Efficient Techniques for Sensitivity and Uncertainty Analysis of Multiscale Air Quality Models

Presented at the
4th Annual CMAS Models-3 User's Conference
Friday Center, UNC-Chapel Hill
September 26-28, 2005

by
Sastry Isukapalli*, Sheng-Wei Wang*, Nilesh Lahoti*
Alper Unal**, and Panos Georgopoulos*

*Computational Chemodynamics Laboratory (CCL) (http://www.ccl.rutgers.edu)
Environmental and Occupational Health Sciences Institute (EOHSI)
(a joint Institute of UMDNJ – Robert Wood Johnson Medical School and Rutgers University)
170 Frelinghuysen Road, Piscataway New Jersey 08854

**MACTEC, Trenton, NJ

Motivation for Efficient Uncertainty Characterization 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

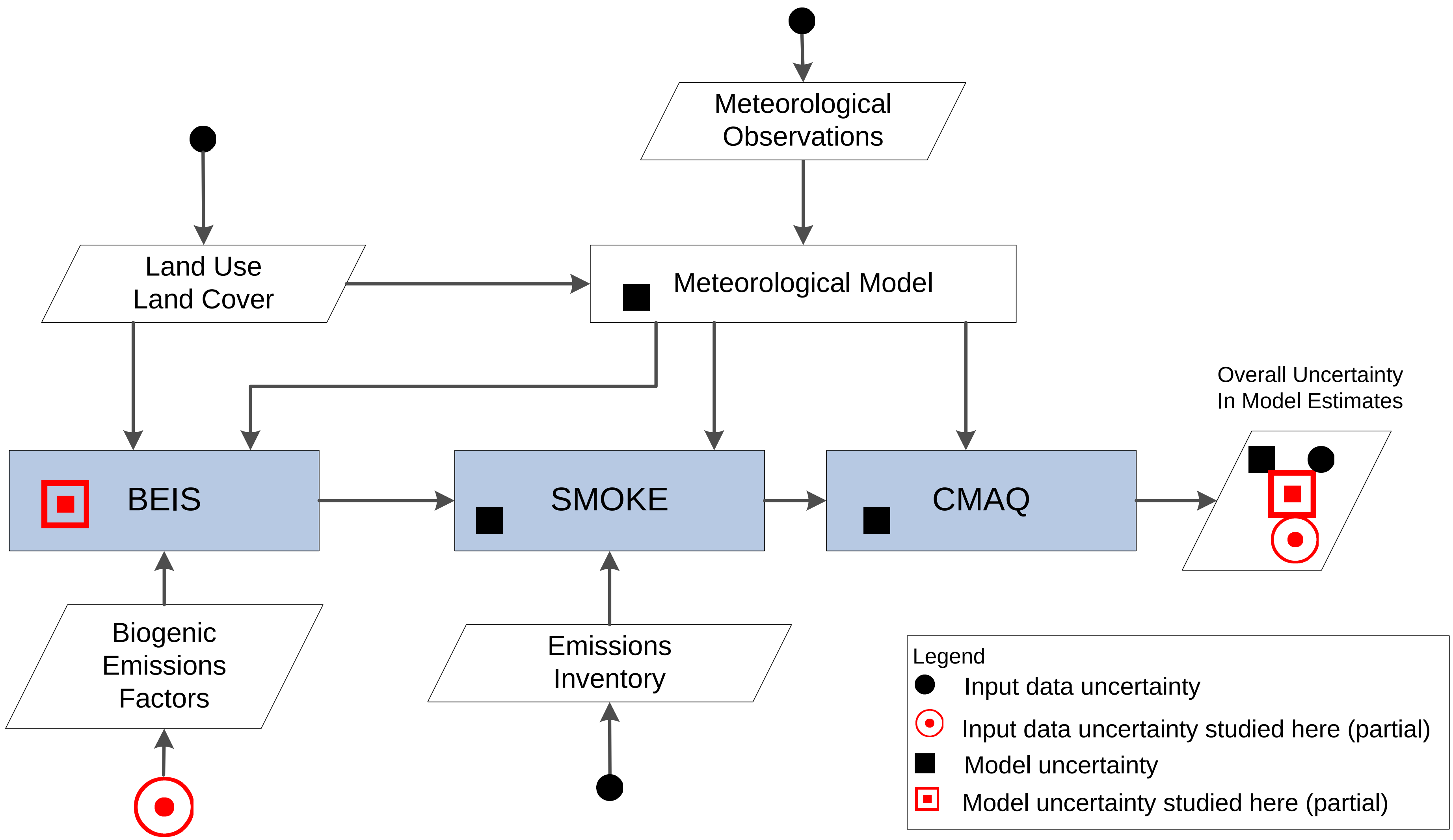
However,

- coupling of multiple computational models results in a "nested system" of uncertainties and variabilities
- each modeling step can require significant computational resources

Uncertainties in air quality modeling include:

- natural uncertainty
- *input/parameter uncertainty*
- model uncertainty
- evaluation data uncertainty

Propagation of Uncertainties in Air Quality Modeling using CMAQ



CMAQ: Community Multi-scale Air Quality modeling system
BEIS: Biogenics Emission Inventory System
SMOKE: Sparse Matrix Operator Kernel Emissions modeling system

Traditional Methods Applied to CMAQ

- Monte Carlo and Latin Hypercube Sampling (LHS)
  - easy to use and apply in a black-box manner
  - computationally demanding (require large number of model simulations)
  - they require even more resources for obtaining "sensitivity information"
  - *past studies with air quality modeling have used very few Monte Carlo runs for studying uncertainties*
    - of the order of 20 - 200 simulations involving 10 - 100 parameters
- Direct Decoupled Method (DDM)
  - provides accurate local sensitivity information
  - significant memory requirements as number of parameters increase
  - large number of simulations for global sensitivity/uncertainty analysis
  - requires re-coding major portions of a model (not a black-box tool)

