SAND2011-9159C

# Uncertainty Quantification in Climate Modeling

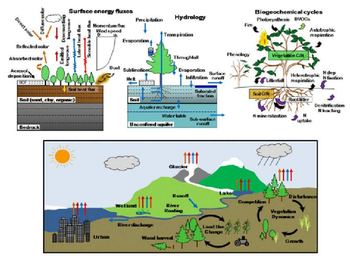
KHACHIK SARGSYAN, COSMIN SAFTA, ROBERT BERRY, BERT DEBUSSCHERE, HABIB NAJM
Sandia National Laboratories

DANIEL RICCIUTO, PETER THORNTON
Oak Ridge National Laboratory

AGU Fall Meeting 2011 Poster ID: GC11A-0899 Corresponding email: ksargsy@sandia.gov

## Uncertainty quantification 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/

## Community Land Model (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

## Problem formulation: surrogate model construction

- Input parameter vector $\boldsymbol{\lambda}$
- Forward function (CLM simulation) $f(\cdot)$
- Given a set of *training* model runs, $(\boldsymbol{\lambda}_i, f(\boldsymbol{\lambda}_i))_{i=1}^N$, build a *surrogate* $f_s(\cdot) \approx f(\cdot)$ that is cheap to evaluate

The surrogate model can be used for

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

## Polynomial chaos spectral representation serves as a surrogate model

To build a surrogate representation for input-output relationship, Polynomial Chaos (PC) spectral expansions are used; see [1].

- 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

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

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

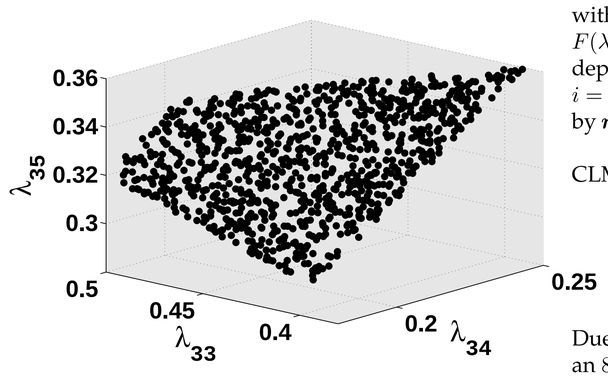
If input parameters are uniform $\lambda_i \sim \text{Uniform}[a_i, b_i]$, then

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

Output is represented with respect to Legendre polynomials

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

## Rosenblatt transformation maps constrained input parameters to an unconstrained space




Given a vector of random variables $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{d'})$ with known joint cumulative distribution function (CDF) $F(\lambda_1, \ldots, \lambda_{d'})$, one can obtain a set of $\eta_i$'s that are independent uniform random variables on $[-1,1]$ for all $i = 1,2,\ldots,d$, using conditional CDFs. This map, denoted by $\boldsymbol{\eta} = R(\boldsymbol{\lambda})$, is called the *Rosenblatt transformation* (RT) [4].

CLM implemented with input parameter constraints

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

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

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

Bayes formula

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

relates the prior distribution $p(\boldsymbol{c})$ of PC modes to the posterior $p(\boldsymbol{c}|\mathcal{D})$, 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 [5],

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

## Iterative Bayesian compressive sensing (BCS): dimensionality reduction by using sparsity priors

The number of polynomial basis terms grows fast; a $p$-th order, $d$-dimensional basis has a total of $(p+d)!/(p!d!)$ terms. With large $d$, one can not afford to build a PC basis of order greater than two.
In order to use as few basis terms as possible, Gaussian *sparsity* priors are taken

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

Then, the posterior is an analytically tractable multivariate normal distribution. The parameters $(\sigma^2, s_0^2, \ldots, s_K^2)$ are fixed by evidence maximization. The optimization leads to very small $s_i^2$ for some indices $i$ - the corresponding bases are discarded. For details, see [2].

*Iterative BCS*: We propose and implement an iterative procedure that allows increasing the order for the relevant basis terms while maintaining the dimensionality reduction. Only basis elements best explained by the data are retained [3].

## Number of training runs needed for reliable detection of important dimensions

Assume $\boldsymbol{\lambda} \sim \text{Uniform}[-1,1]^d$ and consider test function

$$y = \exp\left(\sum_{i=1}^{d} a_i \lambda_i\right),$$

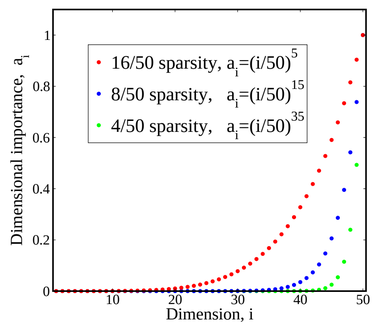
Dimensionality 'importances' dialed-in

$$a_i = \begin{cases} 1 & \text{if } 1 \le i \le d_{\text{imp}}, \\ 0.1 & \text{if } d_{\text{imp}} < i \le d, \end{cases}$$

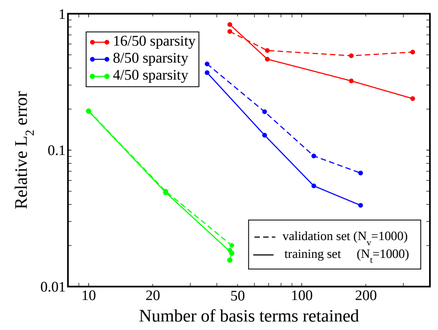
| $d_{\text{imp}}$ | $d$ | $N$ |
|---|---|---|
| 1 | any | 20 |
| 2 | any | 50 |
| 5 | any | 130 |
| 10 | any | 980 |

### Error for different model sparsity levels

Dimensionality 'importances' are chosen so that 90% of energy is in a small subset of dimensions, i.e. model is 'sparse'.



