Stochastic Response Surface Method (SRSM), Automatic Differentiation, and Bayesian Approaches for Uncertainty Propagation and Parameter Estimation

Presented at the NASA Langley Research Center/National Institute of Aerospace as part of the Engineering for Complex Systems (ECS) Lecture Series
March 4, 2005

by
Sastry S. Isukapalli

Computational Chemodynamics Laboratory (CCL) (http://www.ccl.rutgers.edu) Environmental and Occupational Health Sciences Institute (EOHSI)
a joint Institute of University of Medicine and Dentistry, New Jersey – Robert Wood Johnson Medical School and Rutgers, The State University of New Jersey

170 Frelinghuysen Road, Piscataway New Jersey 08854
Acknowledgments

People

- Dr. Panos G. Georgopoulos (CCL-EOHSI; Doctoral Thesis Adviser, and Currently Guide and Supervisor)
- Dr. Amit Roy (Currently at Bristol Myers Squibb; Guidance and Collaboration on SRSM, SRSM-ADIFOR, and Pharmacokinetic Applications)
- Suhrid Balakrishnan (Collaboration on the use of SRSM in Bayesian Inference)
- Several other researchers, and open source tool developers, whose work has extensively utilized in this research.

Major Funding Agencies

- US Environmental Protection Agency (USEPA): Funding for the Center for Exposure and Risk Modeling (CERM)
- New Jersey Department of Environmental Protection (NJDEP): Providing base funding for the Ozone Research Center at EOHSI
- National Institute of Environmental Health Sciences (NIEHS): Proving funding for the NIEHS Center of Excellence at EOHSI
Introduction – Uncertainty in Risk Assessment Modeling

- Natural Uncertainty: inherent randomness and variability
- Model Uncertainty: spatial/temporal resolution, assumptions, and model resolution
- Input Data/Parametric Uncertainty: errors in measurements, surrogate data
- Evaluation Data Uncertainty: can manifest eventually as model/parametric uncertainty
Methods for Characterization of Uncertainty

Existing methods include: interval mathematics, fuzzy theory, and probabilistic analysis.

• **Interval mathematics:**
  - simplest way to represent uncertainty, when only the bounds of parameters under concern are available
  - each uncertain parameter is described by an interval number
  - does not provide adequate information about the nature of output uncertainty
  - all uncertainties are forced into one arithmetic interval

• **Fuzzy theory:**
  - facilitates uncertainty analysis of systems where uncertainty is due to vagueness or “fuzziness” rather than due to randomness alone
  - can be applied to uncertainty analysis with imprecise observation values or with verbal descriptions

• **Probabilistic analysis:**
  - most widely used method for characterizing uncertainty in physical systems, especially when estimates of input probability distributions are available
  - can describe uncertainty arising from stochastic disturbances, variability conditions, and risk considerations
Some Components of Uncertainty Analysis

- Obtaining estimates of uncertainty in inputs (e.g. through measurements)
- Propagating input uncertainty to estimate uncertainty in outputs
- Identifying key input uncertainties w.r.t. different output metrics
- Identifying different possible model formulations and comparing them
- Estimating uncertainty associated with evaluation data
- Characterizing the model/parameter uncertainty arising from uncertainties in evaluation data (e.g. when evaluation data are used in a Bayesian parameter estimation, or model selection process)
- Utilizing new data to reduce model/parameter uncertainty

Uncertainty Propagation

![Uncertainty Propagation Diagram]

\( P(X_i) \rightarrow P(Y_j) \)

\( X_1, X_k \rightarrow Y_1, Y_n \)

Inputs/Parameters

Outputs

Uncertainty Propagation
Motivation for Efficient Uncertainty Propagation Techniques

There is a need to

- provide uncertainty information to decision makers
- identify key factors that contribute to the uncertainties the most
- utilize new data in order to reduce model and parameter uncertainties

However,

- Coupling of multiple computational models results in a “nested system” of uncertainties variabilities, requiring several stages (or dimensions) of traditional Monte Carlo simulations
  - a meteorological model (modeling atmospheric turbulence as well as large scale effects)
  - an atmospheric air quality model that considers transport and transformation of reactive species
  - distributions of population demographics
  - models for microenvironmental transport, human activity patterns, pharmaco- and toxico-kinetics, and
  - detailed dose response models
Motivation for Efficient Uncertainty Propagation Techniques (contd...)

