Uncertainty Quantification and Data Assimilation

Applications in chemical kinetics and climate modeling

Khachik Sargsyan

Sandia National Laboratories, Livermore, CA
Transportation Energy Center
Reacting Flow Research Department (8351)

Background

- 1997-2002, B.S., Applied Mathematics and Applied Physics
  - Moscow Institute of Physics and Technology
- 2002-2007, Ph.D., Applied and Interdisciplinary Math
  - University of Michigan, Dept of Mathematics
  - Thesis: “Mean First Passage Times in the Near-Continuum Limit of Birth-Death Processes”
- since July 2007, Postdoctoral Appointee
  - Sandia National Labs, Reacting Flow Research Dept (8351)

Projects while at Sandia

- ”Stochastic Dynamical Systems: Spectral Methods for the Analysis of Dynamics and Predictability”

  supported by DOE ASCR Applied Math,
  PI: Bert Debusschere

- ”Uncertainty Quantification for Large Scale Ocean Circulation Predictions”

  supported by Sandia Seniors’ Council LDRD,
  PI: Cosmin Safta

- ”Quantifying the Margin of High-Consequence Climate Change”

  supported by NNSA and DOE BER,
  Sandia-CA POC: Khachik Sargsyan

- ”Analysis of Stochasticity in Immune System Signaling Pathways”

  supported by UTMB-Sandia Joint Institute of Biosecurity,
  PI: Bert Debusschere

Uncertainty Quantification: what, how, why?

- What is UQ?
  - The effect of input uncertainties on the outputs of interest.
- Uncertainty sources
  - Model parameters
  - Initial/boundary conditions
  - Model geometry/structure
- Why is it important?
  - Model validation
  - Confidence assessment
  - Optimal design
  - Data assimilation

    - Combination of measurements and model predictions to obtain accurate representations.

Uncertainty Quantification: what, how, why?

- What is UQ?
  - The effect of input uncertainties on the outputs of interest.
- Uncertainty sources
  - Model parameters
  - Initial/boundary conditions
  - Model geometry/structure
- Why is it important?
  - Model validation
  - Confidence assessment
  - Optimal design
  - Data assimilation

    - Combination of measurements and model predictions to obtain accurate representations.

Uncertainty Quantification: what, how, why?

- What is UQ?
  - The effect of input uncertainties on the outputs of interest.
- Uncertainty sources
  - Model parameters
  - Initial/boundary conditions
  - Model geometry/structure
- Why is it important?
  - Model validation
  - Confidence assessment
  - Optimal design
  - Data assimilation
    - Combination of measurements and model predictions to obtain accurate representations.

Uncertainty Quantification: Components and Methods

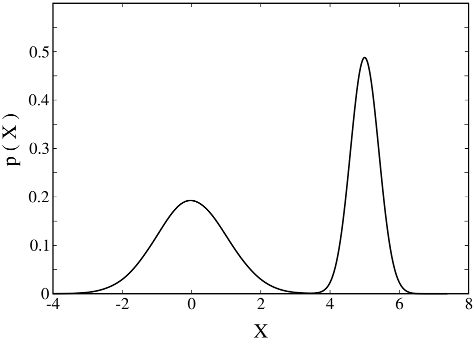
- UQ components
  - Sensitivity analysis
    - Small parameter perturbations
  - Predictability assessment
    - Larger parameter uncertainties
  - Parameter estimation
    - Inverse problem
  - Dynamical analysis

Uncertainty Quantification: Components and Methods

- UQ components
  - Sensitivity analysis
    - Small parameter perturbations
  - Predictability assessment
    - Larger parameter uncertainties
  - Parameter estimation
    - Inverse problem
  - Dynamical analysis
- UQ Methods
  - Direct (intrusive)
    - Derive new forward model
    - Intrusive Spectral Projection (ISP)
  - Sampling (non-intrusive)
    - Monte-Carlo, Quasi Monte-Carlo
    - Non-intrusive Spectral Projection (NISP)

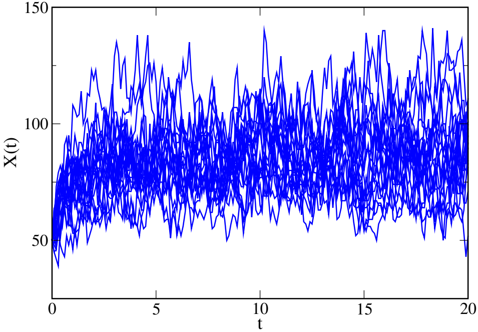
UQ methods are challenged by..

- Nonlinearities, Bifurcations, Bimodalities
- Intrinsic stochasticity
- Limited data
- Tail regions
- Curse of dimensionality



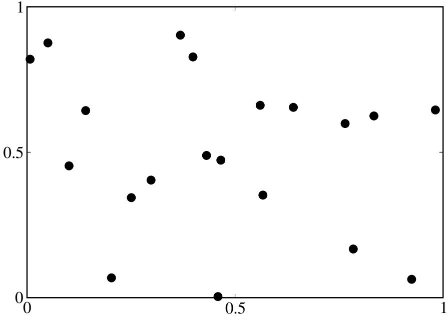
UQ methods are challenged by..

- Nonlinearities, Bifurcations, Bimodalities
- Intrinsic stochasticity
- Limited data
- Tail regions
- Curse of dimensionality



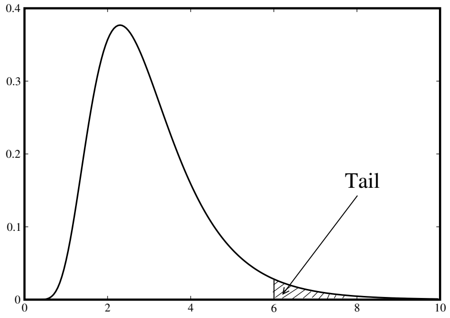
UQ methods are challenged by..

- Nonlinearities, Bifurcations, Bimodalities
- Intrinsic stochasticity
- Limited data
- Tail regions
- Curse of dimensionality



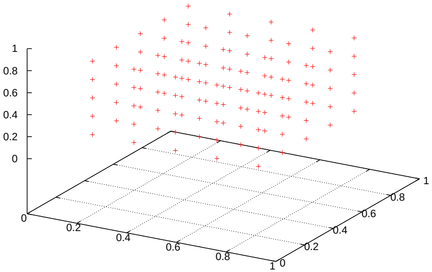
UQ methods are challenged by..

- Nonlinearities,
  Bifurcations,
  Bimodalities
- Intrinsic stochasticity
- Limited data
- Tail regions
- Curse of dimensionality



UQ methods are challenged by..

