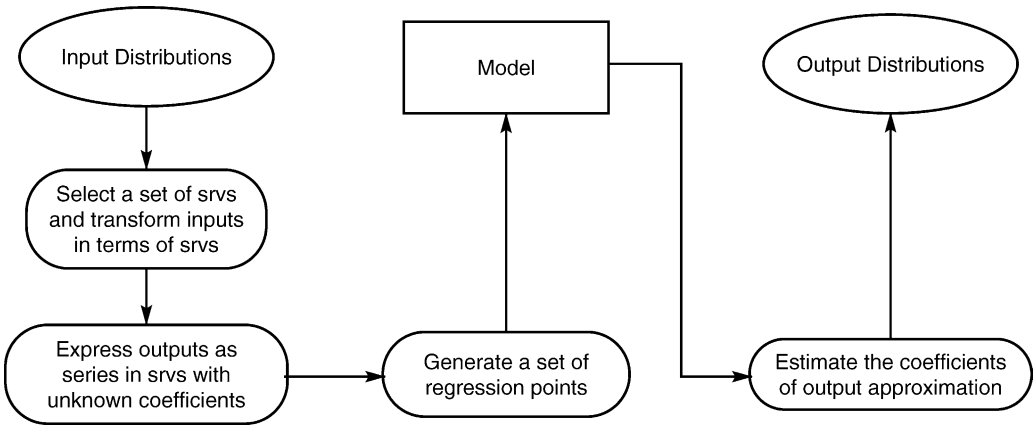

Fig. 4. Flowchart outlining the steps involved in the evaluation of a SRSM polynomial.

Table 2
Representation of common univariate distributions as functions of standard normal random variables, ξ's (which are needed in order to represent input distributions in terms of *srv*s for SRSM)

Distribution type	Transformation[a]
Uniform (a,b)	$a+(b-a)\left(\frac{1}{2}+\frac{1}{2}(\xi/\sqrt{2})\right)$
Normal (μ,σ)	$\mu+\sigma\xi$
log Normal (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)	$y^{1/a}$
Extreme value	$-\log(z)$

[a] ξ is Normal (0,1) and z is Exponential (1) distributed.

$$y = a_0 + \sum_{i_1=1}^{n} a_{i_1}\Gamma_1(\xi_{i_1}) + \sum_{i_1=1}^{n}\sum_{i_2=1}^{i_1} a_{i_1 i_2}\Gamma_2(\xi_{i_1},\xi_{i_2}) + \sum_{i_1=1}^{n}\sum_{i_2=1}^{i_1}\sum_{i_3=1}^{i_2} a_{i_1 i_2 i_3}\Gamma_3(\xi_{i_1},\xi_{i_2},\xi_{i_3}) + \ldots, \quad (1)$$

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

$$\Gamma_p(\xi_{i_1},\ldots,\xi_{i_p}) = (-1)^p \mathrm{e}^{(1/2)\xi^{\mathrm{T}}\xi}\frac{\partial^p}{\partial\xi_{i_1}\ldots\partial\xi_{i_p}}\mathrm{e}^{(-1/2)\xi^{\mathrm{T}}\xi}, \quad (2)$$

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 is convergent 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 U, can be expressed as first, second and third order SRSM polynomial approximations, U_1, U_2 and U_3 as follows

$$U_1 = a_{0,1} + \sum_{i=1}^{n} a_{i,1}\xi_i \quad (3)$$

$$U_2 = a_{0,2} + \sum_{i=1}^{n} a_{i,2}\xi_i + \sum_{i=1}^{n} a_{ii,2}(\xi_i^2-1) + \sum_{i=1}^{n-1}\sum_{j>i}^{n} a_{ij,2}\xi_i\xi_j \quad (4)$$

$$U_3 = a_{0,3} + \sum_{i=1}^{n} a_{i,3}\xi_i + \sum_{i=1}^{n} a_{ii,3}(\xi_i^2-1) + \sum_{i=1}^{n} a_{iii,3}(\xi_i^3-3\xi_i) + \sum_{i=1}^{n-1}\sum_{j>i}^{n} a_{ij,3}\xi_i\xi_j + \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ijj,3}(\xi_i\xi_j^2-\xi_i) + \sum_{i=1}^{n-2}\sum_{j>i}^{n-1}\sum_{k>j}^{n} a_{ijk,3}\xi_i\xi_j\xi_k \quad (5)$$