Traditional Methods

- Monte Carlo and Latin Hypercube Sampling
- computationally demanding (require large number of model simulations)
- they require even more resources for obtaining “sensitivity information”
- Computing power is advancing rapidly, but computational models are becoming more intensive as well*
- Resulting models typically require “a few minutes” to “a few days” CPU time
- Application of straightforward sampling approaches is likely to stay computationally demanding.†

*“640 KB ought to be enough for every body”; “The world needs may be 5 computers”
†Quantum computers can potentially evaluate all combinations in one single “model run”
• Some Alternative methods:
  – Deterministic Equivalent Modeling Method (DEMM) [Tatang, 1995]
  – Stochastic Response Surface Method (SRSM) [Isukapalli et al., 1998]
  – SRSM-ADIFOR (SRSM coupled with automated computer code differentiation techniques) [Isukapalli and Georgopoulos, 2001]

• Recent developments:
  – SRSM coupled with Bayesian parameter estimation techniques for uncertainty reduction [Balakrishnan, 2003]
  – SRSM coupled with moving least squares approach and applied to the design of electric motors [Kim, 2004]
  – Implemented as a computational module in the MENTOR† system
  – Provided as a web interface for application to black box type models (http://ccl.rutgers.edu/srsm.htm)
  – Current work in progress on exploring standard Bayesian techniques for improving the estimates of the SRSM approximations

†MENTOR: Modeling ENvironment for TOtal Risk studies (http://www.cerm.org)
Response Surface Methods (RSMs) [Box and Draper, 1987]

- A polynomial (or other algebraic) function is selected *apriori* to approximate a model
- Unknown coefficients in the polynomial functions are estimated using known model responses
- The fully specified polynomial function is used to estimate model responses as a function of inputs
Stochastic Response Surface Method (SRSM)

- Based on the approach of response surface methods
- Uncertain inputs (model parameters and input variables) are expressed as functions of a set of “standard random variables” ($srv$s; typically iid unit normal random variables, $N(0,1)$; several other types of distributions can also be used (e.g. iid Exponential(1))
- Uncertain responses are expressed as a hermite polynomials of the $srv$s with unknown coefficients (polynomial chaos expansion); other types of orthogonal polynomial expansions can be used depending on the choice of the $srv$s.
- The model is solved for a set of sample points (that depend upon the number of uncertain parameters)
- Polynomial coefficients are estimated by regression on model calculated model responses
- Fully specified polynomials are random variables that represent the uncertainty in model responses
- Monte Carlo simulation is performed on the polynomial functions to estimate the uncertainty in model responses
- Coefficient values encompass a quantitative measure of the contribution of uncertainties in inputs to uncertainties in model outputs
Basic Formulation of the SRSM

Inputs: \( X_i = f(\xi_1, \xi_2, \ldots, \xi_n), \quad i = 1, \ldots, n \)

Responses: 
\[
y = a_0 + \sum_{i_1=1}^{n} a_{i_1} \Gamma_1(\xi_{i_1}) + \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} a_{i_1i_2} \Gamma_2(\xi_{i_1}, \xi_{i_2}) + \ldots
\]
\[
+ \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \sum_{i_3=1}^{n} a_{i_1i_2i_3} \Gamma_3(\xi_{i_1}, \xi_{i_2}, \xi_{i_3}) + \ldots
\]
\[
\Gamma_p(\xi_{i_1}, \ldots, \xi_{i_p}) = (-1)^p e^{\frac{1}{2} \xi^T \xi} \frac{\partial^p}{\partial \xi_{i_1} \ldots \partial \xi_{i_p}} e^{-\frac{1}{2} \xi^T \xi} \quad \text{(Hermite Polynomials)}
\]
Application of the SRSM:

- **Explicit Systems**
  - $Y = G(X)$
  - Direct solution
  - $Y = G(F(\xi))$
  - Mathematically tractable equations
  - Estimate $\alpha$ by standard error minimization methods (e.g., Galerkin's method)