- Nonlinearities, Bifurcations, Bimodalities
- Intrinsic stochasticity
- Limited data
- Tail regions
- Curse of dimensionality



Projects

- ”Stochastic Dynamical Systems: Spectral Methods for the Analysis of Dynamics and Predictability”

  supported by DOE ASCR Applied Math,
  PI: Bert Debusschere

- ”Uncertainty Quantification for Large Scale Ocean Circulation Predictions”

  supported by Sandia Seniors’ Council LDRD,
  PI: Cosmin Safta

- ”Quantifying the Margin of High-Consequence Climate Change”

  supported by DOE BER,
  Sandia-CA POC: Khachik Sargsyan

- ”Analysis of Stochasticity in Immune System Signaling Pathways”

  UTMB-Sandia Joint Institute of Biosecurity,
  PI: Bert Debusschere

Projects

- ”Stochastic Dynamical Systems: Spectral Methods for the Analysis of Dynamics and Predictability”

  supported by DOE ASCR Applied Math,
  PI: Bert Debusschere

- ”Uncertainty Quantification for Large Scale Ocean Circulation Predictions”

  supported by Sandia Seniors’ Council LDRD,
  PI: Cosmin Safta

- ”Quantifying the Margin of High-Consequence Climate Change”

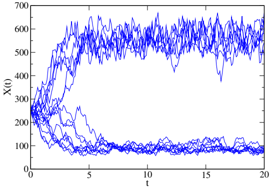
  supported by DOE BER,
  Sandia-CA POC: Khachik Sargsyan

- ”Analysis of Stochasticity in Immune System Signaling Pathways”

  UTMB-Sandia Joint Institute of Biosecurity,
  PI: Bert Debusschere

Stochastic Reaction Networks

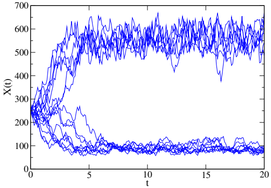
- Reaction networks involving small number of molecules necessitate the use of *stochastic* modeling instead of the *deterministic* one. E.g.
  - Microbial processes (bioenergy, bioremediation)
  - Surface catalytic reactions (fuel cells, batteries)
  - Immune system signaling reactions



- SRNs are modeled as Jump Markov Processes
  - Governed by Chemical Master Equation $\dot{P}(X(t) = n) = \sum_m A_{nm} P(X(t) = n)$
  - Reduces to deterministic Rate Equations in the large volume limit
  - Trajectories simulated by Gillespie's Stochastic Simulation Algorithm (SSA, Gillespie, 1977)

Stochastic Reaction Networks

- Reaction networks involving <u>small number of molecules</u> necessitate the use of *stochastic* modeling instead of the *deterministic* one. E.g.
  - Microbial processes (bioenergy, bioremediation)
  - Surface catalytic reactions (fuel cells, batteries)
  - Immune system signaling reactions



- SRNs are modeled as Jump Markov Processes
  - Governed by Chemical Master Equation $\dot{P}(X(t)=n)=\sum_m A_{nm}P(X(t)=n)$
  - Reduces to deterministic Rate Equations in the large volume limit
  - Trajectories simulated by Gillespie's Stochastic Simulation Algorithm (SSA, Gillespie, 1977)

Goals and Tools

$$X(t, \boldsymbol{\theta}, \boldsymbol{\lambda})$$

- Develop tools for *predictability*(λ) and *dynamical analysis*(t) of SRNs accounting for
  - Inherent stochasticity (θ)
  - Model/parameter variability (λ)
  - Limited data

$$\mathcal{D} = \{X_i\}_{i=1}^{N}$$

- Predictability assessment
  - Fix t, focus on λ dependence
  - Polynomial chaos; Bayesian inference; Domain decomposition

Goals and Tools

$$X(t, \boldsymbol{\theta}, \boldsymbol{\lambda})$$

- Develop tools for *predictability*$(\boldsymbol{\lambda})$ and *dynamical analysis*(t) of SRNs accounting for
  - Inherent stochasticity $(\boldsymbol{\theta})$
  - Model/parameter variability $(\boldsymbol{\lambda})$
  - Limited data

$$\mathcal{D} = \{X_i\}_{i=1}^N$$

- Predictability assessment
  - Fix t, focus on λ dependence
  - Polynomial chaos; Bayesian inference; Domain decomposition

Goals and Tools

$$X(\boldsymbol{\theta}, \boldsymbol{\lambda})$$

- Develop tools for *predictability*$(\boldsymbol{\lambda})$ and *dynamical analysis*(t) of SRNs accounting for
  - Inherent stochasticity $(\boldsymbol{\theta})$
  - Model/parameter variability $(\boldsymbol{\lambda})$
  - Limited data

$$\mathcal{D} = \{X_i\}_{i=1}^{N}$$

- Predictability assessment
  - Fix t, focus on $\boldsymbol{\lambda}$ dependence
  - Polynomial chaos; Bayesian inference; Domain decomposition

Schlögl Model is a prototype bistable model

- Reactions

$$A + 2X \underset{a_2}{\overset{a_1}{\rightleftarrows}} 3X$$

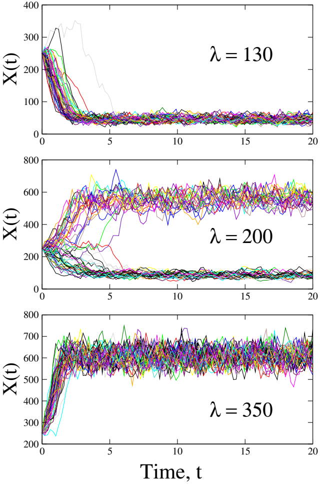
$$B \underset{a_4}{\overset{a_3}{\rightleftarrows}} X$$

- Propensities

$$a_1 = k_1 A X(X-1)/2,$$
$$a_2 = k_2 X(X-1)(X-2)/6,$$
$$a_3 = k_3 B,$$
$$a_4 = k_4 X.$$

- Nominal parameters

$k_1 A$	0.03
k_2	0.0001
$k_3 B = \lambda$	200
k_4	3.5
A	10^5
B	$2 \cdot 10^5$
$X(0)$	250



Polynomial Chaos expansion represents any random variable as a polynomial of a standard random variable

- Truncated PCE: finite dimension n and order p

$$X(\boldsymbol{\theta}) \simeq \sum_{k=0}^{P} c_k \Psi_k(\boldsymbol{\eta})$$

with the number of terms $P + 1 = \frac{(n+p)!}{n!p!}$.

- $\boldsymbol{\eta} = (\eta_1, \cdots, \eta_n)$ standard i.i.d. r.v.
  Ψ_k standard orthogonal polynomials
  c_k spectral modes.

- Most common standard Polynomial-Variable pairs:
  (continuous) Gauss-Hermite, Legendre-Uniform,
  (discrete) Poisson-Charlier.

[Wiener, 1938; Ghanem & Spanos, 1991; Xiu & Karniadakis, 2002]

Polynomial Chaos expansion represents any random variable as a polynomial of a standard random variable

- Truncated PCE: finite dimension n and order p

$$X(\boldsymbol{\theta}) \simeq \sum_{k=0}^{P} c_k \Psi_k(\boldsymbol{\eta})$$

  with the number of terms $P+1 = \frac{(n+p)!}{n!p!}$.