Computationally Efficient, Alternative Techniques

- Stochastic Finite Element Method [Ghanem and Spanos, 1992]
- Deterministic Equivalent Modeling Method (DEMM) [Tatang, 1995]
- Stochastic Response Surface Method (SRSM) [Isukapalli et al., 1998; Isukapalli, 1999]
- High Dimensional Model Representations (HDMR) [Rabitz et al., 1999; Wang et al., 2003]

- DEMM, SRSM, and HDMR can be applied to computational models in a black-box manner
- SRSM and HDMR have been applied to environmental and biological models

Stochastic Response Surface Method (SRSM)

- Based on approach of response surface methods
- Transform uncertain inputs
  - model parameters and input variables expressed as functions of a set of "standard random variables" (*srv*s
  - typically *iid* unit normal random variables, N(0,1)
- Assume functional form for outputs
  - expressed as a hermite polynomials of the *srv*s with unknown coefficients (polynomial chaos expansion)
- Run original model at a set of sample points
  - points depend upon the number of uncertain parameters
- Estimate coefficients of approximation
  - by regression on model calculated model responses
- Use coefficients to assess output uncertainties
  - polynomial chaos expansion represents the uncertainty in model responses
  - Monte Carlo simulation on polynomial functions gives estimate of uncertainty
  - Coefficient encompass a quantitative measure of relative contribution of individual input uncertainties

High Dimensional Model Representations (HDMR)

- HDMR: a systematic method for model reduction
  - can be used to develop a "fast equivalent" model based on the analysis of input/output relations of complex "primary" model
  - Options
    - Cut HDMR
    - Random Sampling HDMR
- reduce the number of required model runs by "optimizing" sampling
- replace the original model with a "fast equivalent" one so that the computational requirements are reduced
- HDMR can be an useful tool in either (1) or (2) framework

Uncertainties in Biogenic Emissions

- Biogenic emissions have a significant impact on local ozone levels
- These estimates are laden with major uncertainties due to:
  - variability in land use and land cover
  - variability in emission rates (variability in sunlight, temperature, etc.)
  - uncertainties introduced when the emision rates are parameterized
- Uncertainties reported to be about a factor of 2
- Isoprene is a major component of biogenic emissions

BEIS Uncertain Inputs/Parameters Considered
7 uncertain parameters, all independent

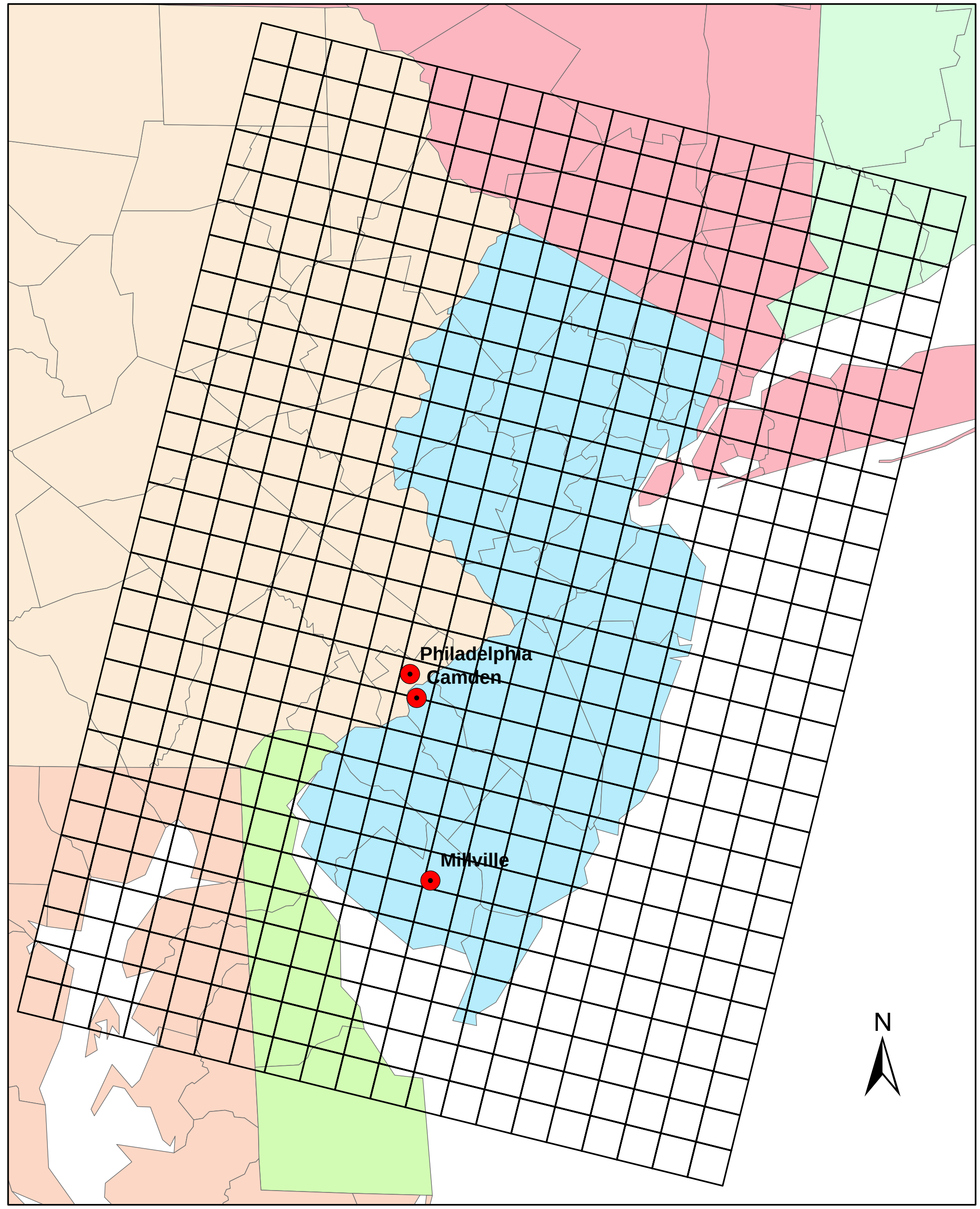
Param	Description	Distribution	Mean/GM**	Std. Dev.	SRSM Transformation
E_s	Emissions flux*	Normal	-NA-	25%	$1 + 0.25\,\xi_1$
LAI	Leaf area index	Normal	-NA-	12.5%	$1 + 0.125\,\xi_2$
		Empiricial Coefficients			
α	light correction	Lognormal	0.0027	0.0015	$\exp(\log(0.0027) + 0.4702\,\xi_3)$
CL_1	light correction	Normal	1.06	0.2	$1.06 + 0.2\,\xi_4$
CT_1	temperature correction	Lognormal	90,000	20,000	$\exp(\log(90000) + 0.2147\,\xi_5)$
CT_2	temperature correction	Lognormal	230,000	20,000	$\exp(\log(230000) + 0.0865\,\xi_6)$
T_M	temperature correction	Normal	314	3	$314 + 3\,\xi_7$