- **Implicit Systems**
  - $G(X, Y) = 0$
  - Assume a form for $Y$
    - (e.g., series expansion)
    - $Y = h(\xi, \alpha)$
  - Black-box models or complex numerical codes
  - Estimate $\alpha$ using:
    - Collocation methods (e.g., ECM or orthogonal collocation)
    - Regression methods (if collocation methods don't converge)

- Evaluate $Y = h(\xi, \alpha)$

- Express inputs as functions of srvs
  - $X = F(\xi)$ (exact, if possible)
  - $X \equiv F(\xi)$ (for complex cases)
**Emphasis on Black Box Models**

- Complex computational models are sometimes better treated as blackbox model

- Modification of computational models to retrofit uncertainty characterization can be expensive (e.g. Quality Assurance)

- Nested models and modeling systems render black box approach necessary

**Application of the SRSM to Black Box Models**

1. Select a set of srvs and transform inputs in terms of srvs
2. Express outputs as series in srvs with unknown coefficients
3. Generate a set of regression points
4. Estimate the coefficients of output approximation
5. Model
6. Input Distributions
7. Output Distributions
Transformation of Uncertain Inputs

<table>
<thead>
<tr>
<th>Distribution Type</th>
<th>Transformation$^a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uniform $(a, b)$</td>
<td>$a + (b - a)\Phi(\xi)$</td>
</tr>
<tr>
<td>Normal $(\mu, \sigma)$</td>
<td>$\mu + \sigma \xi$</td>
</tr>
<tr>
<td>Lognormal $(\mu, \sigma)$</td>
<td>$\exp(\mu + \sigma \xi)$</td>
</tr>
<tr>
<td>Gamma $(a, b)$</td>
<td>$ab \left( \xi \sqrt{\frac{1}{9a}} + 1 - \frac{1}{9a} \right)^3$</td>
</tr>
<tr>
<td>Exponential $(\lambda)$</td>
<td>$-\frac{1}{\lambda} \log(\Phi(\xi))$</td>
</tr>
<tr>
<td>Weibull $(a)$</td>
<td>$y^{1/a}$</td>
</tr>
<tr>
<td>Extreme Value</td>
<td>$-\log(y)$</td>
</tr>
</tbody>
</table>

$^a \xi \sim \text{Normal}(0, 1), \Phi(x) \sim \text{NormCDF}(x),$ and $y \sim \text{Exponential}(1)$

For empirical distributions specified by a cumulative density function, $F(x) = g(x)$

$$x = g^{-1}(\Phi(\xi))$$
Transformation of Correlated Distributions.

• Simple cases: Dirichlet distribution (functions of independent normal random variables)

• Simple cases: Mixtures of distributions

• Simple cases of jointly distributed random variables (e.g. joint normal random variables)

• Jointly distributed with a covariance matrix $\Sigma$ [Based on Devroye, 1986]
  - correlated variables with mean $\mu_i$ and co-variance matrix $\sigma_i$ (common in risk assessment models)
    * create $\Sigma^*$ via $\Sigma^*_{i,j} = \Sigma_{i,j} / (\sigma_i \sigma_j)$

    * construct $Y$ via $Y_i = (x_i - \mu_i) / \sigma_i$

    * construct $Z = HY$, where $HH^T = \Sigma^*$, and

    * express model inputs as $x_i = \mu_i + \sigma_i z_i$. 
SRSM – Illustration with 2 parameter case for a black-box model

For example, for a model with 2 parameters \((X_1, X_2)\) and 2 outputs \((Y_1, Y_2)\) if \(X_1 = \text{Lognormal}(p_1, q_1)\), and \(X_2 = \text{Gamma}(p_2, q_2)\)

Transformation via srvs:

\[
X_1 = \exp(p_1 + q_1 \xi_1) \\
X_2 = p_2q_2 \left( \xi_2 \sqrt{\frac{1}{9p_2}} + 1 - \frac{1}{9p_2} \right)^3
\]

2nd order approximation:

\[
Y_1 = a_0 + a_1 \xi_1 + a_2 \xi_2 + a_3 (\xi_1^2 - 1) + a_4 (\xi_2^2 - 1) + a_5 \xi_1 \xi_2
\]

\[
Y_2 = b_0 + b_1 \xi_1 + b_2 \xi_2 + b_3 (\xi_1^2 - 1) + b_4 (\xi_2^2 - 1) + b_5 \xi_1 \xi_2
\]

