Surrogate-Based Uncertainty Quantification in Climate Models
in presence of High Dimensional, Dependent Inputs and Nonsmooth Outputs

KHACHIK SARGSYAN, COSMIN SAFTA, JAIDEEP RAY,
BERT DEBUSSCHERE, HABIB NAJM
Sandia National Laboratories

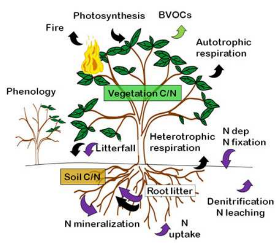
DANIEL RICCIUTO,
PETER THORNTON
Oak Ridge National Laboratory

Corresponding email: ksargsy@sandia.gov

AGU Fall Meeting 2012 Poster ID: GC43E-1082

Community Land Model (CLM)

- Nested computational grid hierarchy
- Represents spatial heterogeneity of the land surface
- A single-site, 1000-yr simulation ~ 10 hrs on 1 CPU
- Involves ~ 80 input parameters



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

Uncertainty quantification challenges

- Computationally expensive model simulations (**1**)
- Physical constraints on some input parameters (**2**)
- High-dimensional input parameter space (**3**)
- Nonsmooth dependence of outputs on inputs (**4**)

1 Surrogate $g_{\boldsymbol{c}}(\boldsymbol{\lambda}) \approx G(\boldsymbol{\lambda})$ is necessary for expensive models

The surrogate model can be queried instead of the CLM for a) Global sensitivity analysis, b) Optimization, c) Forward uncertainty propagation, d) Calibration.

Polynomial chaos model as a surrogate model

- Allows propagation of input parameter uncertainties to outputs of interest
- Interprets input parameters as random variables

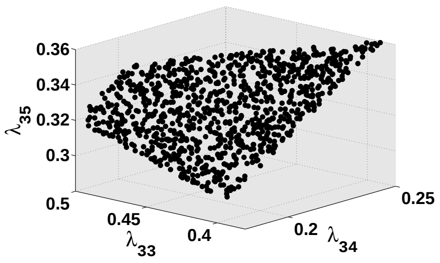
With input parameters modeled as uniform $\lambda_i \sim \mathrm{Uniform}[a_i, b_i]$,

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

Output is represented with respect to Legendre polynomials

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

2 Dependent inputs are mapped to independent ones by Rosenblatt transformation (RT)



Due to two mass fraction constraints in CLM, the RT maps an 81-dimensional parameter vector $\boldsymbol{\lambda}$ to $\boldsymbol{\eta} \in [-1, 1]^{79}$.

3 Bayesian inference of PC modes leads to a probabilistic surrogate

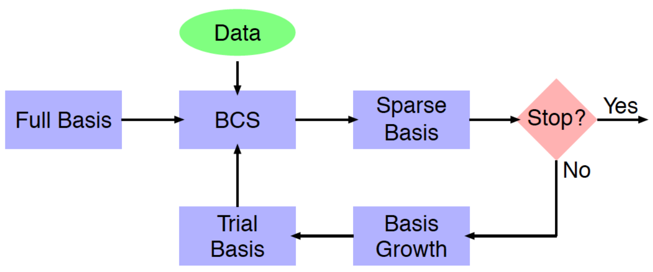
$$p(\boldsymbol{c}|\alpha, \mathcal{D}) \propto \quad L_{\mathcal{D}}(\boldsymbol{c}) \quad \times \quad p(\boldsymbol{c}|\alpha)$$

Data $\mathcal{D}$ is the set of all training runs $\mathcal{D} = (\boldsymbol{\lambda}_i, G(\boldsymbol{\lambda}_i))_{i=1}^{N}$. The size of $\boldsymbol{c}$, i.e. the number of polynomial basis terms grows fast; a p-th order, d-dimensional basis has a total of $(p+d)!/(p!d!)$ terms. Sparsity priors strive to detect the smallest set of basis.

Gaussian likelihood
$$L_{\mathcal{D}}(\boldsymbol{c}) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}s} \exp\left(-\frac{(G(\boldsymbol{\lambda}_i) - g_{\boldsymbol{c}}(\boldsymbol{\eta}_i))^2}{2s^2}\right)$$

Laplace prior
$$p(\boldsymbol{c}|\alpha) = \int \prod_{k=0}^{K-1} p(c_k|\sigma_k^2) p(\sigma_k^2|\alpha) d\sigma_k^2 = \prod_{k=0}^{K-1} \frac{\sqrt{\alpha}}{2} e^{-\sqrt{\alpha}|c_k|}$$

Iterative Bayesian compressive sensing (iBCS): dimensionality reduction using sparsity priors