- $\boldsymbol{\eta} = (\eta_1, \cdots, \eta_n)$ standard i.i.d. r.v.
  Ψ_k standard orthogonal polynomials
  c_k spectral modes.

- Most common standard Polynomial-Variable pairs:
  (continuous) Gauss-Hermite, Legendre-Uniform,
  (discrete) Poisson-Charlier.

[Wiener, 1938; Ghanem & Spanos, 1991; Xiu & Karniadakis, 2002]

Polynomial Chaos expansion represents any random variable as a polynomial of a standard random variable

- Truncated PCE: finite dimension n and order p

$$X(\boldsymbol{\theta}) \simeq \sum_{k=0}^{P} c_k \Psi_k(\boldsymbol{\eta})$$

  with the number of terms $P + 1 = \frac{(n+p)!}{n!p!}$.

- $\boldsymbol{\eta} = (\eta_1, \cdots, \eta_n)$ standard i.i.d. r.v.
  Ψ_k standard orthogonal polynomials
  c_k spectral modes.

- Most common standard Polynomial-Variable pairs:
  (continuous) Gauss-Hermite, Legendre-Uniform,
  (discrete) Poisson-Charlier.

[Wiener, 1938; Ghanem & Spanos, 1991; Xiu & Karniadakis, 2002]

Galerkin Projection is typically needed

PC expansion: $X(\boldsymbol{\theta}) \simeq \sum_{k=0}^{P} c_k \Psi_k(\boldsymbol{\eta}) = g_{\mathcal{D}}(\boldsymbol{\eta})$

Orthogonal projection: $c_k = \frac{\langle X(\boldsymbol{\theta})\Psi_k(\boldsymbol{\eta})\rangle}{\langle \Psi_k^2(\boldsymbol{\eta})\rangle}$

- Intrusive Spectral Projection (ISP)
  - ⋆ Direct projection of governing equations
  - ⋆ Leads to deterministic equations for PC coefficients
  - ∗ No explicit governing equation for SRNs
- Non-intrusive Spectral Projection (NISP)
  - ⋆ Sampling based
  - ⋆ No explicit evolution equation for X needed
  - ∗ Galerkin projection not well-defined for SRNs

Galerkin Projection is typically needed

PC expansion: $X(\boldsymbol{\theta}) \simeq \sum_{k=0}^{P} c_k \Psi_k(\boldsymbol{\eta}) = g_{\mathcal{D}}(\boldsymbol{\eta})$

Orthogonal projection: $c_k = \frac{\langle X(\boldsymbol{\theta})\Psi_k(\boldsymbol{\eta})\rangle}{\langle \Psi_k^2(\boldsymbol{\eta})\rangle}$

- Intrusive Spectral Projection (ISP)
  - ⋆ Direct projection of governing equations
  - ⋆ Leads to deterministic equations for PC coefficients
  - ∗ No explicit governing equation for SRNs
- Non-intrusive Spectral Projection (NISP)
  - ⋆ Sampling based
  - ⋆ No explicit evolution equation for X needed
  - ∗ Galerkin projection not well-defined for SRNs

Galerkin Projection is typically needed

PC expansion: $X(\boldsymbol{\theta}) \simeq \sum_{k=0}^{P} c_k \Psi_k(\boldsymbol{\eta}) = g_{\mathcal{D}}(\boldsymbol{\eta})$

Orthogonal projection: $c_k = \frac{\langle X(\boldsymbol{\theta})\Psi_k(\boldsymbol{\eta})\rangle}{\langle \Psi_k^2(\boldsymbol{\eta})\rangle}$

- Intrusive Spectral Projection (ISP)
  - ⋆ Direct projection of governing equations
  - ⋆ Leads to deterministic equations for PC coefficients
  - ∗ No explicit governing equation for SRNs
- Non-intrusive Spectral Projection (NISP)
  - ⋆ Sampling based
  - ⋆ No explicit evolution equation for X needed
  - ∗ Galerkin projection not well-defined for SRNs

Bayesian inference handles the intrinsic stochasticity well

$$X \simeq \sum_{k=0}^{P} c_k \Psi_k(\eta) = g_{\mathcal{D}}(\eta)$$

$$\overbrace{P(\boldsymbol{c}|\mathcal{D})}^{\text{Posterior}} \propto \overbrace{P(\mathcal{D}|\boldsymbol{c})}^{\text{Likelihood}} \overbrace{P(\boldsymbol{c})}^{\text{Prior}}$$

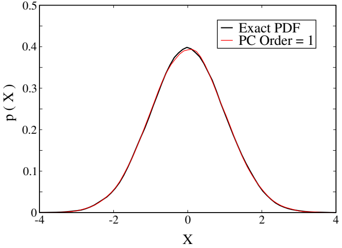
$$L(\boldsymbol{c}) = P(\mathcal{D}|\boldsymbol{c}) = \prod_{i=1}^{N} \mathrm{pdf}_g(X_i)$$

- Noise model is inherent in SSA data $\mathcal{D} = \{X_i\}_{i=1}^{N}$
- Uniformly distributed priors
- Posterior exploration using Markov Chain Monte Carlo (MCMC)
- The whole posterior distribution is accessible
- Maximum a posteriori (MAP) estimate: $\boldsymbol{c}^{MAP} = \mathrm{argmax}_{\boldsymbol{c}} P(\boldsymbol{c}|\mathcal{D})$

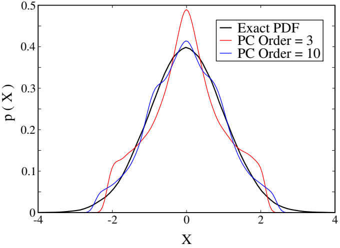
However, global methods are challenged by nonlinear/bimodal systems

Normal Random Variable

Gauss-Hermite PC



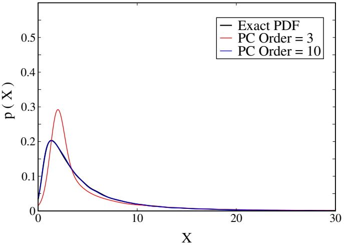
Legendre-Uniform PC



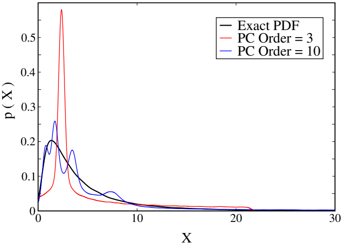
However, global methods are challenged by nonlinear/bimodal systems

Lognormal Random Variable

Gauss-Hermite PC



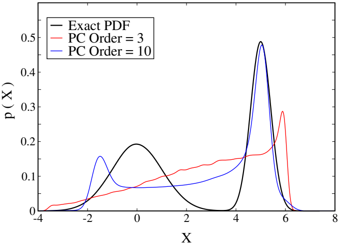
Legendre-Uniform PC



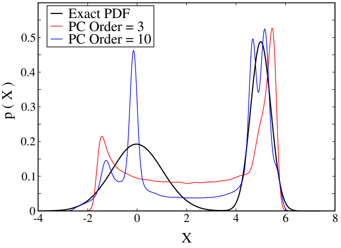
However, global methods are challenged by nonlinear/bimodal systems

