Dimensionality Reduction and Global Sensitivity Analysis for the Community Land Model

C. Safta\textsuperscript{1}, K. Sargsyan\textsuperscript{1}, D. Ricciuto\textsuperscript{2}, B. Debusschere\textsuperscript{1}, H.N. Najm\textsuperscript{1}, P. Thornton\textsuperscript{2}

\textsuperscript{1}Sandia National Laboratories
Livermore, CA, USA

\textsuperscript{2}Oak Ridge National Laboratory
Oak Ridge, TN, USA

The Winter CESM Uncertainty Quantification and Analysis Interest Group Meeting
NCAR Mesa Lab, Boulder CO
February 20-21, 2013
Acknowledgement

This work was supported by the US Department of Energy, Office of Science, under the project “Climate Science for a Sustainable Energy Future”, funded by the Biological and Environmental Research (BER) program.

This is a continuation of a presentation by Khachik Sargsyan in the UQA meeting (“Surrogate construction via Bayesian compressive sensing for the Community Land Model”)

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy’s National Nuclear Security Administration under contract DE-AC04-94AL85000.
Outline

1. Challenges
2. Surrogate Models
   - Polynomial Chaos Expansions
   - Constrained Parameter Space
3. Iterative Bayesian Compressive Sensing (iBCS)
   - Methodology
   - BCS+Classification
4. Community Land Model results
5. Summary
UQ Challenges in Climate Models

- Computationally expensive model simulations
- High-dimensional input parameter space
  - Physical constraints and dependencies for some input parameters
  - Uncertainties in the input parameters are not known
- Non-linear dependence of output quantities of interest on inputs
Community Land Model

http://www.cesm.ucar.edu/models/clm/

- Nested computational grid hierarchy
- Represents spatial heterogeneity of the land surface
- A single-site, 1000-yr simulation takes \( \sim 10 \) hrs on 1 CPU
- Involves \( \sim 70 \) input parameters
Generate initial conditions → several spin-up stages

- total soil organic matter carbon \( [gC/m^2] \) \((TOTSOMC)\)
Community Land Model - Typical Setup (2)

- Sample the parameter space

- Left frame: Contour plot of time-averaged $\text{TOTSOMC}$ values for a range of $(r_{\text{mort}}, f_{\text{root_leaf}})$ values
- Right frame: Time evolution of $\text{TOTSOMC}$ for select $(r_{\text{mort}}, f_{\text{root_leaf}})$ values
Surrogate Models

What are surrogate models?

- Input parameter vector $\lambda$
- Computationally expensive model $f(\cdot)$ (e.g. CLM)
- Given a set of training model runs, $(\lambda_i, f(\lambda_i))_{i=1}^{N}$, a surrogate $f_s(\cdot) \approx f(\cdot)$ is a model that is cheap to evaluate and appropriately represents the underlying detailed, expensive model over a specified range of input parameters.

Why do we need surrogate models?

- Global sensitivity analysis
- Input parameter inference
- Optimization
- Forward uncertainty propagation
To build a surrogate representation for input-output relationship, Polynomial Chaos (PC) spectral expansions are used; see Ghanem and Spanos (1991).

- Interprets input parameters as random variables
- Allows propagation of input parameter uncertainties to outputs of interest
- Serves as a computationally inexpensive surrogate for calibration or optimization
Polynomial Chaos Representations

Input parameters are represented via their cumulative distribution function (CDF) $F(\cdot)$, such that, with $\eta_i \sim \text{Uniform}[-1, 1]$, we have:

$$\lambda_i = F^{-1}_{\lambda_i} \left( \frac{\eta_i + 1}{2} \right), \quad \text{for } i = 1, 2, \ldots, d.$$

If input parameters are uniform $\lambda_i \sim \text{Uniform}[a_i, b_i]$, then

$$\lambda_i = \frac{a_i + b_i}{2} + \frac{b_i - a_i}{2} \eta_i.$$

Output is represented with respect to Legendre polynomials

$$f(\lambda(\eta)) \approx y(\eta) \equiv \sum_{k=0}^{K} c_k \Psi_k(\eta).$$
Map Constrained Parameters to Unconstrained Spaces

- Given a vector of random variables $\lambda = (\lambda_1, \ldots, \lambda_d')$ with known joint cumulative distribution function (CDF) $F(\lambda_1, \ldots, \lambda_d')$
- Use Rosenblatt transformation (RT) to obtain a map $\eta = R(\lambda)$ to a set of $\eta_i$’s that are independent uniform random variables on $[-1, 1]$.

\[
\begin{align*}
\lambda_{18} &< \lambda_{22}, \\
\lambda_{30} + \lambda_{31} + \lambda_{32} &= 1, \\
\lambda_{33} + \lambda_{34} + \lambda_{35} &= 1.
\end{align*}
\]
Bayesian Inference of Polynomial Chaos modes

Bayesian inference of PC modes allows surrogate construction with uncertainties associated with limited sampling

- Bayes formula
  \[ p(c|D) \propto L_D(c)p(c) \]
  relates the prior distribution \( p(c) \) of PC modes to the posterior \( p(c|D) \), where the data \( D \) is the set of all training runs \( D = (\lambda_i, f(\lambda_i))_{i=1}^N \).

- The likelihood accounts for the discrepancy between the simulation data and the surrogate model (Sargsyan et al 2011),
  \[ L_D(c) \propto \exp \left( - \sum_{i=1}^N \frac{(f(\lambda_i) - yc(\eta_i))^2}{2\sigma^2} \right) \]
The number of polynomial basis terms grows fast; a $p$-th order, $d$-dimensional basis has a total of $(p + d)!/(p!d!)$ terms.

