Uncertainty Quantification in the Community Land Model

K. Sargsyan¹, C. Safta¹, B. Debusschere¹, H. Najm¹, D. Ricciuto², P. Thornton²

¹Sandia National Laboratories
Livermore, CA

²Oak Ridge National Laboratory
Oak Ridge, TN

CESM Annual Workshop
Breckenridge, CO
June 18-21, 2012
Outline

1. Introduction
2. Bayesian Inference
3. PC Surrogate
4. Sparse surrogate & sensitivity analysis
5. Calibration plans
The Case for Uncertainty Quantification

UQ is needed in:

- Assessment of confidence in computational predictions
- Validation of physical models
- Design optimization
- Use of computational predictions for decision-support
- Assimilation of observational data and model construction
- Multiscale and multiphysics model coupling
## Overview of UQ Methods

### Estimation of model/parametric uncertainty
- Expert opinion, data collection
- Regression analysis, fitting, parameter estimation
- Bayesian inference of uncertain models/parameters

### Forward propagation of uncertainty in models
- Local sensitivity analysis (SA) and error propagation
- Fuzzy logic; Evidence theory — interval math
- Probabilistic framework — Global SA / stochastic UQ
  - Random sampling, statistical methods
  - Galerkin methods
    - Polynomial Chaos (PC) — intrusive/non-intrusive
  - Collocation, interpolants, regression, fitting ... PC/other
- Random sampling, statistical methods
- Galerkin methods
Plan for CLM UQ

### Single-site computations
- Establish set of relevant uncertain inputs
- Estimate uncertainties in inputs/parameters
  - Build an efficient surrogate model
  - Forward UQ + Global sensitivity analysis
- Calibrate CLM/submodels with available data (Bayes)
- Estimate uncertainties in model output quantities of interest

### Multiple-site/global computations
- Dimensionality reduction in multi-sites and their inputs
- Calibrate multi-site models
- Forward UQ
Bayes formula for Parameter Inference

Data Model

\[ y = f(\lambda) + \epsilon \]

Bayes Formula:

\[ p(\lambda|y) = \frac{p(y|\lambda) p(\lambda)}{p(y)} \]

- Prior: knowledge of \( \lambda \) prior to data
- Likelihood: forward model and measurement noise
- Posterior: combines information from prior and data
- Evidence: normalizing constant for present context
Exploring the Posterior in a Computational Setting

- Given any sample $\lambda$, the un-normalized posterior probability can be easily computed

$$p(\lambda|y) \propto p(y|\lambda)p(\lambda)$$

- Explore posterior w/ Markov Chain Monte Carlo (MCMC)
  - Metropolis-Hastings algorithm:
    - Random walk with proposal PDF & rejection rules
    - Computationally intensive, $\mathcal{O}(10^5)$ samples
    - Each sample: evaluation of the forward model
  - Surrogate models

- Evaluate moments/marginals from the MCMC statistics
Surrogate Models for Bayesian Inference

- Need an inexpensive response surface for
  - Observables of interest $y$
  - as functions of parameters of interest $x$

- Gaussian Process (GP) surrogate
  - GP goes through all data points with probability 1.0
  - Uncertainty between the points

- Fit a convenient polynomial to $y = f(x)$
  - over the range of uncertainty in $x$

- Employ a number of samples $(x_i, y_i)$
- Fit with interpolants, regression, ... global/local
- With uncertain $x$:
  - Construct Polynomial Chaos response surface

Marzouk et al. 2007; Marzouk & Najm, 2009
**Polynomial Chaos (PC) Surrogate in high-D**

$n$-dim, $p$-th order Polynomial

\[ u = f(\eta) = \sum_{k=0}^{K-1} c_k \Psi_k(\eta) \quad \eta \in \mathbb{R}^n; \]

High \((n, p) \Rightarrow K \gg 1\) terms

\(N\) samples \((\eta_i, u_i)\) to fit surface

**Projection:**
- MC/QMC; Sparse Quadrature

**Regression:**
- Require regularization when \(N < K\)
- Compressive sensing: Least-squares L1-regularization
- Bayesian Compressive Sensing (BCS) \((\text{Ji} \ 2008; \ Babacan \ 2010)\)
- Iterative BCS (iBCS) – adaptive sparsity \((\text{Sargsyan} \ 2012)\)
Construction of CLM sparse PC surrogate

- $n \approx 80$ uncertain input parameters – specified ranges
  - Special handling of algebraic constraints
  - Random parameter samples on 80d hypercube
  - Use iBCS to construct sparse PC surrogate

- Challenge with smooth global surrogate
  - Large regions of given input space result in failed vegetation
  - TotVegC response has discontinuous derivative

![Graphs showing data and polynomial fit for input parameter #1]
Clustered Global CLM Surrogate

- Cluster training samples $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2\}$
  - $\mathcal{D}_1 = \{\eta | \text{TotVegC} < \epsilon\}$
- Construct global surrogate based on each data subset
- Clustered surrogate

$$f_s(\eta) = \begin{cases} f_1(\eta) & \text{for } \eta \in \mathcal{D}_1 \\ f_2(\eta) & \text{for } \eta \in \mathcal{D}_2 \end{cases}$$

- Classification step to discover cluster membership for $\eta$
  - $\Rightarrow$ Random Decision Forests classifier (Ho, 1995; Breiman, 2001)
    - collection of decision trees
    - result is mode of the results from individual trees
    - Software: www.alglib.net/dataanalysis/decisionforest.php
CLM Runs – Spinup

- Single-site runs on Jaguar @ ORNL
- Initialization: 3-stage 1000 yr spinup
Given spun-up state:

- 10K random parameter samples on 80-d hypercube
- 1000-yr runs. Each run $\sim 10$ hr on 1-CPU
Sobol Sensitivity Indices – clustered iBCS surrogate

- Indices measure relative fractional contribution to total variance due to different terms in $f_s(\eta)$
- A measure of global importance over the uncertainty range
- Main effects: $\Psi(\eta_i)$ terms; 2-way coupling: $\Psi(\eta_i, \eta_j)$
- $\sim 1000$ out of $2M$ terms retained in $4^{th}$-order surface

Leaf Area Index

Total Vegetation Carbon
Global Sensitivity Analysis for TotVegC

- Dominant main and 2-way coupled effects
- Significant coupling terms
Calibration Effort – in progress

- Examining noise structure in dynamical observables
- Use multiple observables
- Both steady and unsteady observables
- Begin with a minimal set of important parameters
  - Identifiability
This work was supported by:

- US Dept of Energy (DOE), Office of Science, Biological and Environmental Research (BER)

Sandia National Laboratories is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under contract DE-AC04-94-AL85000.