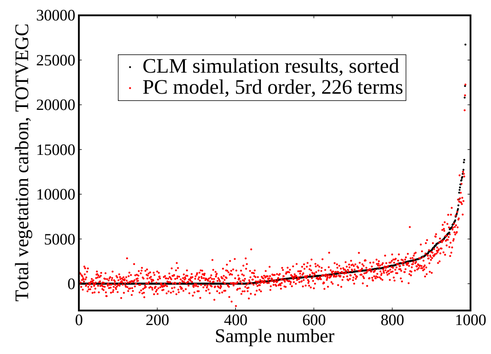

Validation error increase indicates overfitting. $N_t = 1000$ training runs are sufficient if $\sim 10$ dimensions matter.




## Results based on $N = 987$ training runs of CLM

- Single-site mode, $N = 987$ training runs
- 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 |

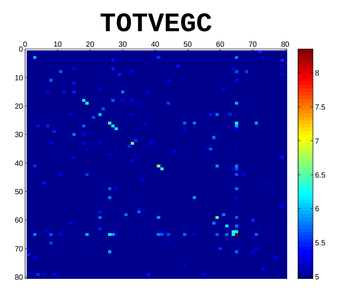


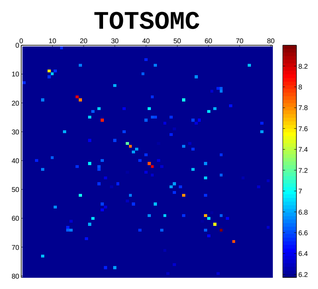

### First order BCS: ranking of the most important input parameters for each output

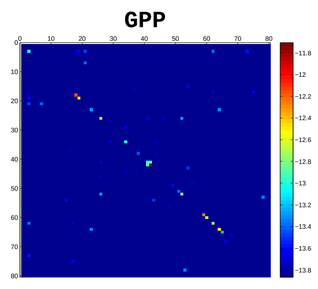
| rank | TOTVEGC | TOTSOMC | GPP | ERR | TLAI | EFLX_LH_TOT | FSH |
|---|---|---|---|---|---|---|---|
| 1 | r_mort | q10_mr | leafcn | k_s4 | froot_leaf | leafcn | rholnir |
| 2 | q10_mr | leafcn | k_s4 | froot_leaf | q10_mr | q10_mr | q10_mr |
| 3 | froot_leaf | froot_leaf | froot_leaf | q10_hr | q10_hr | froot_leaf | leafcn |
| 4 | br_mr | br_mr | flnr | fflnr | leaf_long | k_s4 | br_mr |
| 5 | q10_mr | fflnr | q10_mr | q10_mr | k_s4 | br_mr | flnr |
| 6 | leafcn | dnp | q10_hr | dnp | br_mr | flnr | k_s4 |
| 7 | k_s4 | q10_hr | dnp | rf_s3s4 | dnp | leaf_long | taulnir |
| 8 | stem_leaf | leaf_long | rf_s3s4 | leaf_long | stem_leaf | q10_hr | froot_leaf |
| 9 | flnr | k_s4 | leaf_long | mp | r_mort | rf_s3s4 | frootcn |
| 10 | dnp | frootcn | br_mr | bdnr | rf_s3s4 | stem_leaf | f_frag |

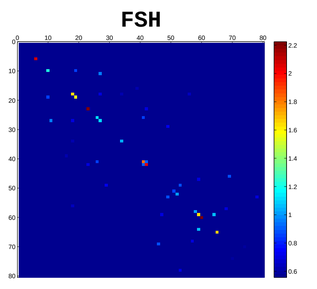
### Second order BCS: most influential input parameter couplings for each output

- Each axis corresponds to the input parameter list. (i,j) element corresponds to the $\lambda_i \lambda_j$ basis term
- While full second order basis has $\sim 3000$ terms, the iterative BCS algorithm picks only $\sim 100$ of them that are able to capture the data well
- Higher order results leads to more accurate surrogate models. However, strong output non-linearities and smooth bases make input domain decomposition methods necessary




## Conclusions

- 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

## Future work

- Targeted sampling for relevant basis terms to build a more accurate surrogate
- Clustering techniques for efficient domain decomposition to relieve the non-linear effects
- Spatially distributed input parameters, global CLM

## References

[1] R.G. Ghanem and P.D. Spanos. *Stochastic Finite Elements: A Spectral Approach*. Springer Verlag, New York, 1991.

[2] S. Ji, Y. Xue, and L. Carin. Bayesian compressive sensing. *IEEE Transactions on Signal Processing*, 56(6):2346–2356, 2008.

[3] R. Berry J. Ray B. Debusschere K. Sargsyan, C. Safta and H. Najm. Efficient uncertainty quantification methodologies for high-dimensional climate land models. Technical report, SAND 2011-8757, November 2011.

[4] M. Rosenblatt. Remarks on a multivariate transformation. *Annals of Mathematical Statistics*, 23(3):470 – 472, 1952.

[5] K. Sargsyan, C. Safta, B. Debusschere, and H. Najm. Multiparameter spectral representation of noise-induced competence in Bacillus subtilis. *in preparation*, 2011.

*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 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.*