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 (SRSM expansion), which is done using an extension to collocation methods based on a combination of regression and an improved input collocation scheme called the Efficient Collocation Method (ECM) (Isukapalli, 1999). In the ECM, points are selected based on a modification of the standard orthogonal collocation method of (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. A simple heuristic technique is used to select the required number of points from the large number of potential candidates: for each term of the series expansion, a 'corresponding' collocation point is selected. For example, the collocation point corresponding to the constant is the origin; i.e. all the standard normal

variables (ξ_i's) are set to value zero. For terms involving only one variable, the collocation points are selected by setting all other ξ_i's to zero value, and by letting the corresponding variable take values as the roots of the higher order Hermite polynomial. For terms involving two or more random variables, the values of the corresponding variables are set to the values of the roots of the higher order polynomial, and so on. If more points 'corresponding' to a set of terms are available than needed, the points which are closer to the origin are preferred, as they fall in regions of higher probability. Further, when there is still an unresolved choice, the collocation points are selected such that the overall distribution of the collocation points is more symmetric with respect to the origin. If still more points are available, the collocation point is selected randomly. 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 et al. (1998, 2000); Isukapalli (1999); Isukapalli and Georgopoulos (1999) and Balakrishnan et al. (2002).

After the set of sample inputs 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 the singular value decomposition method.

The result is a large increase of efficiency in terms of computational time required for uncertainty analysis if model runs are expensive (due to the small number of model simulations required). A large sample Monte Carlo analysis of the outputs as approximated by SRSM polynomial expansions (of the appropriate order) will accurately represent the full model output Monte Carlo analysis and can be evaluated at a fraction of the computational cost (polynomial evaluations vs full model evaluations).

3.3. High Dimensional Model Representation

The HDMR method is a family of tools (Rabitz et al., 1998; Rabitz and Alis, 1999), which prescribe systematic sampling procedures to map out the relationships between sets of input and output model variables. Let the n-dimensional vector $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$ represent the input variables of the model under consideration, and $f(\mathbf{x})$ be one of the output variables. Since the influence of the input variables on the output variable can be independent and/or cooperative, using HDMR one expresses the output $f(\mathbf{x})$ as a hierarchical correlated function expansion in terms of the input variables as

$$f(\mathbf{x}) = f_0 + \sum_{i=1}^{n} f_i(x_i) + \sum_{1\le i<j\le n} f_{ij}(x_i, x_j) + \sum_{1\le i<j<k\le n} f_{ijk}(x_i, x_j, x_k) + \cdots + f_{12\ldots n}(x_1, x_2, \ldots, x_n) \quad (6)$$

Here f_0 denotes the mean effect which is a constant. The function $f_i(x_i)$ is a first-order term expressing the effect of variable x_i acting independently, although generally nonlinearly, upon the output $f(x)$. The function $f_{ij}(x_i,x_j)$ is a second-order term describing the cooperative effects of the variables x_i and x_j upon the output $f(\mathbf{x})$. The higher-order terms reflect the cooperative effects of increasing numbers of input variables acting together to influence the output $f(\mathbf{x})$. The last term $f_{12,\ldots,n}(x_1,x_2,\ldots,x_n)$ gives any residual dependence of all the input variables locked together in a cooperative way to influence the output $f(\mathbf{x})$. After the relevant component functions in Eq. (6) are determined and suitably represented, then the expressions constitute the HDMR, thereby replacing the original method of calculating $f(\mathbf{x})$ by the computationally expensive model. Usually only low order correlations amongst the input variables are typically adequate in describing the output behavior and therefore it is expected that the HDMR expansion converges very rapidly. This has been verified in a number of computational studies (Shim and Rabitz, 1998; Rabitz and Shim, 1999; Shorter et al., 1999; Wang et al., 1999; Li et al., 2001) where the HDMR expansions up to second order are often sufficient to describe the outputs of many realistic systems.

