SAND2012-2679C

Efficient Surrogate Construction for High-Dimensional Climate Models

C. Safta[1], K. Sargsyan[1], D. Ricciuto[2], R.D. Berry[1]
B.Debusschere[1],H. Najm[1],P. Thornton[2]

[1]Sandia National Laboratories
Livermore, CA, USA

[2]Oak Ridge National Laboratory
Oak Ridge, TN, USA

SIAM Conference on Uncertainty Quantification
Raleigh, North Carolina
April 4, 2012

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.

- See related SIAM UQ presentations on algorithm development by K. Sargsyan (MS23) and challenges related to climate models by D. Ricciuto (MS59)

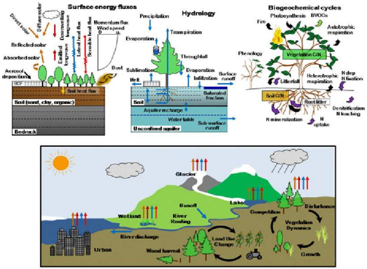
Sandia National Laboratories is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under contract DE-AC04-94-AL85000.

Outline

UQ Challenges in Climate Models

- Computationally expensive model simulations
- High-dimensional input parameter space
- Physical constraints and dependencies for some input parameters
- 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 ~ 10 hrs on 1 CPU
- Involves ~ 80 input parameters

Surrogate Models

What do we need surrogate models for ?

- Global sensitivity analysis
- Input parameter inference
- Optimization
- Forward uncertainty propagation

What are surrogate models ?

- Input parameter vector $\boldsymbol{\lambda}$
- Computationally expensive model $f(\cdot)$ (e.g. climate models)
- Given a set of *training* model runs, $(\boldsymbol{\lambda}_i, f(\boldsymbol{\lambda}_i))_{i=1}^{N}$, a *surrogate* $f_s(\cdot) \approx f(\cdot)$ is a model that is cheap to evaluate

Polynomial Chaos Representations

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 $\eta_i \sim \mathrm{Uniform}[-1,1]$, we have:

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

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

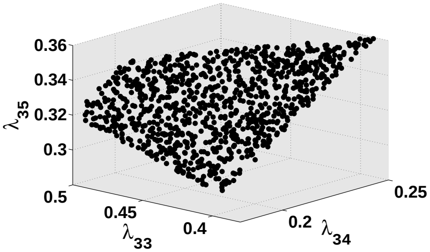
$$\lambda_i = \frac{a_i + b_i}{2} + \frac{b_i - a_i}{2}\,\eta_i.$$

Output is represented with respect to Legendre polynomials

$$f(\boldsymbol{\lambda}(\boldsymbol{\eta})) \approx y_{\mathbf{c}}(\boldsymbol{\eta}) \equiv \sum_{k=0}^{K} c_k \Psi_k(\boldsymbol{\eta}).$$

Map Constrained Parameters to Unconstrained Spaces

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



$$\begin{aligned} \lambda_{18} &< \lambda_{22}, \\ \lambda_{30} + \lambda_{31} + \lambda_{32} &= 1, \\ \lambda_{33} + \lambda_{34} + \lambda_{35} &= 1. \end{aligned}$$

Bayesian Inference of PC modes

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

- Bayes formula

$$p(\mathbf{c}|D) \propto L_{\mathcal{D}}(\mathbf{c})p(\mathbf{c})$$

  relates the prior distribution $p(\mathbf{c})$ of PC modes to the posterior $p(\mathbf{c}|\mathcal{D})$, where the data $\mathcal{D}$ is the set of all training runs $\mathcal{D} = (\boldsymbol{\lambda}_i, f(\boldsymbol{\lambda}_i))_{i=1}^{N}$.
- The likelihood accounts for the discrepancy between the simulation data and the surrogate model (Sargsyan *et al* 2011),

$$L_{\mathcal{D}}(\mathbf{c}) \propto \exp\left(-\sum_{i=1}^{N} \frac{(f(\boldsymbol{\lambda}_i) - y_{\mathbf{c}}(\boldsymbol{\eta}_i))^2}{2\sigma^2}\right)$$

Iterative Bayesian Compressive Sensing (iBCS)

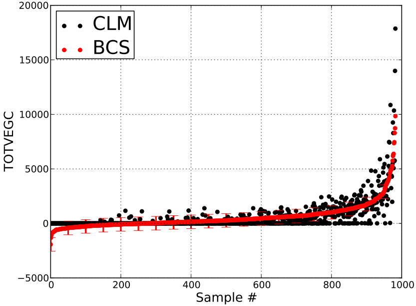
- The number of polynomial basis terms grows fast; a p-th order, d-dimensional basis has a total of $(p+d)!/(p!d!)$ terms.
- Dimensionality reduction by using Gaussian *sparsity* priors.