Notes:
* Emissions flux is species specific
** For Lognormal distribution, the Geometric Mean (GM) is shown here. A value of "-NA-" implies that a multiplication factor is shown here
Truncation at 2.5 standard deviations are assumed

Source for parameter distributions: Hanna et al. (2005), Monte Carlo estimation of uncertainties in BEIS3 emission outputs and their effects on uncertainties in chemical transport model predictions, J. Geophys. Res., 110, D01302.

Study Domain

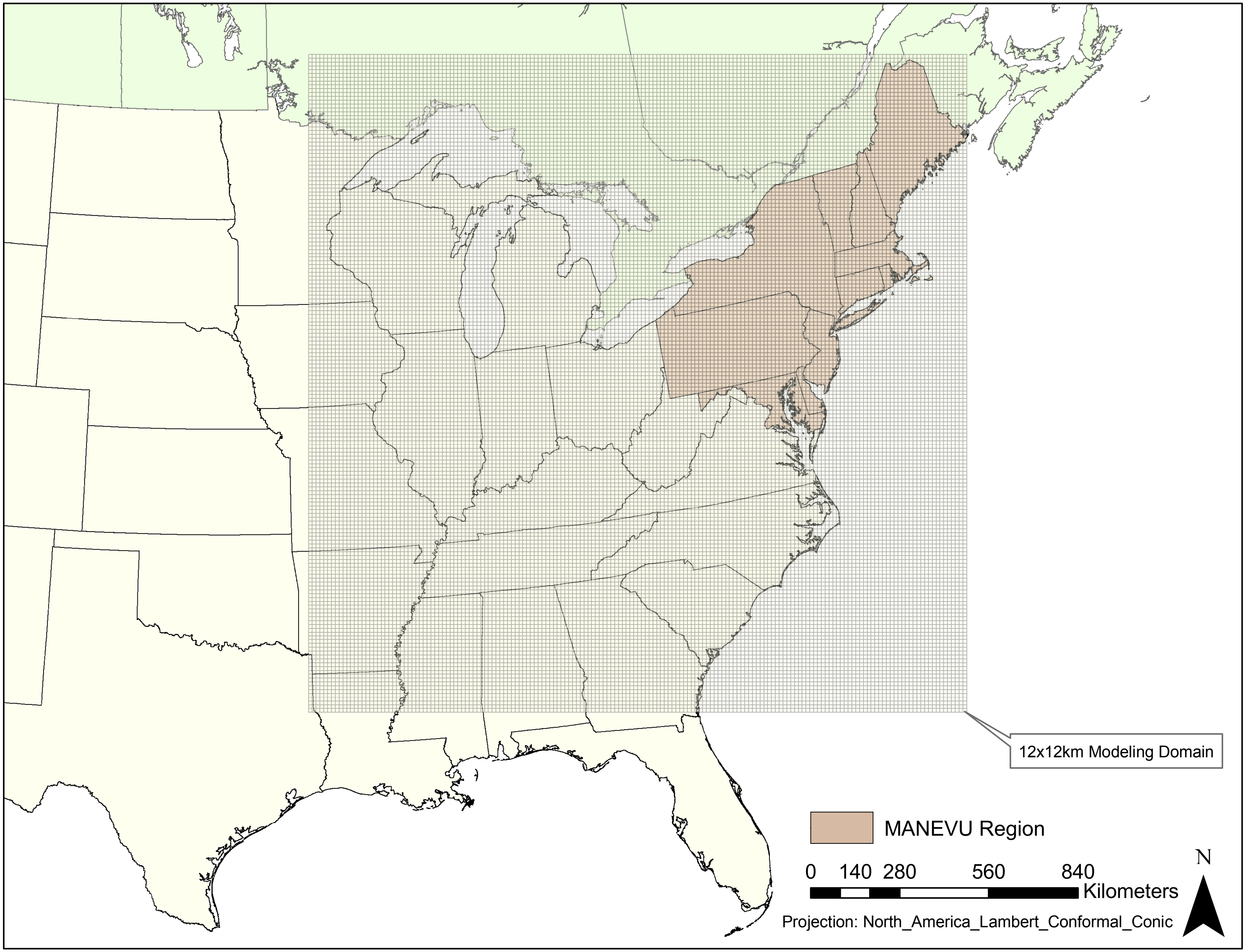
- Covers the entire NJ and urban Philadelphia, PA region
- 12 km x 12 km resolution
- 20 cells in the east-west direction
- 28 cells in the north-south direction
- Grid Projection: Lambert Conformal α = 33, β = 45, γ = -97, Origin Latitude = -97, Origin Longitude = 40



