

# *Efficient Surrogate Construction for High-Dimensional Climate Models*

C. Safta<sup>1</sup>, K. Sargsyan<sup>1</sup>, D. Ricciuto<sup>2</sup>, R.D. Berry<sup>1</sup>  
B. Debusschere<sup>1</sup>, H. Najm<sup>1</sup>, P. Thornton<sup>2</sup>

<sup>1</sup>Sandia National Laboratories  
Livermore, CA, USA

<sup>2</sup>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)

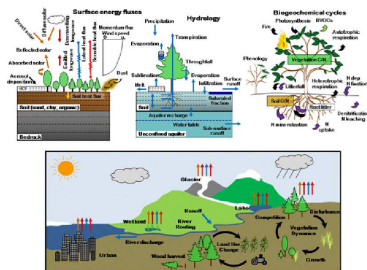
Sandia National Laboratories is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under contract DE-AC04-94-AL85000.

# Outline

- 1 Challenges
- 2 Surrogate Models
- 3 Constrained Parameter Space
- 4 Bayesian Compressive Sensing
  - Methodology
  - CLM results
- 5 Classification
- 6 Adaptive Sparse Quadrature
- 7 Summary

# 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

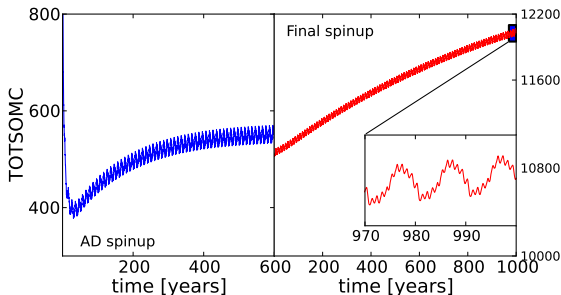


<http://www.cesm.ucar.edu/models/clm/>

- 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  $\sim 80$  input parameters

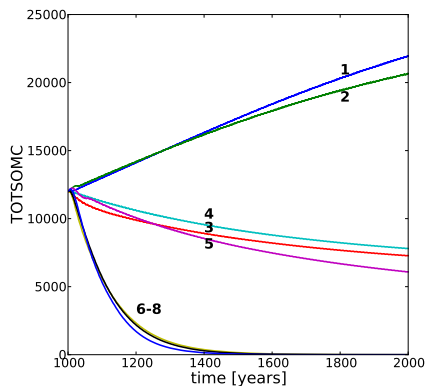
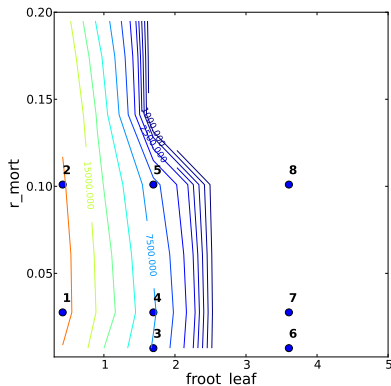
# Community Land Model - Typical Setup

- Generate initial conditions



# Community Land Model - Typical Setup

- Sample the parameter space



# 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  $\lambda$
- Computationally expensive model  $f(\cdot)$  (e.g. climate models)
- 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



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

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

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

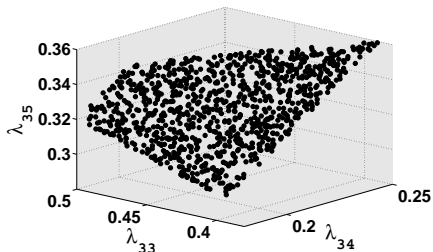
$$\lambda_i = \frac{a_i + b_i}{2} + \frac{b_i - a_i}{2} \eta_i.$$

Output is represented with respect to Legendre polynomials

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

# Map Constrained Parameters to Unconstrained Spaces

- Given a vector of random variables  $\lambda = (\lambda_1, \dots, \lambda_{d'})$  with known joint cumulative distribution function (CDF)  $F(\lambda_1, \dots, \lambda_{d'})$
- 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  $[-1, 1]$ .