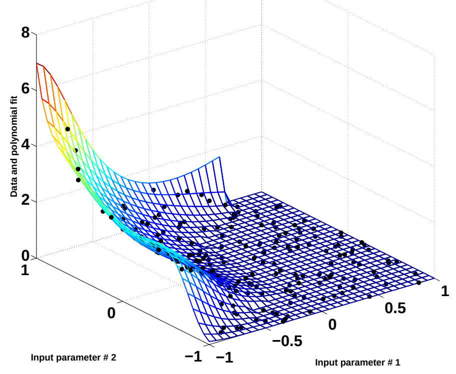


4 Data clustering and classification-enhanced piecewise PC surrogate for nonsmooth outputs

Declare a surrogate

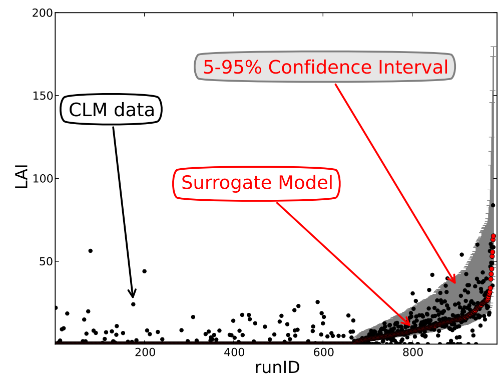
$$g_s(\boldsymbol{x}) = \begin{cases} g_{\boldsymbol{c}_1}(\boldsymbol{x}) & \text{if } \boldsymbol{x} \in^* \mathcal{D}_1 \\ g_{\boldsymbol{c}_2}(\boldsymbol{x}) & \text{if } \boldsymbol{x} \in^* \mathcal{D}_2 \end{cases}$$

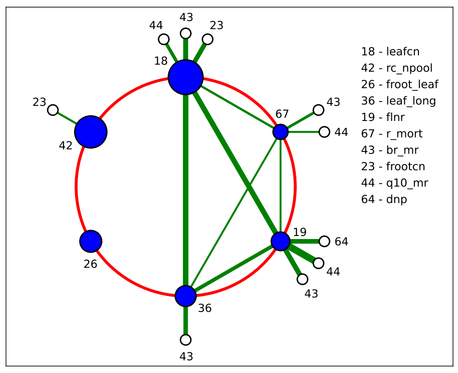
* Requires a classification step to find out which cluster $\boldsymbol{x}$ belongs to. We applied Random Decision Forests (RDF).



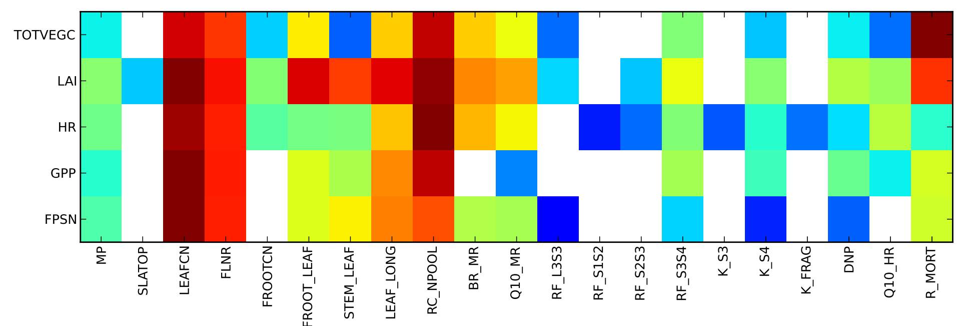
Surrogate model and global sensitivity indices for Leaf Area Index

- Surrogate constructed using only 10000 training simulations
- While full second order basis has ~ 3000 terms, the iterative BCS algorithm picks only ~ 100 of them that are able to capture the data well
- Surrogate sensitivity indices computed via Monte Carlo
- Circle sizes correspond to main effect sensitivity indices
- Line widths correspond to joint sensitivity indices





Sensitivity ranking of the most important input parameters for each output



Highlights

- Surrogate models are necessary for complex climate models
- Polynomial Chaos surrogate is inferred using Bayesian machinery
- High-dimensionality is addressed by the iterative Bayesian compressive sensing (iBCS) algorithm
- Constrained/dependent input parameters are mapped to an unconstrained input parameter set via Rosenblatt transformation
- Data clustering and classification employed for nonsmooth models to obtain a piecewise-PC surrogate model

Future work

- Sampling in the reduced space to build a more accurate surrogate
- Calibration of input parameters given observational data
- Optimal computational design to improve BCS efficiency
- Local surrogate construction tailored to given observational data

References

[1] S. Babacan, R. Molina, and A. Katsaggelos. Bayesian compressive sensing using Laplace priors. *IEEE Transactions on Image Processing*, 19(1):53–63, 2010.

[2] K. Sargsyan, C. Safta, H. Najm, B. Debusschere, D. Ricciuto, and P. Thornton. Dimensionality reduction for complex models via Bayesian compressive sensing. *in preparation*.

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.

Sandia National Laboratories

Sandia National Laboratories is a multiprogram 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 No. DE-AC04-94AL85000.