Binormal Random Variable

Gauss-Hermite PC



Legendre-Uniform PC



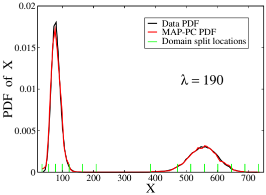
Adaptivity criterion for domain decomposition

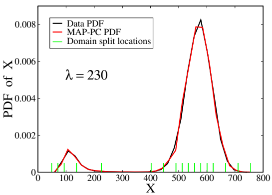
- Domain decomposition methods reduce the effect of nonlinearities/modalities
- Adaptivity criterion based on Kullback-Leibler divergence (or *relative entropy*):

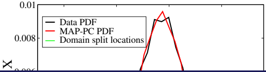
$$\rho(P_X, P_Y) = \int P(z) \log \frac{P_X(z)}{P_Y(z)} dz \simeq \frac{1}{N} \sum_{i=1}^{N} \log \frac{P_X(X_i)}{P_Y(X_i)}$$

PC Inference for fixed parameter values

- Fix the parameter λ
- Gather SSA data $\mathcal{D} = \{X_i\}_{i=1}^{N}$
- Infer the model parameters c_k's, where $X = \sum_{k=0}^{P} c_k \Psi_k(\eta)$
- If the representation is not satisfactory (see the criterion), split the data domain and proceed recursively

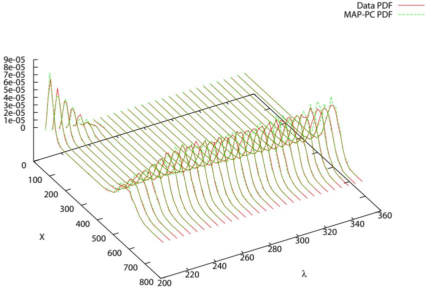


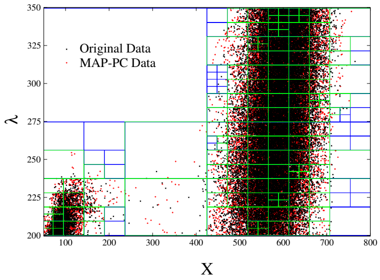




Parametric uncertainty propagation through PCE

- Postulate parametric uncertainty $\lambda = \lambda_0 + \Delta\lambda\eta_1$
- Gather two-dimensional data $\mathcal{D} = \{(X_i, \lambda_i)\}_{i=1}^{N}$
- Infer the model parameters c_k's, where $X = \sum_{k=0}^{P} c_k \Psi_k(\eta_1, \eta_2)$
- If the representation is not satisfactory (see the criterion), split the data domain and proceed recursively





Goals and Tools

$$X(t, \boldsymbol{\theta}, \boldsymbol{\lambda})$$

- Develop tools for *predictability*(λ) and *dynamical analysis*(t) of SRNs accounting for
  - Inherent stochasticity (θ)
  - Model/parameter variability (λ)
  - Limited data

$$\mathcal{D} = \{X_i\}_{i=1}^N$$

- Dynamical analysis
  - Fix λ, focus on t dependence
  - Polynomial chaos; Karhunen-Loève decomposition; Rosenblatt transformation; Data clustering

Goals and Tools

$$X(t, \boldsymbol{\theta}, \boldsymbol{\lambda})$$

- Develop tools for *predictability*$(\boldsymbol{\lambda})$ and *dynamical analysis*(t) of SRNs accounting for
  - Inherent stochasticity $(\boldsymbol{\theta})$
  - Model/parameter variability $(\boldsymbol{\lambda})$
  - Limited data

$$\mathcal{D} = \{X_i\}_{i=1}^N$$

- Dynamical analysis
  - Fix λ, focus on t dependence
  - Polynomial chaos; Karhunen-Loève decomposition; Rosenblatt transformation; Data clustering

Goals and Tools

$$X(t, \boldsymbol{\theta})$$

- Develop tools for *predictability*$(\boldsymbol{\lambda})$ and *dynamical analysis*(t) of SRNs accounting for
    - Inherent stochasticity $(\boldsymbol{\theta})$
    - Model/parameter variability $(\boldsymbol{\lambda})$
    - Limited data
$$\mathcal{D} = \{X_i\}_{i=1}^N$$
- Dynamical analysis
    - Fix $\boldsymbol{\lambda}$, focus on t dependence
    - Polynomial chaos; Karhunen-Loève decomposition; Rosenblatt transformation; Data clustering

Karhunen-Loève decomposition reduces stochastic process to a finite number of random variables

- KL decomposition:

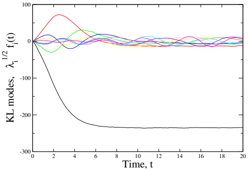
$$X(t,\theta) = \bar{X}(t) + \sum_{n=1}^{\infty} \xi_n(\theta)\sqrt{\lambda_n} f_n(t)$$

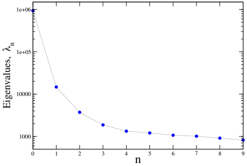
- Uncorrelated, zero-mean KL variables:

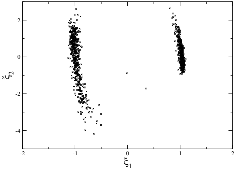
$$\langle \xi_n \rangle = 0, \qquad \langle \xi_n \xi_m \rangle = \delta_{nm}$$

- SSA(continuum) $\longleftrightarrow$ KL(discrete)

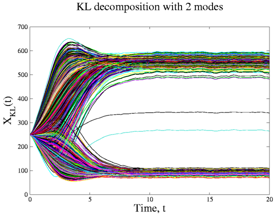
$$X(t) \longleftrightarrow \boldsymbol{\xi} = (\xi_1, \xi_2, \dots)$$

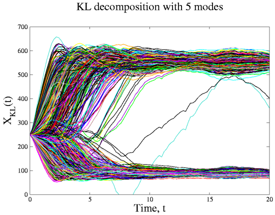


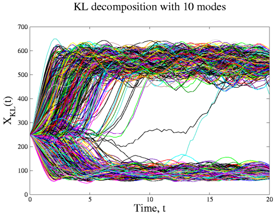


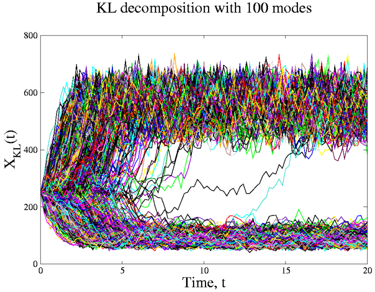


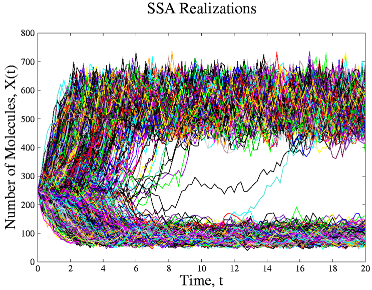
K-L decomposition captures each realization