In this work, the Cut-HDMR procedure will be used to compute the expansion terms. With the Cut-HDMR method, first a reference point $\bar{\mathbf{x}} = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n)$ is

defined in the variable space. In the convergence limit, the Cut-HDMR is invariant to the choice of reference point $\bar{\mathbf{x}}$. In practice, $\bar{\mathbf{x}}$ is chosen within the neighborhood of interest in the input space. The expansion functions are determined by evaluating the input–output responses of the system relative to the defined reference point $\bar{\mathbf{x}}$ along associated lines, surfaces, subvolumes, etc. (i.e. cuts) in the input variable space. This process reduces to the following relationship for the component functions in Eq. (6)

$$f_0 = f(\bar{\mathbf{x}}), \tag{7}$$

$$f_i(x_i) = f(x_i, \bar{\mathbf{x}}^i) - f_0, \tag{8}$$

$$f_{ij}(x_i, x_j) = f(x_i, x_j, \bar{\mathbf{x}}^{ij}) - f_i(x_i) - f_j(x_j) - f_0, \ldots \tag{9}$$

where the notation $f(x_i, \bar{\mathbf{x}}^i) \equiv f(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_{i-1}, x_i, \bar{x}_{i+1}, \ldots, \bar{x}_n)$ denotes that all the input variables are at their reference point values except x_i.

The f_0 term is the output response of the system evaluated at the reference point $\bar{\mathbf{x}}$. The higher-order terms are evaluated as cuts in the input variable space through the reference point. Therefore, each first-order term $f_i(x_i)$ is evaluated along its variable axis through the reference point. Each second-order term $f_{ij}(x_i,x_j)$ is evaluated in a plane defined by the binary set of input variables x_i, x_j through the reference point, etc. The process of subtracting off the lower-order expansion functions removes their dependence to assure a unique contribution from the new expansion function.

In practice, each of the HDMR expansion functions is numerically represented as a low-dimensional look-up table over its variables. Note that the HDMR in Eq. (6) is exact along any of the cuts, and the output response $f(\mathbf{x})$ at a point $\mathbf{x}$ off of the cuts can be obtained by the following procedure:

1. Interpolate each of the low dimensional HDMR expansion terms in the look-up tables with respect to the input values of the point $\mathbf{x}$, and
2. Sum the interpolated values of the HDMR terms from zeroth order to the highest order retained in keeping with the desired accuracy.

Uncertainty analysis using HDMR relies on an accurate reduced model being generated with a small number of full model simulations. An arbitrarily large sample Monte Carlo analysis can be performed on the outputs as approximated by HDMR and should accurately result in the same distributions as obtained through the Monte Carlo analysis of the full model. The tremendous computational savings result from just having to perform interpolation instead of full model simulations for output determination.

4. Application of methods for uncertainty analysis

With the problem formally outlined in the problem formulation subsection, five independent uncertain parameters with known distributions were identified: the global multipliers for the GCU and TCCZ K_v fields, the LAZ and UAZ K_h fields and that for the recharge rate field (Table 1). A small sample ($N=1000$, which represents the approximate computational limit for this elaborate model) Monte Carlo analysis was carried out by running the full model (FACT simulated GSA model) at random instantiations of the input distributions. Well location hydraulic head values for all three aquifers and stream baseflow values were collated corresponding to these input points and the output empirical distribution functions and representative probability density functions obtained.

For the application of the SRSM, the same independent input parameter distributions were used. Collocation points were generated via the ECM, transformed to appropriate input points, the model run at these specified points and results collated in order to generate SRSM polynomials for outputs under consideration. For this study, second order SRSM polynomials were generated for each of the 667 well hydraulic head values as well as for each of the stream baseflow values. Thus, in effect, a second order SRSM approximate model for the GSA as simulated by FACT was obtained which required only 51 full model simulations.