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

  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).
- *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

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



Name	Units	Description
TOTVEGC	gC/m^2	Total vegetation carbon
TOTSOMC	gC/m^2	Total soil carbon
GPP	$gC/m^2/s$	Gross primary production
ERR	W/m^2	Energy conservation error
TLAI	none	Total leaf area index
EFLX_LH_TOT	W/m^2	Total latent heat flux
FSH	W/m^2	Sensible heat flux

Climate Land Model - 1st order BCS

- Ranking of the most important input parameters for each output

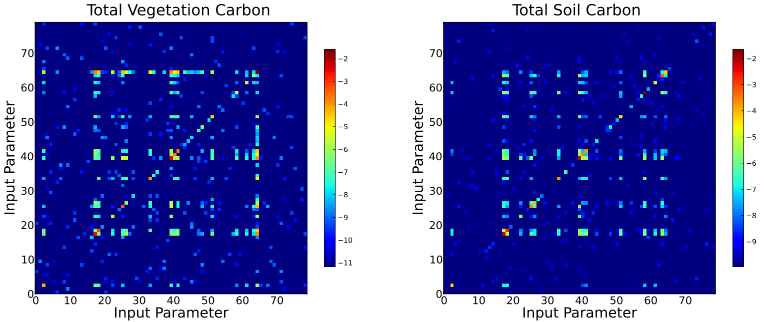
$$S_i = \frac{\sum_{k\in\mathbb{I}_i} c_k^2 ||\Psi_k||^2}{\sum_{k>0} c_k^2 ||\Psi_k||^2}$$

rank	TOTVEGC	TOTSOMC	GPP
1	r_mort	q10_mr	leafcn
2	q10_mr	leafcn	k_s4
3	froot_leaf	froot_leaf	froot_leaf
4	br_mr	br_mr	flnr
5	q10_hr	fflnr	q10_mr
6	leafcn	dnp	q10_hr
7	k_s4	q10_hr	dnp
8	stem_leaf	leaf_long	rf_s3s4
9	flnr	k_s4	leaf_long
10	dnp	frootcn	br_mr

Climate Land Model - 2^{nd} order BCS

- Most influential input parameter couplings for each output - enery contained in each parameter pair

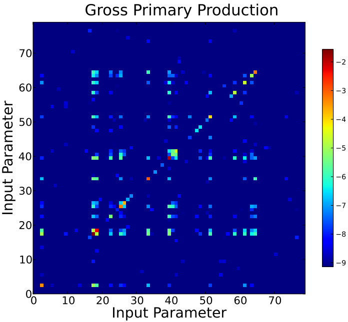
$$S_{ij} = \frac{\sum_{k\in\mathbb{I}_{ij}} c_k^2 ||\Psi_k||^2}{\sum_{k>0} c_k^2 ||\Psi_k||^2}$$

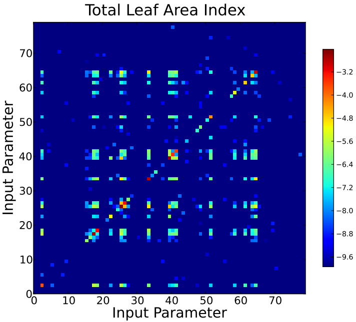


Climate Land Model - 2^{nd} order BCS

- Most influential input parameter couplings for each output - enery contained in each parameter pair

$$S_{ij} = \frac{\sum_{k \in \mathbb{I}_{ij}} c_k^2 ||\Psi_k||^2}{\sum_{k>0} c_k^2 ||\Psi_k||^2}$$