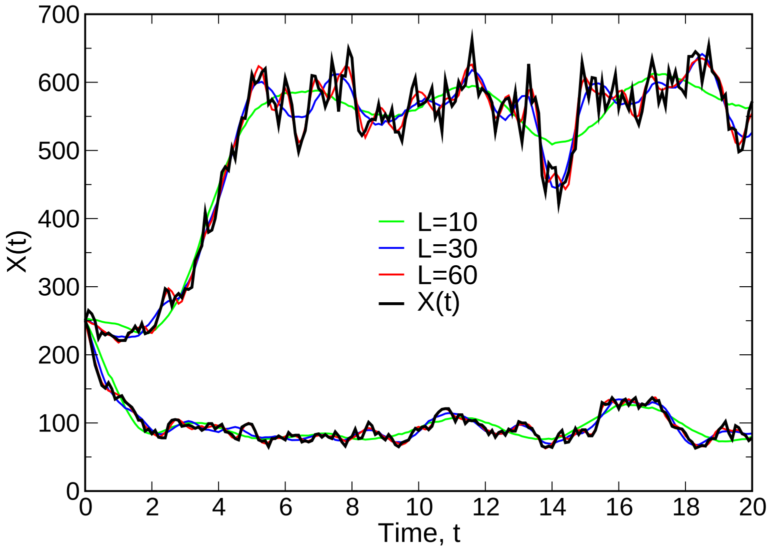








K-L decomposition captures each realization



PC expansion of a random vector

$$\boldsymbol{\xi} = \sum_{k=0}^{P} \boldsymbol{c}_k \Psi_k(\boldsymbol{\eta})$$

Galerkin projection

$$\boldsymbol{c}_k = \frac{\langle \boldsymbol{\xi} \Psi_k(\boldsymbol{\eta}) \rangle}{\langle \Psi_k^2(\boldsymbol{\eta}) \rangle}$$

is not well-defined,

since ξ and η do not belong to the same stochastic space.

PC expansion of a random vector

$$\boldsymbol{\xi} = \sum_{k=0}^{P} \boldsymbol{c}_k \Psi_k(\boldsymbol{\eta})$$

Galerkin projection

$$\boldsymbol{c}_k = \frac{\langle \boldsymbol{\xi} \Psi_k(\boldsymbol{\eta}) \rangle}{\langle \Psi_k^2(\boldsymbol{\eta}) \rangle}$$

is not well-defined,
since $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ do not belong to the same stochastic space.

Need a map $\boldsymbol{\xi} \leftrightarrow \boldsymbol{\eta}$.

Rosenblatt transformation

- Rosenblatt transformation maps any (not necessarily independent) set of random variables $(\xi_1, \ldots, \xi_n)$ to uniform i.i.d.'s $\{\eta_i\}_{i=1}^n$ (Rosenblatt, 1952).

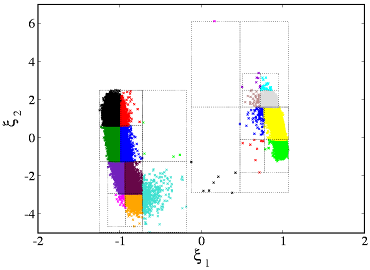
$$\begin{aligned}
\eta_1 &= F_1(\xi_1) \\
\eta_2 &= F_{2|1}(\xi_2|\xi_1) \\
\eta_3 &= F_{3|2,1}(\xi_3|\xi_2,\xi_1) \\
&\vdots \\
\eta_n &= F_{n|n-1,\ldots,1}(\xi_n|\xi_{n-1},\ldots,\xi_1)
\end{aligned}$$

- Inverse Rosenblatt transformation $\boldsymbol{\xi} = R^{-1}(\boldsymbol{\eta})$ ensures a well-defined quadrature integration

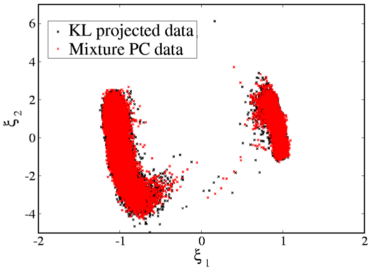
$$\langle \xi_i \Psi_k(\boldsymbol{\eta}) \rangle = \int R^{-1}(\boldsymbol{\eta})_i \Psi_k(\boldsymbol{\eta}) d\boldsymbol{\eta}$$

KL+PC+Data Partitioning represent the dynamics of a bimodal process

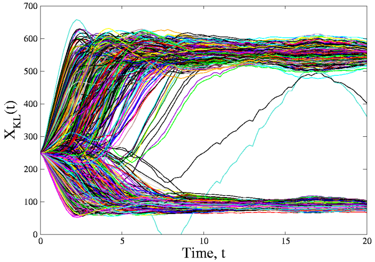
a)



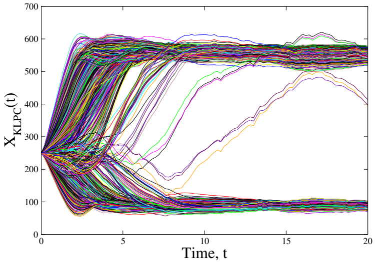
b)



KL decomposition with 5 modes



KL-PC representation, 5 KL modes, 3rd PC order



Conclusions and Future Work

- Lessons learned...
  - Bayesian methods are well-suited to deal with *intrinsic stochasticity* and *limited data*.
  - Data-based partitioning algorithms help to capture *nonlinearities* and *bimodalities*.

- Still plenty to cover...
  - Combine parametric uncertainty and time dependence
  - Sparse grid PC projection, HDMR expansion, smarter domain decomposition algorithms
  - Predict optimal partitioning
  - Direct CME solution, continuous approximations (Fokker-Planck)
  - PC with discrete random variables

Conclusions and Future Work

- Lessons learned...
  - Bayesian methods are well-suited to deal with *intrinsic stochasticity* and *limited data*.
  - Data-based partitioning algorithms help to capture *nonlinearities* and *bimodalities*.

- Still plenty to cover...
  - Combine parametric uncertainty and time dependence
  - Sparse grid PC projection, HDMR expansion, smarter domain decomposition algorithms
  - Predict optimal partitioning
  - Direct CME solution, continuous approximations (Fokker-Planck)
  - PC with discrete random variables

Details can be found at..

- K. Sargsyan, B. Debusschere, H. Najm and O. Le Maître, *"Spectral representation and reduced order modeling of the dynamics of stochastic reaction networks via adaptive data partitioning"*. SIAM Journal on Scientific Computing, accepted, 2009.

- K. Sargsyan, B. Debusschere, H. Najm and Y. Marzouk, *"Bayesian inference of spectral expansions for predictability assessment in stochastic reaction networks"*. Journal of Computational and Theoretical Nanoscience, 6:10, 2009.