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

# 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  $p(\mathbf{c}|D)$ , 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)

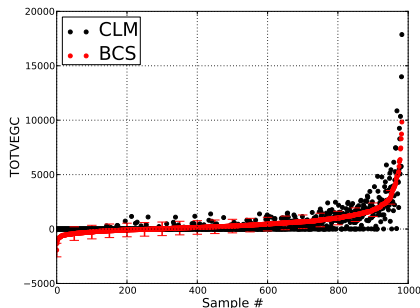
- The number of polynomial basis terms grows fast; a  $p$ -th order,  $d$ -dimensional basis has a total of  $(p + d)!/(p!d!)$  terms.
- Dimensionality reduction by using hierarchical priors.

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

- The parameter  $\alpha$  can be further modeled hierarchically, or fixed.
- The parameters  $(\sigma^2, s_0^2, \dots, s_K^2)$  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

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



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

# Climate Land Model - 1<sup>st</sup> order BCS

- Ranking of the most important input parameters for each output

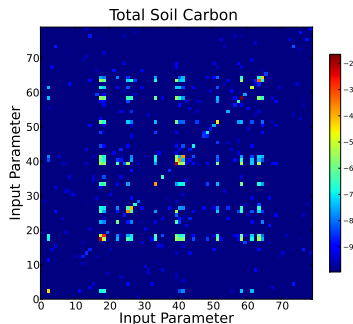
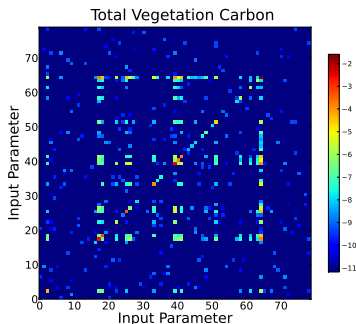
$$S_i = \frac{\sum_{k \in \mathbb{I}_i} c_k^2 \|\Psi_k\|^2}{\sum_{k > 0} c_k^2 \|\Psi_k\|^2}$$

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<sup>nd</sup> order BCS

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

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

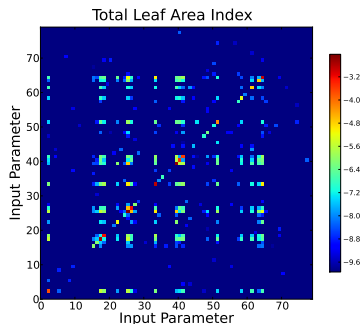
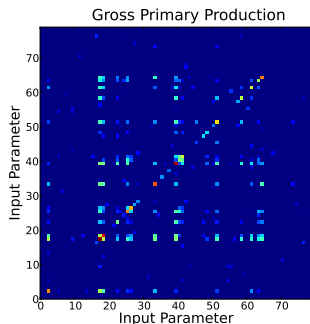




# Climate Land Model - 2<sup>nd</sup> order BCS

- Most influential input parameter couplings for each output - energy 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

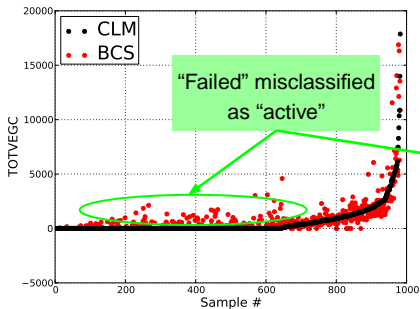
- Large regions of the original quasi-hypercube parameter space lead to simulations with failed vegetation.
- Partition the space using a classification algorithm
  - Cons: Classification will introduce errors
  - Pros: iBCS algorithm will avoid the “failed vegetation” plateau. Will only use the “active”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