Projection: NAD 1983 UTM Zone 18N

Simulation Details

- Simulation Period: August 10 - 14, 2002 (UTC)
- Meteorological outputs from MM5 (Mesoscale Meteorological Model, Ver. 5)
  - MM5 model results obtained from the NJDEP simulations
- Emission Inventory obtained from the NJDEP simulations
- Initial and boundary conditions obtained from a "parent simulation"
  - A bench mark CMAQ simulation for the Eastern United States
  - Simulation Period: August 6 - 16, 2002
  - Note: Biogenic emissions in the study region are at the nominal values during the "parent simulation"
- Number of uncertain parameters: 7 (all independent)
- Output metrics: ozone levels
  - Maximum predicted hourly average ozone concentration over the entire episode and domain
  - Maximum predicted eight-hour running average of ozone concentration over the entire episode and domain
  - Ozone profiles in two grid cells covering Philadelphia, PA, and Millville, NJ



Domain used for "parent simulation"

Number of simulation steps for SRSM approximation

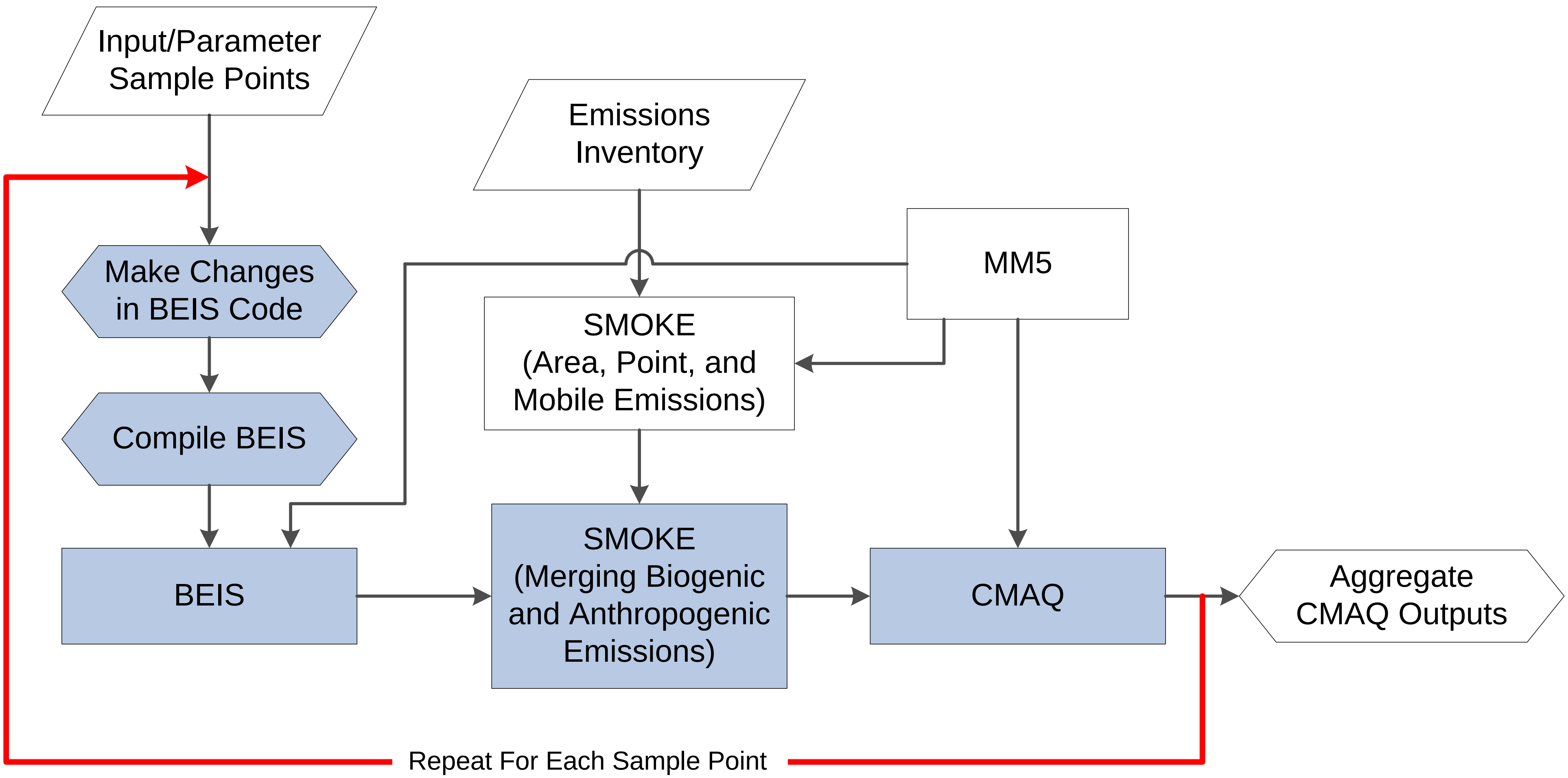
- Second order approximation
- Number of SRSM coefficients to estimate for 7 input variables ($n = 7$): $1 + 2n + n(n-1)/2 = 36$
- Number of steps used for regression: twice the number of coefficients = 72