Solution:

\[
\begin{bmatrix}
1 & \xi_{1,1} & \xi_{2,1} & \xi_{1,1}^2 - 1 & \xi_{2,1}^2 - 1 & \xi_{1,1}\xi_{2,1} \\
1 & \xi_{1,2} & \xi_{1,2} & \xi_{2,2}^2 - 1 & \xi_{2,2}^2 - 1 & \xi_{1,2}\xi_{2,2} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
1 & \xi_{1,r} & \xi_{2,r} & \xi_{1,r}^2 - 1 & \xi_{2,r}^2 - 1 & \xi_{1,r}\xi_{2,r}
\end{bmatrix}
\begin{bmatrix}
a_0 \\
a_1 \\
a_2 \\
a_3 \\
a_4 \\
a_5
\end{bmatrix}
= 
\begin{bmatrix}
y(1,1) \\
y(1,2) \\
y(1,3) \\
y(1,r)
\end{bmatrix}
\begin{bmatrix}
y(2,1) \\
y(2,2) \\
y(2,3) \\
y(2,r)
\end{bmatrix}
\]
Utilizing Local Sensitivity Information for Uncertainty Propagation

Techniques for Calculating Derivatives

- analytical methods and symbolic manipulation methods (e.g. using Maple)
- divided difference methods:

\[
\frac{\partial u_i}{\partial k_j} \approx \frac{u_{i,r} - u_{i,l}}{2h} ; \ h \to 0
\]

- requires more model simulations
- truncation errors for large step sizes ("\(\Delta h\)") and roundoff errors for small \(\Delta h\)

Automatic Differentiation [Bischof, 1994]

every function on computer as a sequence of elementary operations

using chain rule: 
\[
\left. \frac{\partial}{\partial t} f(g(t)) \right|_{t=t_0} = \left( \left. \frac{\partial}{\partial s} f(s) \right|_{s=g(t_0)} \right) \left( \left. \frac{\partial}{\partial t} g(t) \right|_{t=t_0} \right)
\]

in a mechanical fashion, derivatives can be computed (up to machine precision)
SRSM-ADIFOR: SRSM coupled with sensitivity analysis methods

- local sensitivity analysis methods provide partial derivatives of the outputs w.r.t. inputs
- when coupled with SRSM, this can reduce required model runs by a factor of $M + 1$ for $M$ inputs

Available Systems for Automatic Differentiation

- ADIC: Automatic Differentiation for C programs [Bischoff, 94]
- ADIFOR: Automatic Differentiation for FORTRAN programs [Bischoff, 94]
- MAD/ADMAT: Automatic Differentiation for Matlab

§Complete, regularly updated list can be found at http://www.autodiff.org
Schematic of Automatic Differentiation using ADIFOR

COMPUTER MODEL
(To Be Analyzed)

Automatic
Derivative
Calculator
(e.g. ADIFOR)

DERIVATIVE CODE

Further Analysis

Compile and Link

Derivative Computing Code
y(1) = 1.0
y(2) = 1.0
do i = 1,n
    if (x(i) > 0.0) then
        y(1) = x(i) * y(1) * y(1)
    else
        y(2) = x(i) * y(2) * y(2)
    endif
endo

(a)

(y(1) = 1.0
y(2) = 1.0
do i = 1,n
    if (x(i) > 0.0) then
        temp = x(i) * y(i)
        y(1) = temp * y(1)
    else
        temp = x(i) * y(2)
        y(2) = temp * y(2)
    endif
endo

(b)

(a) Example code fragment and (b) automatically modified for preparation for derivative calculation (adapted from Bischoff, 1994)
\begin{verbatim}
dy(1) = 0.0
y(1) = 1.0
dy(2) = 0.0
y(2) = 1.0

do i = 1,n
    if (x(i) > 0.0) then
        dtemp = y(1)*dx(i) + x(i)*dy(1)
        temp = x(i) * y(1)
        dy(1) = y(1)*dtemp + temp*dy(1)
        y(1) = temp * y(1)
    else
        dtemp = y(2)*dx(i) + x(i)*dy(2)
        temp = x(i) * y(2)
        dy(2) = y(2)*dtemp + temp*dy(2)
        y(2) = temp * y(2)
    endif
enddo
\end{verbatim}