Projects

- ”Stochastic Dynamical Systems: Spectral Methods for the Analysis of Dynamics and Predictability”

  supported by DOE ASCR Applied Math,
  PI: Bert Debusschere

- ”Uncertainty Quantification for Large Scale Ocean Circulation Predictions”

  supported by Sandia Seniors’ Council LDRD,
  PI: Cosmin Safta

- ”Quantifying the Margin of High-Consequence Climate Change”

  supported by DOE BER,
  Sandia-CA POC: Khachik Sargsyan

- ”Analysis of Stochasticity in Immune System Signaling Pathways”

  UTMB-Sandia Joint Institute of Biosecurity,
  PI: Bert Debusschere

Projects

- "Stochastic Dynamical Systems: Spectral Methods for the Analysis of Dynamics and Predictability"

  supported by DOE ASCR Applied Math,
  PI: Bert Debusschere

- "Uncertainty Quantification for Large Scale Ocean Circulation Predictions"

  supported by Sandia Seniors' Council LDRD,
  PI: Cosmin Safta

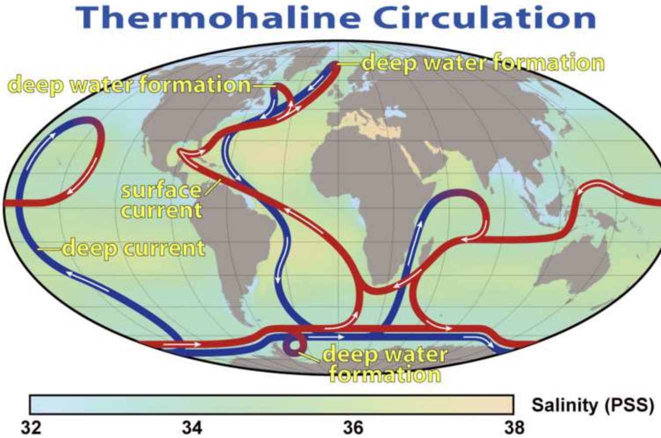
- "Quantifying the Margin of High-Consequence Climate Change"

  supported by DOE BER,
  Sandia-CA POC: Khachik Sargsyan

- "Analysis of Stochasticity in Immune System Signaling Pathways"

  UTMB-Sandia Joint Institute of Biosecurity,
  PI: Bert Debusschere

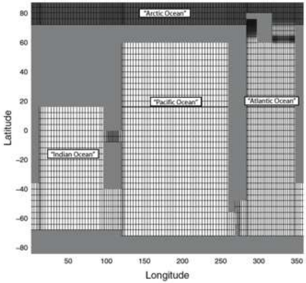
Meridional Overturning Circulation



SOURCE: HTTP://EN.WIKIPEDIA.ORG/WIKI/THERMOHALINE_CIRCULATION

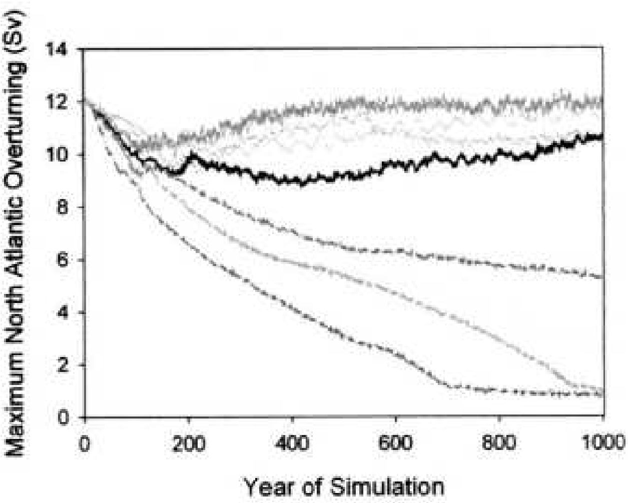
Webster et al - J. Environ. Syst. 31: 39-59, 2007

- Computational Model
  - 3D Ocean general circulation model
  - Zonally-averaged atmospheric model
  - Thermodynamic sea-ice model
  - Simplified models for river runoff

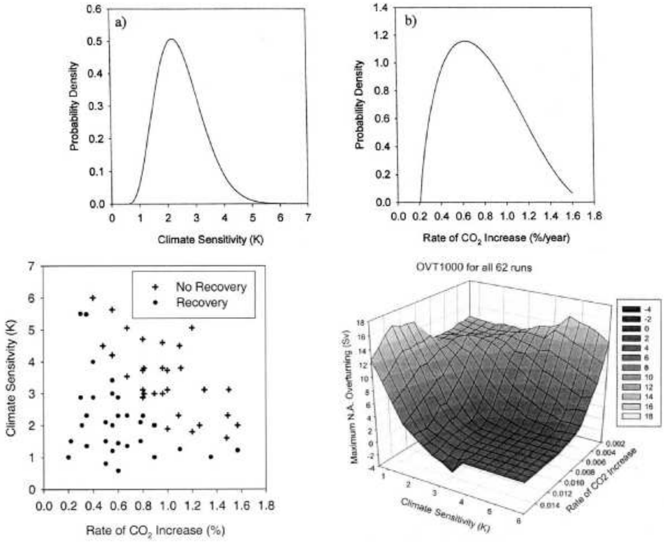


- Parameters
  - Rate of CO_2 increase (r)
  - Climate sensitivity (λ)

Webster et al - J. Environ. Syst. 31: 39-59, 2007



Webster et al - J. Environ. Syst. 31: 39-59, 2007



Bayesian Inference of the Location of Discontinuity

- Parameterize the discontinuity:

$$r \approx p_{\mathbf{c}}(\lambda) = \sum_{k=0}^{K} c_k P_k(\lambda)$$

- Approximation model:

$$\mathcal{M}_{\mathbf{c}} \equiv g(\lambda, r) = m_L + (m_R - m_L) \tanh\left(\alpha(r - p_{\mathbf{c}}(\lambda))\right)$$

- Statistical noise model:

$$\sigma(\lambda, r) \quad = \quad \sigma_L + (\sigma_R - \sigma_L) \tanh\left(\alpha(r - p_{\mathbf{c}}(\lambda))\right) + \exp\left(-\frac{(r - p_{\mathbf{c}}(\lambda))^2}{2\delta^2}\right)$$

- Likelihood function:

$$\log P(\mathcal{D}|\mathcal{M}_{\mathbf{c}}) = \sum_{i=1}^{N} \log\left(P(z_i|\mathcal{M}_{\mathbf{c}})\right) = -\sum_{i=1}^{N} \frac{(z_i - g(\lambda, r))^2}{2\sigma(\lambda, r)^2}.$$

Bayesian Inference of the Location of Discontinuity

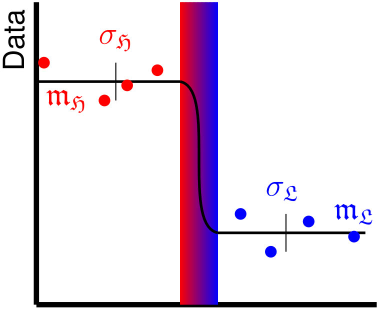
$$r \approx p_{\boldsymbol{c}}(\lambda) = \sum_{k=0}^{K} c_k P_k(\lambda)$$