Number of simulation steps for HDMR approximation

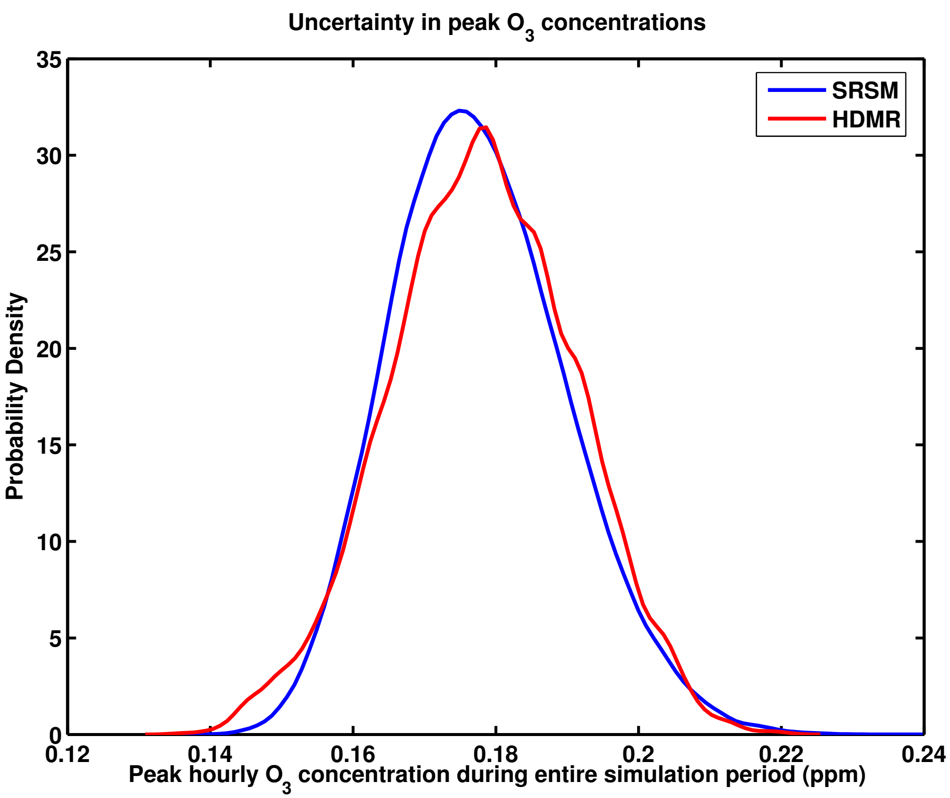
- First order approximation
- 9 cuts ($c = 9$) in each dimension
- Cut percentiles: from 2 to 98: [2 14 26 38 50 62 74 86 98]
- Total number of simulations for 7 input variables ($n = 7$): $1 + n(c-1) = 57$

Total: 72 simulations for SRSM and 57 simulations for HDMR

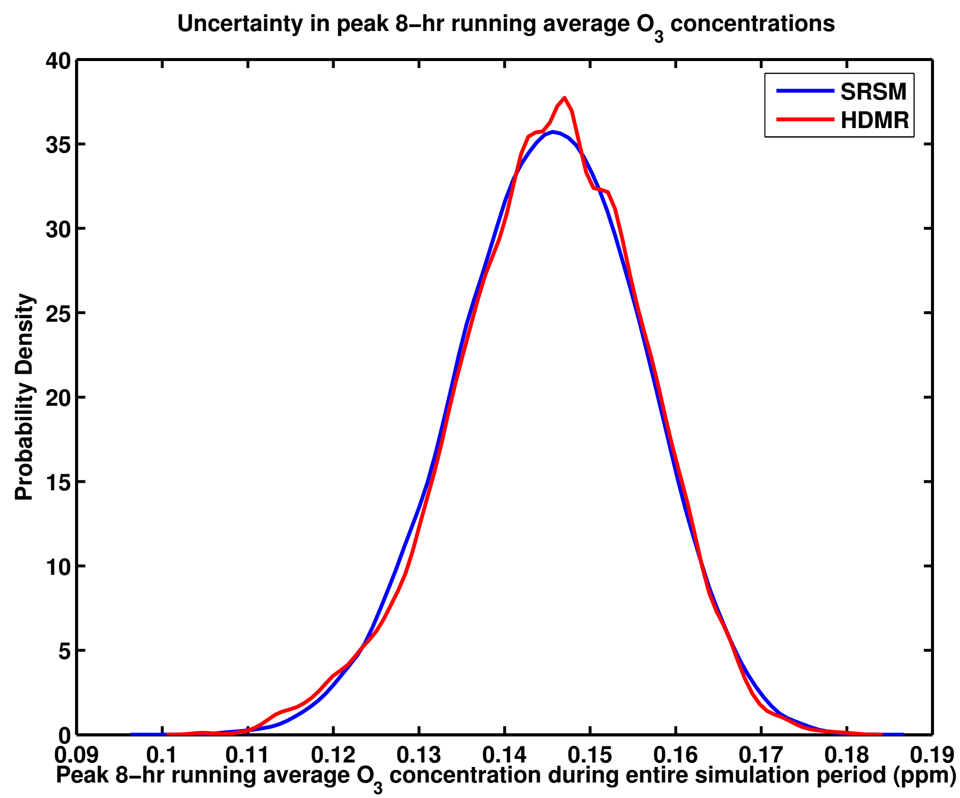
Uncertainty Propagation Steps Used in this Study



Uncertainty estimates for peak hourly and peak 8-hr running average O_3 concentrations