ADIFOR generated code for calculating derivatives (adapted from Bischoff, 1994)
SRSM-ADIFOR: Formulation

Inputs: \( X_i = f(\xi_1, \xi_2, \ldots, \xi_n), \quad i = 1, \ldots, n \)

Outputs: \[
y = a_0 + \sum_{i_1=1}^{n} a_{i_1} \Gamma_1(\xi_{i_1}) + \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} a_{i_1 i_2} \Gamma_2(\xi_{i_1}, \xi_{i_2}) \\
+ \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \sum_{i_3=1}^{n} a_{i_1 i_2 i_3} \Gamma_3(\xi_{i_1}, \xi_{i_2}, \xi_{i_3}) + \ldots
\]

Derivatives: \[
\frac{\partial y}{\partial \xi_i} = \sum_{i_1=1}^{n} a_{i_1} \frac{\partial \Gamma_1(\xi_{i_1})}{\partial \xi_i} + \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} a_{i_1 i_2} \frac{\partial \Gamma_2(\xi_{i_1}, \xi_{i_2})}{\partial \xi_i} \\
+ \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \sum_{i_3=1}^{n} a_{i_1 i_2 i_3} \frac{\partial \Gamma_3(\xi_{i_1}, \xi_{i_2}, \xi_{i_3})}{\partial \xi_i} + \ldots
\]

\[
\Gamma_p(\xi_{i_1}, \ldots, \xi_{i_p}) = (-1)^p e^{\frac{1}{2} \xi^T \xi} \frac{\partial^p}{\partial \xi_{i_1} \ldots \partial \xi_{i_p}} e^{-\frac{1}{2} \xi^T \xi}
\]
Steps in the Application of SRSM-ADIFOR

1. Input Distributions
   - Select a set of srvs and transform inputs in terms of srvs
   - Generate a set of regression points

2. Output Distributions
   - Estimate the coefficients of output approximation

3. Model
   - ADIFOR application
   - Derivative Code
   - Express outputs as series in srvs with unknown coefficients

4. Additional steps:
   - Generate a set of regression points
Monte Carlo, SRSM and SRSM-ADIFOR steps

1. Generate a set of random inputs
2. Input Distributions
3. Select a set of svrs and transform inputs in terms of svrs
4. Express outputs as series in svrs with unknown coefficients
5. Generate a set of regression points
6. Compute the inputs corresponding to the regression points
7. Original Model
8. Wrap model with svrs as inputs
9. Derivative code (w.r.t. svrs)
10. ADIFOR application
11. Estimate coefficients of output approximation
12. Output distributions

Legend: → (SRSM), →→ (SRSM-ADIFOR), → (Monte Carlo)
SRSM/SRSM-ADIFOR: Case study

Uncertainty Analysis of a Physiologically Based Pharmacokinetic Model

- **model**: Physiologically based pharmacokinetic (PBPK) model for perchloroethylene (PERC). This model is used to describe the uptake and metabolism of PERC in humans.

- **sources of uncertainty**
  - interindividual variability; intraindividual variability
  - human PBPK parameters are estimated by inter-species scale-up of animal PBPK model parameters
  - *in vitro* data are used to estimate parameters (*in vivo* data not available)

- **input random variables**: PBPK model parameters
  - 7 uncorrelated parameters (partition coefficients, blood flows and metabolic constants)
  - 4 correlated parameters (compartmental mass fractions)

- **output random variables**: (surrogates for biologically effective dose)
  - cumulative amount of PERC metabolized in liver (CML)
  - area under arterial concentration (AUCA)
Schematic of the PERC PBPK model [Roy, 1996]

\[
V_j \frac{dc_j}{dt} = Q_j \left( c_{\text{arterial}} - \frac{c_j}{P_j/\text{blood}} \right) - R_j \quad ; \quad \text{where} \quad R_j = \frac{V_{\text{max},j} c_{v,j}}{K_m_j + c_{v,j}}
\]

\[
c_{\text{arterial}} = \frac{Q_{\text{cardiac}} c_{\text{venous}} + Q_{\text{alveolar}} C_{\text{air(inhaled)}}}{Q_{\text{cardiac}} + Q_{\text{alveolar}}/P_{\text{blood/air}}} \quad ; \quad \text{where} \quad c_{\text{venous}} = \frac{1}{Q_{\text{cardiac}}} \sum_{j=1}^{n} Q_j c_j
\]
### Parameters in the PERC PBPK Model (Farrar et al. 1989)