"Likelihood" "Prior"

Bayes' formula:

$$P(\mathcal{M}|\mathcal{D}) = \frac{P(\mathcal{D}|\mathcal{M})P(\mathcal{M})}{P(\mathcal{D})}$$

"Posterior" "Evidence"



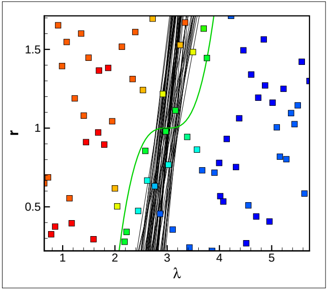
Inference of Discontinuity - 3^{rd} order polynomial

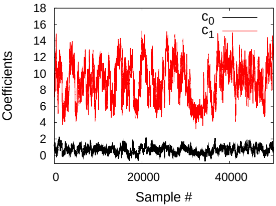
- Synthetic discontinuous data

$$z_i = (1+\sigma\xi)\tanh\left(\beta(r_i - \tilde{r}(\lambda_i))\right).$$

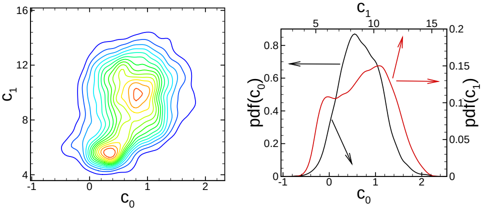
- Use straight lines to infer the discontinuity

$$\tilde{r}(\lambda) = c_0 + c_1\lambda.$$





MCMC samples



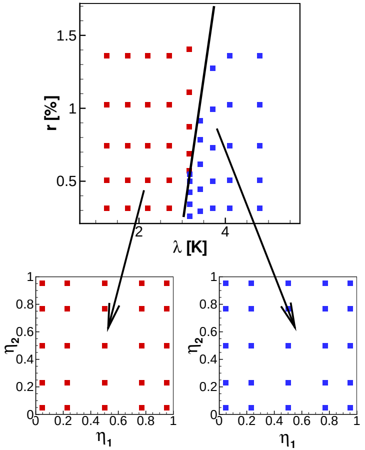
Joint and Marginal Posterior Distributions

Parameter Domain Mapping

- Use Rosenblatt Transformation (RT) to map the pair of uncertain parameters (r,λ) to i.i.d. standard random variables η_1 and η_2:

$$\begin{aligned} \lambda &= F_{\lambda}^{-1}(\eta_1), \\ r &= F_{r|\lambda}^{-1}(\eta_2|\eta_1) \end{aligned}$$

- Apply the RT mapping to both sides of the discontinuity



ROSENBLATT TRANSFORMATION: $(r, \lambda) \rightarrow (\eta_1, \eta_2)$

PC expansion, averaged over discontinuity curves

- PC expansion for each discontinuity curve sample:

$$Z_{\boldsymbol{c}}^{L,R}(\lambda, r) = \tilde{Z}_{\boldsymbol{c}}(\eta_1, \eta_2) = \sum_{p=0}^{P} z_p \Psi_p^{(2)}(\eta_1, \eta_2)$$

- Model expansion depends on the parameter location:

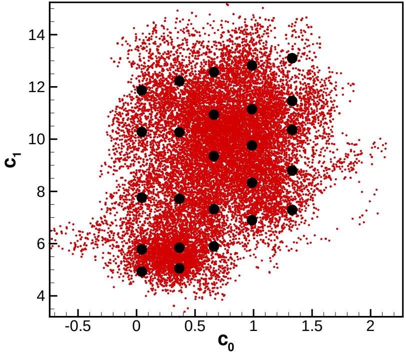
$$Z_{\boldsymbol{c}}(\lambda, r) = \begin{cases} Z_{\boldsymbol{c}}^{L}(\lambda, r) & \text{if } (\lambda, r) \in D_L \\ Z_{\boldsymbol{c}}^{R}(\lambda, r) & \text{if } (\lambda, r) \in D_R \end{cases}.$$

- Average over all PC expansions via RT:

$$\hat{Z}(\lambda, r) = \int_C p(\boldsymbol{c}) Z_{\boldsymbol{c}}(\lambda, r) d\boldsymbol{c} = \int_{[0,1]^{K+1}} Z_{R^{-1}(\vec{\eta})}(\lambda, r) d\vec{\eta}$$

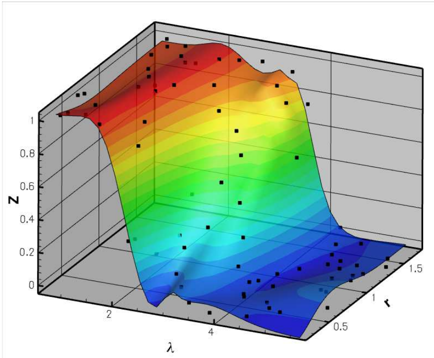
Discontinuous data represented well with the averaged PC

QUADRATURE IN (c_0, c_1) DOMAIN



Quadrature points necessary for integrating/averaging over all discontinuity curves

PCE IN (η_1, η_2) DOMAIN



Averaged-PC representation through discontinuous data

As a conclusion..

- A methodology for uncertainty quantification in climate models with limited data and discontinuities was proposed
  - Bayesian approach to detect and parameterize the discontinuity as well as the uncertainty associated with it.
  - Rosenblatt transformation maps each of the irregular domains to rectangular ones where the application of the local spectral methods of uncertainty propagation is feasible.

- Work-in-progress paper accepted to “Knowledge Discovery from Climate Data: Prediction, Extremes, and Impacts” - 9th IEEE International Conference on Data Mining
- Abstract submitted to “Uncertainty Quantification and its Application to Climate Change” - American Geophysical Union 2009 Fall Meeting

As a conclusion..

- A methodology for uncertainty quantification in climate models with limited data and discontinuities was proposed
  - Bayesian approach to detect and parameterize the discontinuity as well as the uncertainty associated with it.
  - Rosenblatt transformation maps each of the irregular domains to rectangular ones where the application of the local spectral methods of uncertainty propagation is feasible.