Dimensionality reduction by using hierarchical priors.

$$p(c | s_k^2) \propto \prod_{k=0}^{K} \exp \left( -\frac{c_k^2}{2s_k^2} \right) \quad p(s_k^2 | \alpha) = \frac{\alpha}{2} \exp \left( -\frac{\alpha s_k^2}{2} \right)$$

The parameter $\alpha$ can be further modeled hierarchically, or fixed.

The parameters $(\sigma^2, s_0^2, \ldots, s_K^2)$ are fixed by evidence maximization, and bases corresponding to small $s_i^2$ are discarded (Ji et al. 2008, Babacan et al., 2010).

**Iterative BCS:** We implement an iterative procedure that allows increasing the order for the relevant basis terms while maintaining the dimensionality reduction (Sargsyan et al. 2011, 2012).
Climate Land Model - Single site mode for Niwot Ridge

- $N = 10,000$ training runs based on uniformly LHS distributed parameter values.
- Outputs: steady-state, 10-year averages of 7 quantities

### iBCS for one observable

<table>
<thead>
<tr>
<th>Name</th>
<th>Units</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOTVEGC</td>
<td>gC/m$^2$</td>
<td>Total vegetation carbon</td>
</tr>
<tr>
<td>TOTSOMC</td>
<td>gC/m$^2$</td>
<td>Total soil carbon</td>
</tr>
<tr>
<td>GPP</td>
<td>gC/m$^2$/s</td>
<td>Gross primary production</td>
</tr>
<tr>
<td>ERR</td>
<td>W/m$^2$</td>
<td>Energy conservation error</td>
</tr>
<tr>
<td>TLAI</td>
<td>none</td>
<td>Total leaf area index</td>
</tr>
<tr>
<td>EFLX_LH_TOT</td>
<td>W/m$^2$</td>
<td>Total latent heat flux</td>
</tr>
<tr>
<td>FSH</td>
<td>W/m$^2$</td>
<td>Sensible heat flux</td>
</tr>
</tbody>
</table>
Classify Parameter Space

- Large regions of the original quasi-hypercube parameter space lead to simulations with failed vegetation.

- Partition the space using a classification algorithm
  - Classification using Random Decision Forests implemented in the AlgLib software library (http://www.alglib.net)
  - the result is the mode of the results from individual decision trees

- Calibration using 9K samples/Validation using 1K samples
- Shift accuracy from “failed vegetation” plateau to “active vegetation” regions
- Apply the iBCS algorithm on “active vegetation” results
Classification+iBCS

- Clustering/classification-based piecewise Polynomial Chaos construction to accommodate non-smooth transition between dead and live vegetation regions

- Classification errors are approximately 10-15%

- Posterior predictive distribution of the surrogate model output covers the spread of simulation data
Climate Land Model - Global Sensitivity Analysis

- Ranking of the most influential input parameters for each output of interest

\[ S_i = \frac{\sum_{k \in I} c_k^2 \| \Psi_k \|^2}{\sum_{k > 0} c_k^2 \| \Psi_k \|^2} \]

<table>
<thead>
<tr>
<th>rank</th>
<th>TOTVEGC</th>
<th>TOTSOMC</th>
<th>GPP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>r_mort</td>
<td>q10_mr</td>
<td>leafcn</td>
</tr>
<tr>
<td>2</td>
<td>q10_mr</td>
<td>leafcn</td>
<td>k_s4</td>
</tr>
<tr>
<td>3</td>
<td>froot_leaf</td>
<td>froot_leaf</td>
<td>froot_leaf</td>
</tr>
<tr>
<td>4</td>
<td>br_mr</td>
<td>br_mr</td>
<td>flnr</td>
</tr>
<tr>
<td>5</td>
<td>q10_hr</td>
<td>flnr</td>
<td>q10_mr</td>
</tr>
<tr>
<td>6</td>
<td>leafcn</td>
<td>dnp</td>
<td>q10_hr</td>
</tr>
<tr>
<td>7</td>
<td>k_s4</td>
<td>q10_hr</td>
<td>dnp</td>
</tr>
<tr>
<td>8</td>
<td>stem_leaf</td>
<td>leaf_long</td>
<td>rf_s3s4</td>
</tr>
<tr>
<td>9</td>
<td>flnr</td>
<td>k_s4</td>
<td>leaf_long</td>
</tr>
<tr>
<td>10</td>
<td>dnp</td>
<td>frootcn</td>
<td>br_mr</td>
</tr>
</tbody>
</table>
Climate Land Model - Global Sensitivity Analysis

- Most influential input parameter couplings for each output - energy contained in each parameter pair
- Results below correspond to Leaf Area Index (LAI)

\[ S_{ij} = \frac{\sum_{k \in \Pi_{ij}} c_k^2 \| \Psi_k \|^2}{\sum_{k > 0} c_k^2 \| \Psi_k \|^2} \]

- Blue discs sizes are proportional to \( S_i \)
- Thickness of green lines is proportional to \( S_{ij} \)
Most influential input parameter couplings for each output - energy contained in each parameter pair

TOTVEGC

GPP
Sensitivity indices used to discard unimportant parameters

Combine analysis for several outputs of interest, \{TOTVEGC, LAI, ER, GPP\}, to arrive to a reduced input parameter space.
Summary

Sensitivity analysis for complex, expensive, climate models is enabled by cheap surrogate models

- Polynomial Chaos surrogate model is constructed using Bayesian techniques
- Constrained/dependent input parameters are mapped to an unconstrained input set via Rosenblatt transformation
- High-dimensionality is tackled by iterative Bayesian compressive sensing algorithm
- Classification for efficient domain decomposition to relieve the non-linear effects

Future plans include running CLM ensembles on lower-dimensional parameter spaces.
- Goal is to increase predictive fidelity of the CLM surrogate, for reliable parameter calibration.