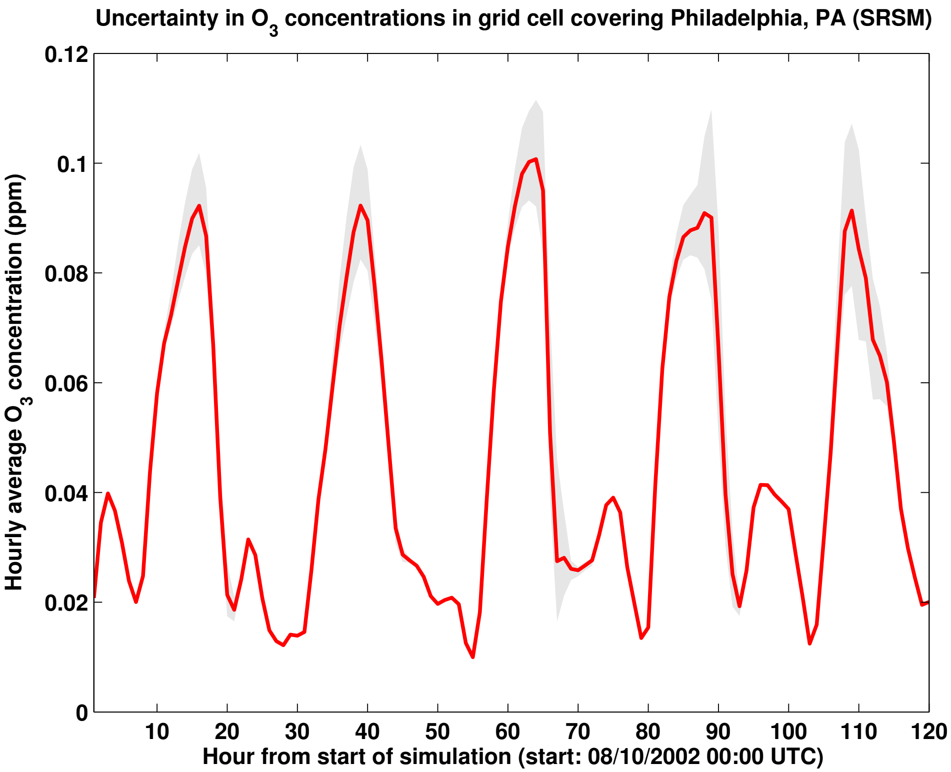
Peak hourly average



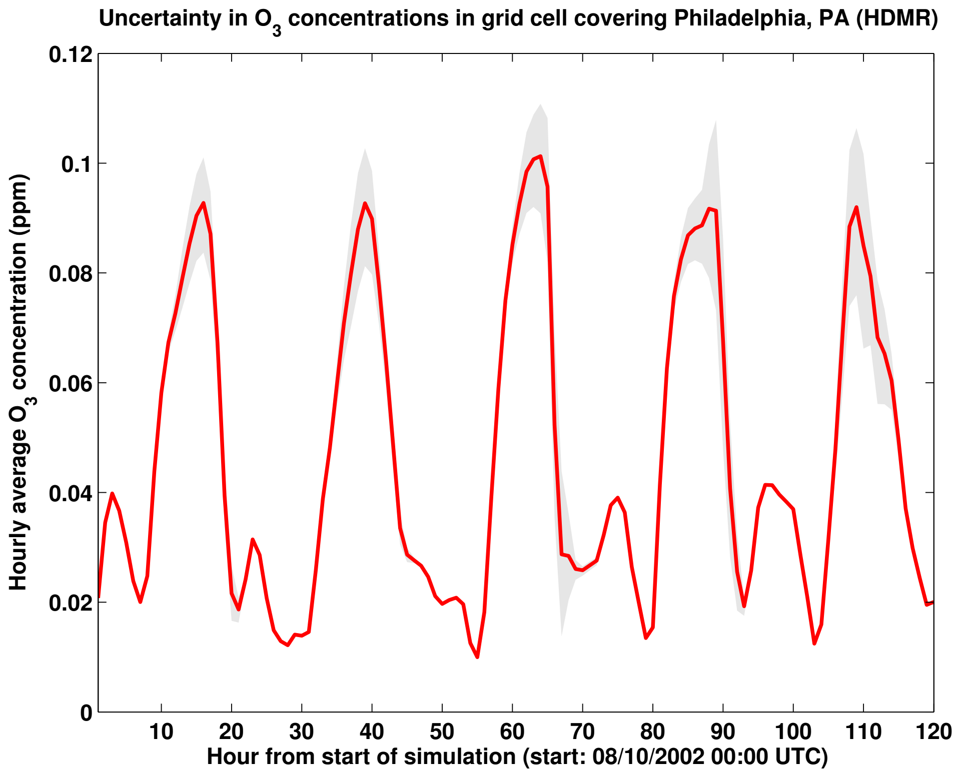
Peak 8-hr running average

Uncertainty estimates for hourly O_3 concentrations at Philadelphia, PA

(median and 95th confidence intervals are shown)