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
<th>Distribution</th>
<th>Preferred Value (PV)</th>
<th>UF</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>BW</strong></td>
<td>body weight [Kg]</td>
<td></td>
<td>70</td>
<td></td>
</tr>
<tr>
<td><strong>Partition Coefficients</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$P_{\text{blood/air}}$</td>
<td>blood:air partition coefficient</td>
<td>LN</td>
<td>12</td>
<td>1.7</td>
</tr>
<tr>
<td>$P_{\text{fat/blood}}$</td>
<td>fat:blood partition coefficient</td>
<td>LN</td>
<td>102</td>
<td>2.15</td>
</tr>
<tr>
<td>$P_{\text{sp/blood}}$</td>
<td>slowly perfused tissue:blood partition coefficient</td>
<td>LN</td>
<td>2.66</td>
<td>11.0</td>
</tr>
<tr>
<td>$P_{\text{rp/blood}}$</td>
<td>rapidly perfused tissue:blood partition coefficient</td>
<td>LN</td>
<td>5.05</td>
<td>5.69</td>
</tr>
<tr>
<td>$P_{\text{liv/blood}}$</td>
<td>liver:blood partition coefficient</td>
<td>LN</td>
<td>5.05</td>
<td>9.37</td>
</tr>
<tr>
<td><strong>Blood Flows</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$Q_c$</td>
<td>cardiac output [liters/hr]</td>
<td></td>
<td>348</td>
<td>1.12</td>
</tr>
<tr>
<td>$Q_p$</td>
<td>alveolar ventilation [liters/hr]</td>
<td></td>
<td>288</td>
<td>1.50</td>
</tr>
<tr>
<td>$Q_{\text{fat}}$</td>
<td>blood flow to fat [liters/hr]</td>
<td></td>
<td>17.4</td>
<td>1.09</td>
</tr>
<tr>
<td>$Q_{\text{sp}}$</td>
<td>blood flow to slowly perfused tissue [liters/hr]</td>
<td></td>
<td>87.0</td>
<td>1.04</td>
</tr>
<tr>
<td>$Q_{\text{rp}}$</td>
<td>blood flow to rapidly perfused tissue [liters/hr]</td>
<td></td>
<td>153</td>
<td>1.25</td>
</tr>
<tr>
<td>$Q_{\text{liv}}$</td>
<td>blood flow to liver [liters/hr]</td>
<td></td>
<td>90.6</td>
<td>1.35</td>
</tr>
<tr>
<td><strong>Compartment Mass Fractions</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$V_{\text{fat}}$</td>
<td>mass of fat compartment [Kg]</td>
<td>Dirichlet</td>
<td>23.0%</td>
<td>1.09</td>
</tr>
<tr>
<td>$V_{\text{sp}}$</td>
<td>mass of slowly perfused tissue compartment [Kg]</td>
<td>Dirichlet</td>
<td>62.0%</td>
<td>1.04</td>
</tr>
<tr>
<td>$V_{\text{rp}}$</td>
<td>mass of rapidly perfused tissue compartment [Kg]</td>
<td>Dirichlet</td>
<td>5.0%</td>
<td>1.25</td>
</tr>
<tr>
<td>$V_{\text{liv}}$</td>
<td>mass of liver compartment [Kg]</td>
<td>Dirichlet</td>
<td>2.6%</td>
<td>1.35</td>
</tr>
<tr>
<td><strong>Metabolic Constants</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$K_m$</td>
<td>Michaelis-Menten constant for metabolism [mg/liter]</td>
<td>LN</td>
<td>1.47</td>
<td>12.3</td>
</tr>
<tr>
<td>$V_{\text{max}}$</td>
<td>maximum rate of metabolism [mg/hr]</td>
<td>LN</td>
<td>0.33</td>
<td>2.84</td>
</tr>
</tbody>
</table>

LN - Lognormal; UF - Uncertainty factor. The range of PV/UF to PV*UF contains 95% of the population.
Probability density estimates of dose surrogates

Cumulative Amount of PERC metabolized (CML)