# Classification+iBCS

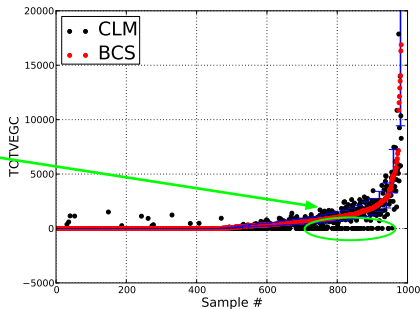
- Classification using Random Decision Forests
  - consists of a collection of decision trees with controlled variation
  - the result is the mode of the results from individual decision trees
  - AlgLib software library (<http://www.alglib.net>)
- Calibration using 9K samples/Validation using 1K samples
- Shift accuracy from “failed vegetation” plateau to “active vegetation” regions
- Apply the iBCS algorithm on “active vegetation” results

# Classification+iBCS - cont'd

## Ascending *CLM* results



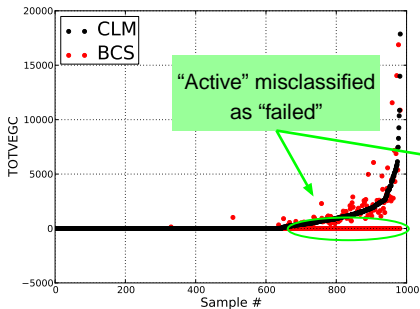
## Ascending *iBCS* results



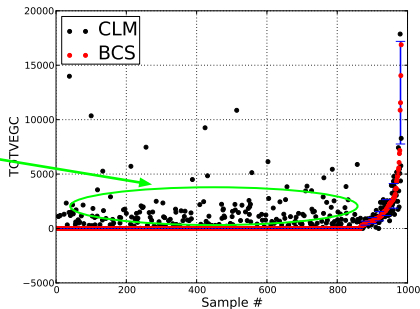
- Bias towards the “active” simulations.

# Classification+iBCS - cont'd

## Ascending *CLM* results



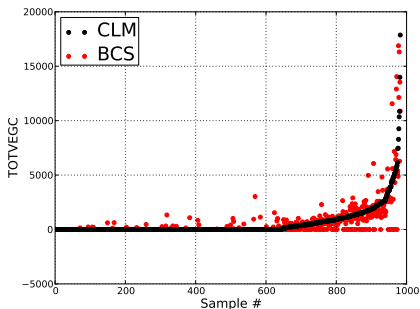
## Ascending *iBCS* results



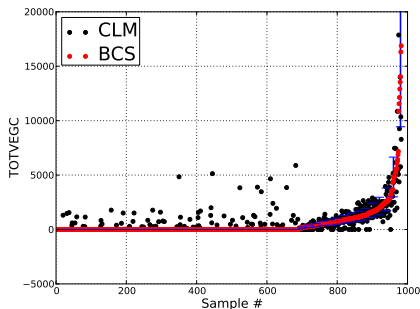
- Bias towards the "failed" simulations.

# Classification+iBCS - cont'd

## Ascending *CLM* results



## Ascending *iBCS* results



- The minimum classification error for the ensemble of CLM simulations is around 11%

# Adaptive Sparse Quadrature - In progress

- Improve the iBCS surrogate using Galerkin projection → efficient techniques to avoid or at least delay the curse of dimensionality
  - For example, an 80D/700-term surrogate employs terms of the form  $\Psi_4(\lambda_i), \Psi_3(\lambda_i)\Psi_1(\lambda_j), \dots$
  - 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)

(Nobile *et al.*, Gerstner & Griebel, Conrad & Marzouk, Paul Constantine)

# Adaptive Sparse Quadrature - In progress

- Improve the iBCS surrogate using Galerkin projection → efficient techniques to avoid or at least delay the curse of dimensionality
  - For example, an 80D/700-term surrogate employs terms of the form  $\Psi_4(\lambda_i), \Psi_3(\lambda_i)\Psi_1(\lambda_j), \dots$
  - 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)

(Nobile *et al.*, Gerstner & Griebel, Conrad & Marzouk, 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