Cut-Plane HDMR was also used to create a functional approximation to the full model for the hydraulic head values and stream baseflow values. This required quantization of the input parameter space for calculation of output functions. The same input distributions' ranges were used for this purpose with upper and lower limits defined by nominal $\pm 2\sigma$ values, respectively, the justification being that probabilistically this range (4σ) encompassed more

than 95% of the data variability. First order HDMR was used to create a reduced model for all of the well hydraulic head values as well as for each of the stream baseflow values and required 45 full model simulations (with nine cut points per variable). The reduced models generated by SRSM and HDMR were then subsequently used for the uncertainty analysis of the outputs. A large sample Monte Carlo analysis ($N=10{,}000$) was used to estimate model output empirical distribution functions as well as representative probability density functions.

5. Results and discussion

Empirical distribution functions and representative probability densities were obtained from the small sample ($N=1000$) Monte Carlo analysis (MC) carried out for all hydraulic head values as well as for the stream baseflow rates. These small sample MC empirical distribution functions obtained were bounded using the KS goodness of fit criterion (for 99% confidence limits). The empirical distribution functions were then plotted against those obtained via SRSM and HDMR (through a Monte Carlo analysis on the corresponding reduced models). Probability density plots were also obtained using a Gaussian kernel density estimate with automatic bandwidth selection via a two-stage Solve-the-Equation Plug-In approach, which has been recommended as being the most reliable in terms of overall performance by a number of authors (Jones et al., 1996; Wand and Jones, 1995).

As can be clearly seen from Figs. 5–9, the results show excellent agreement, implying that the uncertainty characteristics have been well captured by both SRSM and HDMR for the stream baseflow rates and all the hydraulic head values. This claim is further strengthened by the results of the application of the two-sample KS tests performed to check whether the responses obtained through the small sample Monte Carlo analysis (on the full model) and those obtained from the SRSM and HDMR analyses are derived from the same population. For 99% confidence limits, the hypothesis that the two sets of outputs, one from the small sample Monte Carlo analysis and one from either SRSM or HDMR, come from the same population, is

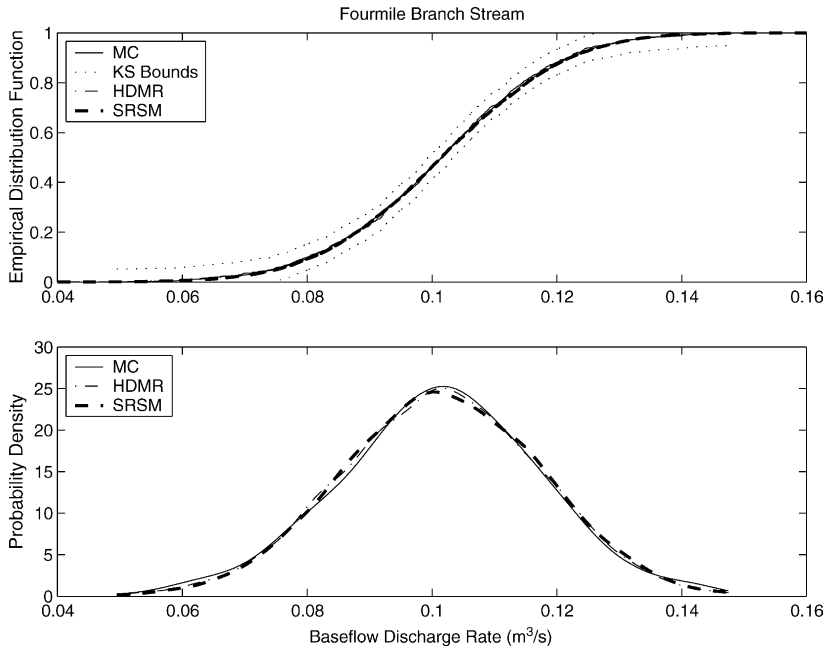

Fig. 5. Fourmile Branch stream baseflow rate (m^3/s) empirical distribution function and corresponding representative probability density for the small sample Monte Carlo run for the full model (MC) and associated KS bounds (KS Bounds) along with those obtained from the HDMR and SRSM Monte Carlo analysis.