Evaluation of SRSM

- Monte Carlo - 100,000 runs
- SRSM - 3rd Order (600 runs)
- LHS - 1,000 runs
- SRSM-ADIFOR - 3rd Order (70 runs)
Probability density estimates of dose surrogates

Area Under the Arterial Concentration Curve (AUCA)

Evaluation of SRSM

- Monte Carlo - 100,000 runs
- SRSM - 3rd Order (600 runs)
- LHS - 1,000 runs
- SRSM-ADIFOR - 3rd Order (70 runs)
Uncertainty Analysis of a Reactive Plume Model (RPM-IV)

- Model: Regulatory model for estimating pollutant concentrations downwind of strong point sources; uses Carbon Bond IV chemical mechanism (95 reactions among 35 chemical species)


- Model Inputs: Emission estimates and meteorological variables

- Model Outputs: Primary and secondary pollutant concentration estimates

- Uncertainty: Emission estimates and chemical composition of emissions

- Case Study: Measured Data at Marathon Refinery, Robinson IL, June-July, 1977 [Sexton, 1983]

Uncertain Parameters

- Amounts of emissions of VOCs (volatile organics) and NO\textsubscript{x} (nitrous oxides): Truncated normal distributions (40% standard deviation from nominal values)

- mole fractions of VOCs and NO\textsubscript{x}: Dirichlet distributions (satisfies the condition that the random mole fractions should always add to unity)

- total number of uncertain parameters: 9 (5 fractions of VOCs, 2 fractions of NO\textsubscript{x} and amounts of VOCs and NO\textsubscript{x})
Ozone concentration at the plume centerline

- Monte Carlo --- 10000 runs
- SRSM --- 600 runs
- SRSM−ADIFOR --- 80 runs
Use of SRSM in Bayesian Parameter Estimation

Advantages of Bayesian Parameter Estimation

- Assimilates prior information on parameter values and information contained in the data.
  - prior information on parameters are specified by probability distribution functions (pdfs).

- Convenient for mechanistic, biological and environmental process models.
  - assignment of prior distributions is facilitated for models having physically meaningful parameters.

- Parameters are regarded as random variables, and are characterized by probability density functions (pdfs).
  - in contrast to classical parameter estimation, parameters do not have a single “true,” but unknown value
Terminology

• Bayesian inference on model parameters ($\theta$) given observed data ($y$) involves the estimation of either:
  
  – the posterior distribution of the parameters conditioned on the data, $P(\theta|y)$, or
  
  – the joint posterior distribution, $P(y, \theta) = P(\theta|y)/P(y)$.

• The above distributions are determined using Bayes' Theorem:

$$P(\theta|y) \propto P(y|\theta)P(\theta),$$

where $P(\theta)$ is the prior distribution, and $P(y|\theta)$ is the sampling distribution. The sampling distribution is known as the likelihood function when $y$ is fixed, and $\theta$ is unknown.

Markov Chain Monte Carlo (MCMC) Simulation

• A class of Bayesian inference methods that generate numerical approximations of the joint posterior distribution of parameters, $P(y, \theta)$

• Several algorithms have been proposed for MCMC simulation:
  
  – Metropolis; Metropolis-Hastings; Gibbs Sampling
Metropolis Algorithm

1. Obtain $\theta^{(0)}$, an initial realization of the parameter vector

2. Obtain $\theta^{(t)}$, the $t^{th}$ realization of the parameter vector by:
   
   (a) sampling from a proposal distribution, $J_t(\theta^*|\theta^{(t-1)})$, to obtain a candidate parameter vector, $\theta^*$
   
   (b) calculating
   
   $$r = \frac{P(\theta^*|y)}{P(\theta^{(t)}|y)}$$
   
   and

   (c) setting
   
   $$\theta^{(t)} = \begin{cases} \theta^* & \text{with probability } \min(r,1), \text{ and} \\ \theta^{(t-1)} & \text{otherwise} \end{cases}$$
Gibbs Sampling

- A computationally efficient algorithm for generating a Markov Chain that converges to a joint posterior distribution of the parameters, conditioned on the data