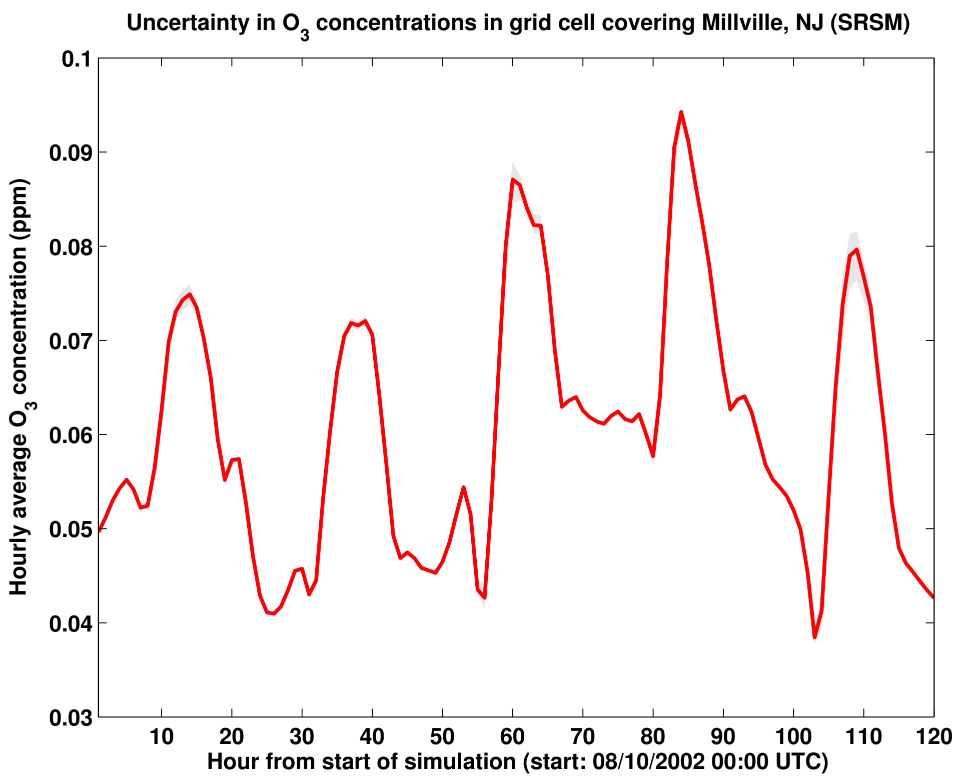
SRSM



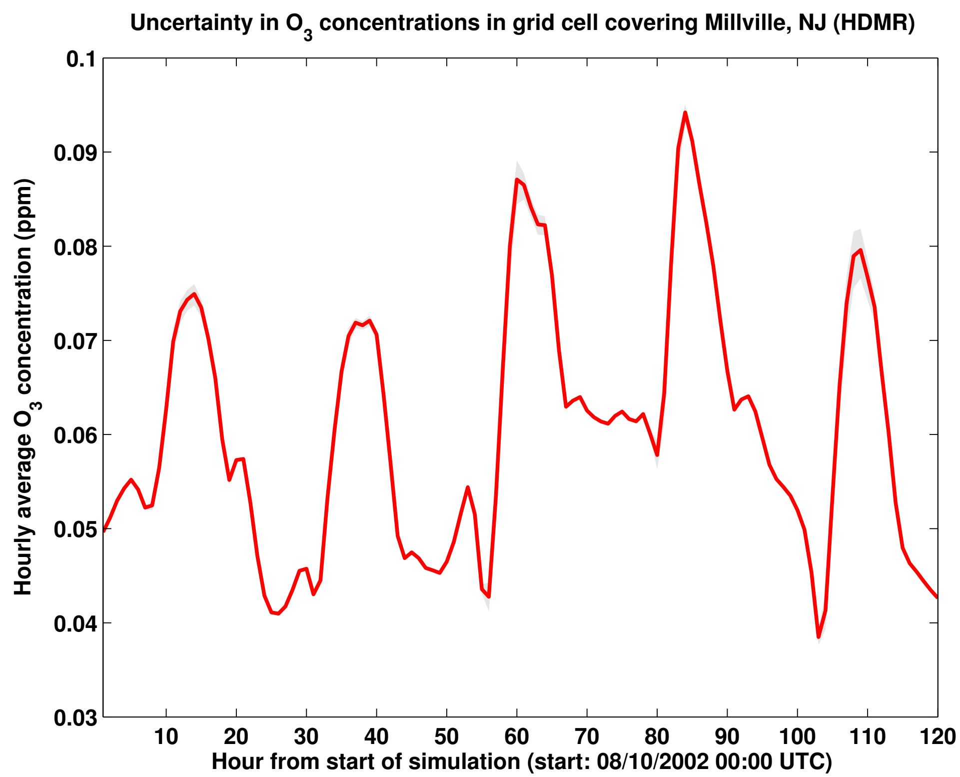
HDMR

Uncertainty estimates for hourly O_3 concentrations at Millville, NJ

(median and 95th confidence intervals are shown)



SRSM



HDMR

Discusssion

- SRSM and HDMR provide similar estimates of uncertainties in O_3 concentrations due to uncertainties in a subset of biogenic emissions
- Using either SRSM or HDMR, uncertainties in different types of outputs and output metrics can be estimated through a small number of simulations
- The response surfaces from HDMR and SRSM can be readily used to estimate individual contributions of input uncertainties to outputs
- SRSM and HDMR can be used as a replacement of the ambiguous use of very small number of Monte Carlo simulations
- The effect of input uncertainty on "overall peak" is higher than the uncertainties at the two specific locations considered

Acknowledgments

- US Environmental Protection Agency (USEPA): Funding for the Center for Exposure and Risk Modeling (CERM) at EOHSI
- New Jersey Department of Environmental Protection (NJDEP): Base funding for the Ozone Research Center at EOHSI, and for the year 2002 benchmark simulation data

Important Disclaimer

- Viewpoints expressed here do not necessarily reflect the views of USEPA, NJDEP, or their contractors
- The purpose of the case study used here is solely for illustrating the applicability of the SRSM and HDMR methods to complex models such as CMAQ

Supporting Slides

Basic Forumalation of the SRSM

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

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

$$\Gamma_p(\xi_{i_1}, \ldots, \xi_{i_p}) = (-1)^p e^{\frac{1}{2}\boldsymbol{\xi}^T\boldsymbol{\xi}} \frac{\partial^p}{\partial\xi_{i_1}\ldots\partial\xi_{i_p}} e^{-\frac{1}{2}\boldsymbol{\xi}^T\boldsymbol{\xi}} \qquad \text{(Hermite Polynomials)}$$

Transformation of Uncertain Inputs

Distribution Type	Transformation[a]
Uniform (a, b)	$a + (b - a)\Phi(\xi)$
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)	$-\frac{1}{\lambda}\log(\Phi(\xi))$
Weibull (a)	$y^{1/a}$
Extreme Value	$-\log(y)$