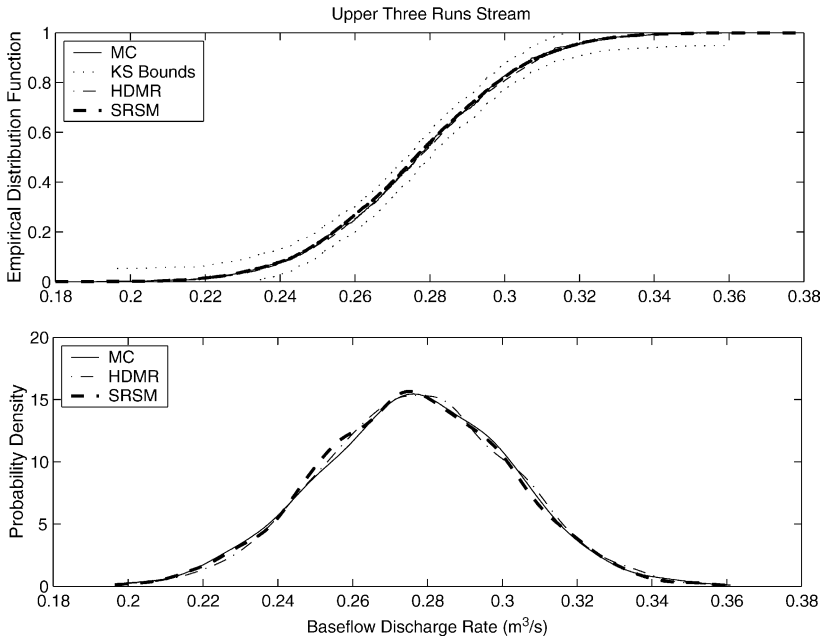

Fig. 6. Upper Three Runs stream baseflow rate (m^3/s) empirical distribution function and corresponding representative probability density for the small sample Monte Carlo run for the full model (MC) and associated KS bounds (KS Bounds) along with those obtained from the HDMR and SRSM Monte Carlo analysis.

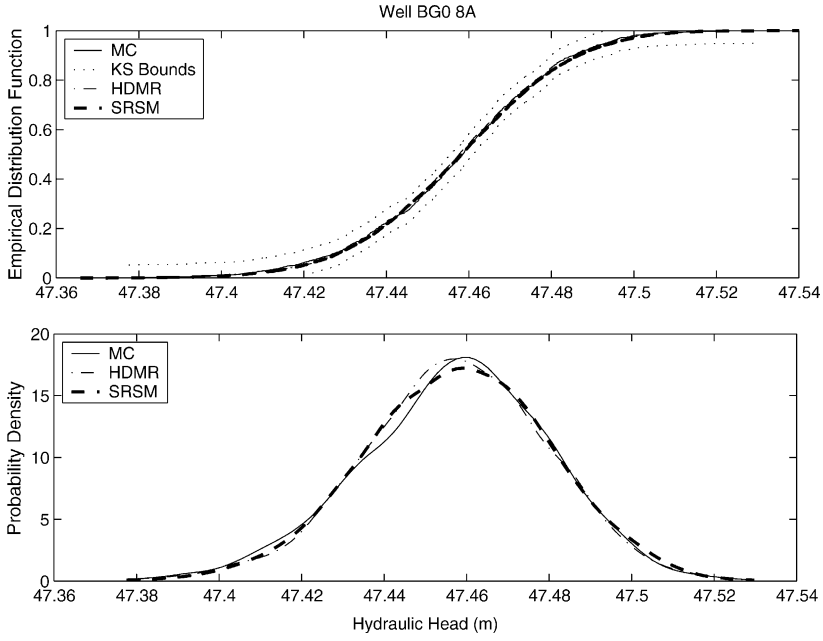

Fig. 7. Well BGO 8A (located in the upper UTR aquifer, UAZ) hydraulic head value empirical distribution function and corresponding representative probability density for the small sample Monte Carlo run for the full model (MC) and associated KS bounds (KS Bounds) along with those obtained from the HDMR and SRSM Monte Carlo analysis.