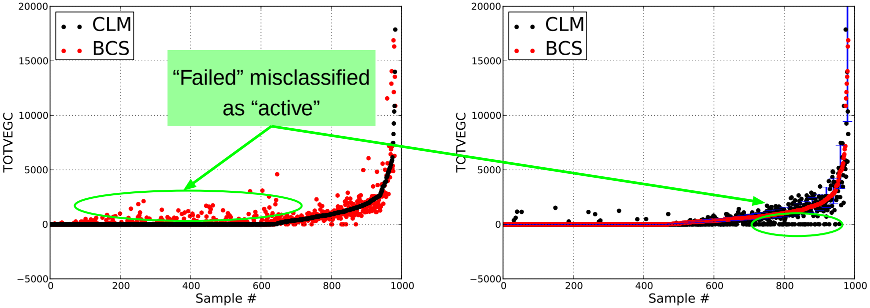


Classify Parameter Space

- Large regions of the original quasi-hypercube parameter space lead to simulations with failed vegetation.
- Partition the space using a classification algorithm
    - Cons: Classification will introduce errors
    - Pros: iBCS algorithm will avoid the “failed vegetation” plateau. Will only use the “active” simulations.
    - Will the “active” parameter sets form a continuous region ?

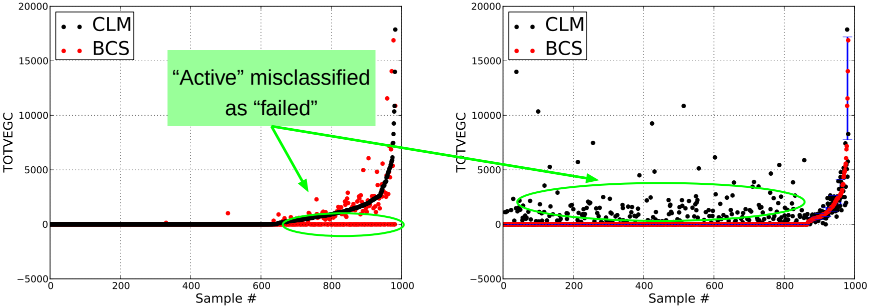
Classification+iBCS

- Classification using Random Decision Forests
  - Calibrate 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



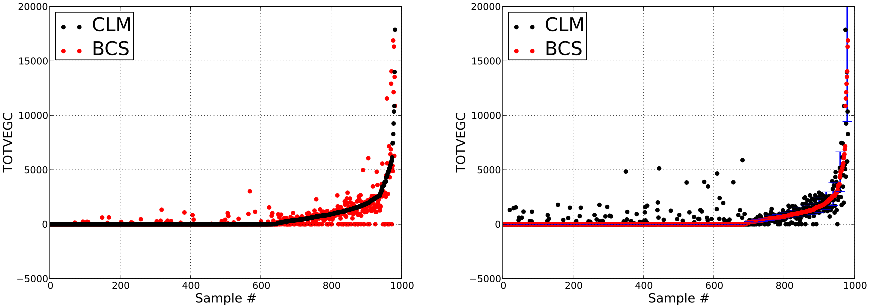
Classification+iBCS

- Classification using Random Decision Forests
  - Calibrate 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

- Classification using Random Decision Forests
    - Calibrate 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



Adaptive Sparse Quadrature - Future Work

- Improve the iBCS surrogate using Galerkin projection $\rightarrow$ efficient techniques to avoid or at least delay the curse of simensionality
  - For example, an 80D/700-term surrogate employs terms of the form λ_i^4, $\lambda_i^3\lambda_j$, $\ldots$.
  - An adaptive set of sampling points require about 3200 additional simulations.
- How do we position these sample points in the "active vegetation" region to actually improve the BCS surrogate ? (80D domain mapping)

(see recent talk by Patrick Conrad/Youssef Marzouk, and pre-print by Paul Constantine)

Adaptive Sparse Quadrature - Future Work

- Improve the iBCS surrogate using Galerkin projection $\rightarrow$ efficient techniques to avoid or at least delay the curse of simensionality
  - For example, an 80D/700-term surrogate employs terms of the form λ_i^4, $\lambda_i^3\lambda_j$, $\ldots$.
  - An adaptive set of sampling points require about 3200 additional simulations.
- How do we position these sample points in the "active vegetation" region to actually improve the BCS surrogate ? (80D domain mapping)

(see recent talk by Patrick Conrad/Youssef Marzouk, and pre-print by Paul Constantine)

Summary

Surrogate models are necessary for complex climate 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
- Adaptive sparse quadrature for relevant basis terms to build a more accurate surrogate