- First generates a point in parameter space $\theta^{(1)} = (\theta_1^{(1)}, \theta_2^{(1)}, \ldots, \theta_d^{(1)})$ from an arbitrary initial point $\theta^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_d^{(0)})$, by:

  \[
  \begin{align*}
  \theta^{(1)} &\sim P(\theta_1^{(1)}|\theta_2^{(1)}, \theta_3^{(0)}, \ldots, \theta_d^{(0)}, y) \\
  \theta_2^{(1)} &\sim P(\theta_2^{(1)}|\theta_1^{(1)}, \theta_3^{(0)}, \ldots, \theta_d^{(0)}, y) \\
  \vdots &\vdots \\
  \theta_d^{(1)} &\sim P(\theta_d^{(1)}|\theta_1^{(1)}, \theta_2^{(1)}, \ldots, \theta_{d-1}^{(1)}, y)
  \end{align*}
  \]

- This process is repeated to generate the sequence $\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(N)}$, which tends to $P(\theta|y)$ as $N \rightarrow \infty$. 
Utilizing SRSM in Bayesian Inference

• Need
  – The computational model needs to evaluated at each step of the MCMC for chain
  – This could become computationally expensive (e.g. 1 minute per model run for 10,000 runs with 6 MCMC chains results in about 1 week per chain).
  – Many environmental models are much more computationally intensive.
  – An approximate model (e.g. SRSM) can potentially be used in conjunction with MCMC

• Approach
  – Obtain SRSM approximation
  – Use the approximate model for MCMC application
  – Evaluate the convergence and posterior distributions†

• Applications
  – A preliminary application to the FACT (Flow and Contaminant Transport) model, a groundwater model, using SRSM with MCMC
  – Current effort is on detailed evaluation of utilizing the SRSM with MCMC through a simple PBPK model

†Avoid “computationally efficient,” but wrong answers
Uncertainty Reduction (Parameter Estimation) in Ground Water Modeling

- Model: Subsurface Flow and Contaminant Transport (FACT) code, a transient 2-D, finite element code designed to simulate isothermal groundwater flow, moisture movement, and solute transport in variably saturated and fully saturated subsurface porous media.

- Model inputs: 3-D conductivity fields and 2-D recharge rate field, which are inherently uncertain.

- Substantial prior knowledge exists for the inputs of interest.

- Computationally quite demanding (1/2 hr per run on a Sun Fire 280R with two 750 MHz UltraSPARC III CPUs and 4GB memory).

- Model outputs: hydraulic head values (for which actual data are available from field measurements) and stream baseflow rates.
FACT model schematic

- Rainfall
- Evapotranspiration
- Runoff
- Recharge
- Water Table Surface
- Seepage Face
- Fourmile Branch
- UTRA - “upper zone”
- UTRA - “lower zone”
- Gordon confining unit
- Gordon Aquifer
- Meyers Branch Confining System

Layers of the model:
- Top layer: Water Table Surface
- Middle layer 1: UTRA - “upper zone”
- Middle layer 2: UTRA - “lower zone”
- Bottom layer: Gordon Aquifer

Key processes:
- Rainfall: Water entering the system
- Evapotranspiration: Water lost to the atmosphere
- Runoff: Water moving overland to streams
- Recharge: Water infiltrating into the soil

Groundwater flow directions:
- Water moves from upper to lower zones
- Seepage faces at boundaries between layers

NASA/NIA ECS Lecture Series
Using SRSM approximation as a reduced model
Probability densities estimated using MCMC with the SRSM reduced model

- Log10 GCU Kv
- Log10 LAZ Kh
- Log10 TCZ Kv
- Log10 UAZ Kh

Units of
Kv, Kh : ft/d
Recharge : in/yr
Discussion

Improving the parameterization and accuracy of the SRSM

- Polynomial expansion may become unstable as order of approximation increases

- Many coefficients (especially for higher order terms), may be estimated to be small, but non-zero.

- Instead of approximating the output directly, there should be a mechanism to estimate multiple transformations of outputs w.r.t. the polynomial expansion and automatically select the “best” approximation.

- Bayesian techniques for model averaging can be useful in this context.
Use of SRSM (and other approximate models) for Bayesian parameter estimation and uncertainty reduction

- Adequacy of the model approximation (e.g. for performing an MCMC simulation)

- Adequacy of the Bayesian parameter estimates
  - Utility of this approach depends greatly on the comparative cost of finding a “good estimate” versus evaluating whether an estimate is “good”.