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

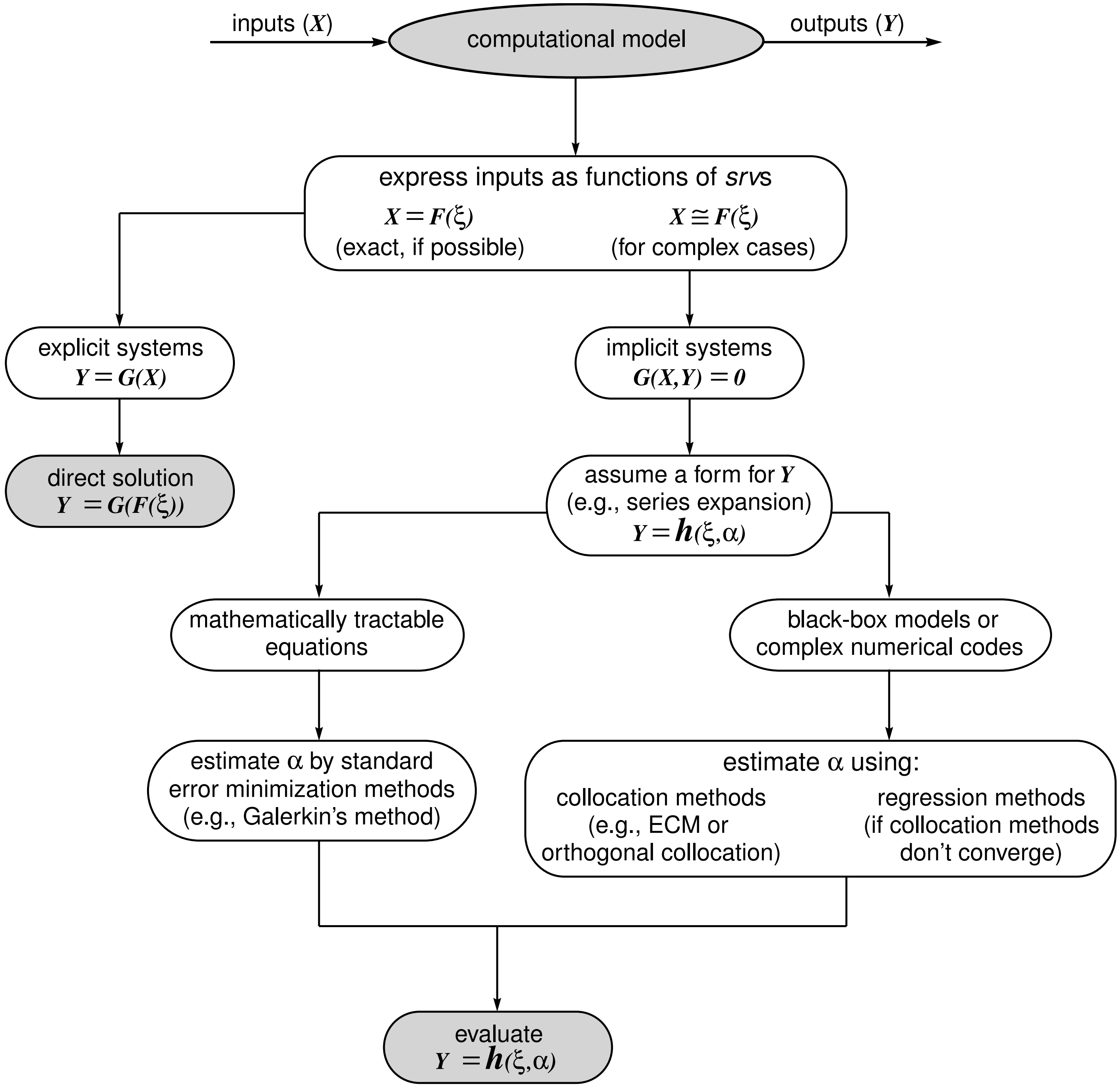
For empirical distributions specified by a cumulative density function, $F_{\boldsymbol{x}}(x) = g(x)$

$$\boldsymbol{x} = g^{-1}(\Phi(\xi))$$

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 $\mathbf{\Sigma}$ [Based on Devroye, 1986]
  - correlated variables with mean μ_i and co-variance matrix σ_i (common in risk assessment models)
    - create $\mathbf{\Sigma}^*$ via $\Sigma^*_{i,j} = \Sigma_{i,j}/(\sigma_i \sigma_j)$
    - construct $\boldsymbol{Y}$ via $Y_i = (\boldsymbol{x}_i - \mu_i)/\sigma_i$
    - construct $\boldsymbol{Z} = \boldsymbol{H}\boldsymbol{Y}$, where $\boldsymbol{H}\boldsymbol{H}^T = \mathbf{\Sigma}^*$, and
    - express model inputs as $\boldsymbol{x}_i = \mu_i + \sigma_i \boldsymbol{z}_i$.

Application of the SRSM:



HDMR Approach

$$\mathbf{I} \longrightarrow \boxed{\textbf{SYSTEM}} \longrightarrow \mathbf{O}$$

- System (a mathematical model):
  - **Input I:** $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$
  - **Output O:** $g(\mathbf{x}) = g(x_1, x_2, \ldots, x_n)$
- Exponential difficulty in traditional sampling:
  - sampling $x_i \rightarrow x_i^1, x_i^2, \ldots, x_i^s, \qquad (i = 1, 2, \ldots, n)$
  - exponential effort $\sim s^n$
- The HDMR method expresses a model output as a expansion of correlated functions:

$$\begin{aligned} g(\mathbf{x}) &= f_0 + \sum_{i=1}^{n} f_i(x_i) + \sum_{1 \leq i < j \leq n} f_{ij}(x_i, x_j) + \cdots \\ &+ f_{12\ldots n}(x_1, x_2, \ldots, x_n) \end{aligned}$$

HDMR Rationale

- *the outputs of most physical systems do not draw on high order cooperativity amongst the input variables*
- Cut-HDMR:
  - $f_0 = g(\mathbf{a})$
  - $f_i(x_i) = g(x_i, \mathbf{a}^i) - f_0$
  - $f_{ij}(x_i, x_j) = g(x_i, x_j, \mathbf{a}^{ij}) - f_i(x_i) - f_j(x_j) - f_0$
  - where $\mathbf{a} = \{a_1, a_2, \ldots, a_n\}$ is a chosen reference (cut) point in the desired domain Ω of $\mathbf{x}$ and
  - $\{x_i, \mathbf{a}^i\} = \{a_1, \ldots, a_{i-1}, x_i, a_{i+1}, \ldots, a_n\}$
  - $\{x_i, x_j, \mathbf{a}^{ij}\} = \{a_1, \ldots, a_{i-1}, x_i, a_{i+1}, \ldots, a_{j-1}, x_j, a_{j+1}, \ldots, a_n\}$