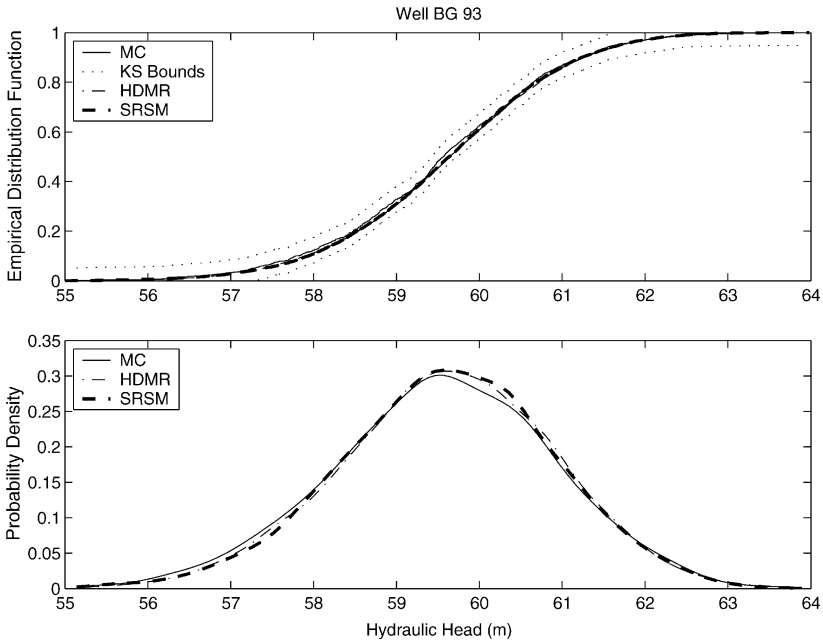

Fig. 8. Well BG 93 (located in the lower UTR Aquifer, LAZ) hydraulic head value empirical distribution function and corresponding representative probability density for the small sample Monte Carlo run for the full model (MC) and associated KS bounds (KS Bounds) along with those obtained from the HDMR and SRSM Monte Carlo analysis.

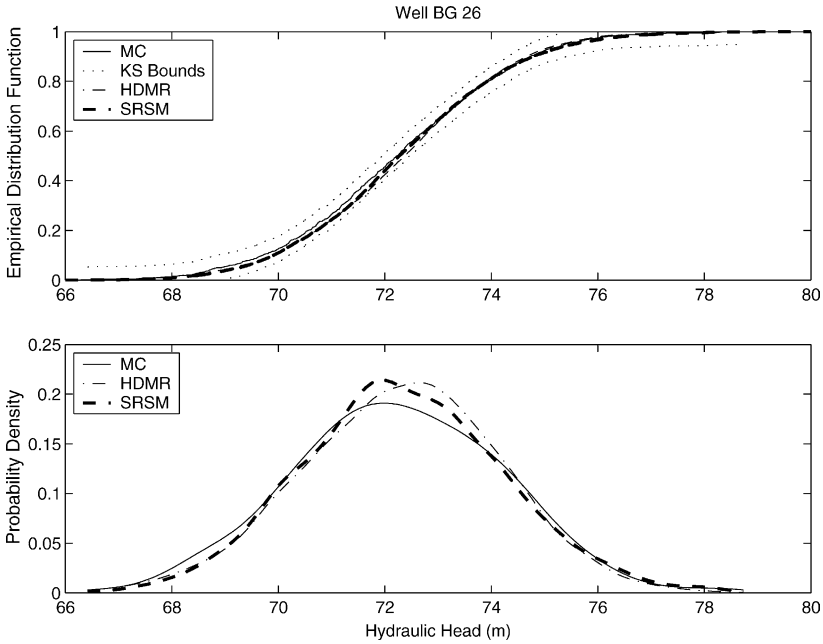

Fig. 9. Well BG 26 (located in the Gordon aquifer) hydraulic head value empirical distribution function and corresponding representative probability density for the small sample Monte Carlo run for the full model (MC) and associated KS bounds (KS Bounds) along with those obtained from the HDMR and SRSM Monte Carlo analysis.

accepted for all cases. This also established both convergence and sufficiency of second order SRSM and first order HDMR for this case study.

The motivation of this work was to highlight the effectiveness of SRSM and HDMR as efficient and accurate methods of uncertainty analysis for computationally intensive models where tight bounds on accuracy of estimates are often difficult or impossible to achieve using traditional Monte Carlo methods for investigation. The application of these methods to the GSA modeled by the FACT code illustrates how they require orders of magnitude less computational time and capture output uncertainties with high accuracy. Future work in this direction will include incorporation of field data into the uncertainty analysis.

Acknowledgments

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 with deep insight into key areas of the problem and its formulation. This work has been funded in part by the US 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 US Department of Energy, Instrument DE-FG2600NT 40938 and the Petroleum Research Fund (administered by the ACS). The viewpoints expressed in this work are solely the responsibility of the authors and do not necessarily reflect the views of the US Department of Energy, the US Environmental Protection Agency, or their contractors.

References

Balakrishnan, S., Georgopoulos, P., Banerjee, I., Ierapetritou, M., 2002. Uncertainty considerations for describing complex reaction systems. Aiche Journal 48 (12), 2875–2889.

Balakrishnan, S., Roy, A., Ierapetritou, M.G., Flach, G.P., Georgopoulos, P.G., 2003. Uncertainty reduction and characterization of complex environmental fate and transport models: an empirical Bayesian framework incorporating the stochastic response surface method. Water Resources Research 39 (12), 1350–1362.

D'Agostino, R., Stephens, M., 1986. Goodness-of-Fit Techniques. Marcel Dekker, Inc., New York.

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

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

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

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

Isukapalli, S.S., Georgopoulos, P.G., 1999. Computational Methods for Efficient Sensitivity and Uncertainty Analysis of Models for Environmental and Biological Systems. CCL/EDMAS-03, Piscataway, NJ.

Isukapalli, S.S., Roy, A., Georgopoulos, P.G., 1998. Stochastic Response Surface Methods (SRSMs) for uncertainty propagation: application to environmental and biological systems. Risk Analysis 18 (3), 351–363.

Isukapalli, S.S., Roy, A., Georgopoulos, P.G., 2000. Efficient sensitivity/uncertainty analysis using the combined stochastic response surface method and automated differentiation: application to environmental and biological systems. Risk Analysis 20 (5), 591–602.

Jones, M.C., Marron, J.S., Sheather, S.J., 1996. A brief survey of bandwidth selection for density estimation. Journal of the American Statistical Association 91 (433), 401–407.

Kolmogorov, A., 1941. Confidence limits for an unknown distribution function. Annals of Mathematical Statistics 12, 461–463.

Li, G., Rabitz, H., Wang, S., Jaffe, P., Wang, S.W., Georgopoulos, P.G., 2001. Efficient Sensitivity/uncertainty Analysis Using the High Dimensional Model Representation (HDMR) Method: Application to the Princeton Groundwater Model. CCL-TR-01.01. Computational Chemodynamics Laboratory, Piscataway, NJ.

Massey, F.J., 1951. The Kolmogorov–Smirnov test for goodness of fit. Journal of the American Statistical Association 46 (253), 68–78.

Rabitz, H., Alis, O., 1999. General foundations of high dimensional model representations. Journal of Mathematical Chemistry 25, 197–233.

Rabitz, H., Shim, K., 1999. Multicomponent semiconductor material discovery using a generalized correlated function expansion. Journal of Physical Chemistry A 111 (23), 10640–10651.

Rabitz, H., Alis, O., Shorter, J., Shim, K., 1998. Efficient input–output model representations. Computer Physics Communications 115, 1–10.

Shim, K., Rabitz, H., 1998. Independent and correlated composition behavior of the energy band gaps for the ga alpha alloys. Physical Review B 58, 1940–1946.

Shorter, J., Ip, P.C., Rabitz, H., 1999. An efficient chemical kinetics solver using high dimensional model representations. Journal of Physical Chemistry A 103 (36), 7192–7198.

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

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

Wand, M.P., Jones, M.C., 1995. Kernel Smoothing. Chapman and Hall, London.

Wang, S.W., Levy II., H., Li, G., Rabitz, H., 1999. Fully equivalent operational models for atmospheric chemical kinetics within global chemistry-transport models. Journal of Geophysical Research 104 (D23), 30417–30426.