- Demonstrate the methodology with data from climate research groups (in touch with MIT group)
- Explore generalizations of this approach for situations where climate model data is not available for Gaussian quadrature

Projects

- ”Stochastic Dynamical Systems: Spectral Methods for the Analysis of Dynamics and Predictability”

  supported by DOE ASCR Applied Math,
  PI: Bert Debusschere

- ”Uncertainty Quantification for Large Scale Ocean Circulation Predictions”

  supported by Sandia Seniors’ Council LDRD,
  PI: Cosmin Safta

- ”Quantifying the Margin of High-Consequence Climate Change”

  supported by DOE BER,
  Sandia-CA POC: Khachik Sargsyan

- ”Analysis of Stochasticity in Immune System Signaling Pathways”

  UTMB-Sandia Joint Institute of Biosecurity,
  PI: Bert Debusschere

Projects

- ”Stochastic Dynamical Systems: Spectral Methods for the Analysis of Dynamics and Predictability”

  supported by DOE ASCR Applied Math,
  PI: Bert Debusschere

- ”Uncertainty Quantification for Large Scale Ocean Circulation Predictions”

  supported by Sandia Seniors’ Council LDRD,
  PI: Cosmin Safta

- ”Quantifying the Margin of High-Consequence Climate Change”

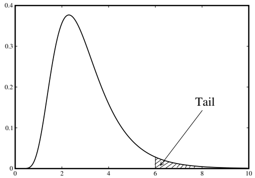
  supported by DOE BER,
  Sandia-CA POC: Khachik Sargsyan

- ”Analysis of Stochasticity in Immune System Signaling Pathways”

  UTMB-Sandia Joint Institute of Biosecurity,
  PI: Bert Debusschere

UQ of High-Consequence Climate Events

- Develop advance UQ tools that target “tail” events
  - “Tails” are low-probability, high-consequence events
  - Current UQ methods do not properly capture the “tails”



- Methods proposed
  - Surrogate modeling via PC expansions
  - Alternate PC bases

Really big picture

- Uncertainty Quantification and Data Assimilation go hand in hand
  - Spectral methods as the most appropriate tool for forward UQ
  - Bayesian methods are well-suited for handling inverse problems
- Relevant application areas
  - *Stochastic* chemical kinetics
    - Gene regulation, immune system signaling, bacterial/viral behavior
    - Interfacial electrochemistry, electrical storage
  - Climate models

Really big picture

- Uncertainty Quantification and Data Assimilation go hand in hand
  - Spectral methods as the most appropriate tool for forward UQ
  - Bayesian methods are well-suited for handling inverse problems
- Relevant application areas
  - *Stochastic* chemical kinetics
    - Gene regulation,
      immune system signaling,
      bacterial/viral behavior
    - Interfacial electrochemistry,
      electrical storage
  - Climate models

Acknowledgements

- Bert Debusschere (8351)
- Habib Najm (8351)
- Cosmin Safta (8964)
- Youssef Marzouk (MIT)

- DOE ASCR
- Sandia LDRD
- DOE BER

Thank You!

Adaptivity criterion for domain decomposition

Data: $\mathcal{D} = \{X_i\}_{i=1}^N$

Model: $X \simeq \sum_{k=0}^P c_k \Psi_k(\eta) = g_{\mathcal{D}}(\eta)$

MAP-PC samples: $\{Y_i\}_{i=1}^N$, where $Y_i = g_{\mathcal{D}}(\eta_i)$

- Log-likelihood:

$$\log L = \log P(\text{Data}|\text{Model}) = \sum_{i=1}^N \log P_Y(X_i)$$

- Target log-likelihood (the *perfect match* log-likelihood, i.e. for $\{Y_i\}_{i=1}^N = \mathcal{D}$):

$$\log L_T = \sum_{i=1}^N \log P_X(X_i)$$

- Kullback-Leibler divergence (or *relative entropy*):

$$\rho(P_X, P_Y) = \int P(z) \log \frac{P_X(z)}{P_Y(z)} dz \simeq \frac{1}{N} \sum_{i=1}^N \log \frac{P_X(X_i)}{P_Y(X_i)}$$

Karhunen-Loève decomposition reduces stochastic process to a finite number of random variables

- Separate the average:

$$X_0(t,\theta) = X(t,\theta) - \bar{X}(t)$$

- The covariance function is symmetric, bounded and positive definite. Hence, it can be expanded as a sum

$$C(t_1,t_2) = \langle X_0(t_1,\theta)X_0(t_2,\theta)\rangle = \sum_{n=1}^{\infty} \lambda_n f_n(t_1) f_n(t_2)$$

- Positive eigenvalues:

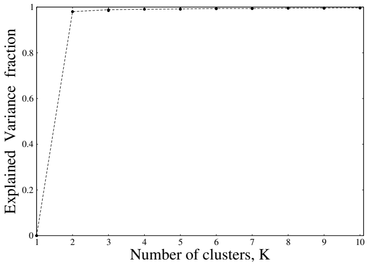
$$\int_0^T C(t_1,t_2) f_n(t_1)\,dt_1 = \lambda_n f_n(t_2).$$

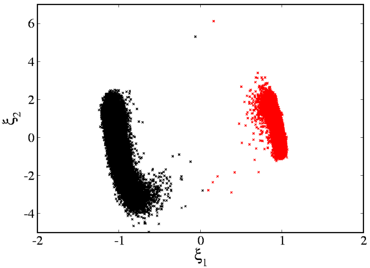
- KL decomposition:

$$X(t,\theta) = \bar{X}(t) + \sum_{n=1}^{\infty} \xi_n(\theta)\sqrt{\lambda_n} f_n(t)$$

Clustering precedes data domain decomposition

- Finite number of KL variables: $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_L)$
- Multidimensional data: $\{\boldsymbol{\xi}^{(i)}\}_{i=1}^{N}$
- K-Center clustering (Gonzalez, 1985)
- Distance measure scaled with KL eigenvalues
- 'Elbow' criterion with Explained Variance to pick the optimal number of clusters
- E.V. $=$ Variance of dataset with all points replaced with their corresponding cluster's center





The final representation is a Mixture PC model

- Divide data into K partitions with fractions p_j:

$$p_1 + p_2 + \cdots + p_K = 1$$

- Find PC expansion for ξ in each partition:

$$\xi_{PC}^{(j)} = \sum_{k=0}^{P} c_k^{(j)} \Psi_k(\zeta^{(j)})$$

- Superpose the results to obtain PC mixture model (assuming data points are of equal importance/weight):

$$\xi = \xi_{PC}^{(j)} \text{ w. prob. } p_j$$

- Probability distribution function is a mixture of PC PDFs:

$$\mathrm{Pdf}_\xi(x) = p_1 \mathrm{Pdf}_{\xi_{PC}^{(1)}}(x) + \cdots + p_K \mathrm{Pdf}_{\xi_{PC}^{(K)}}(x)$$

Dynamical Analysis: Big Picture

Fix the parameter $X(t, \theta, \mathbf{\Lambda}) \equiv X(t, \theta)$

SSA $\longrightarrow$ KL $\longrightarrow$ PCE

Random process $\longrightarrow$ L random v. $\longrightarrow$ L(P+1) deterministic v.

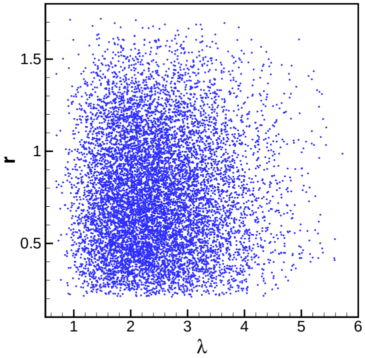
$$X(t, \theta) \longrightarrow \xi_i(\theta)(i = \overline{1, L}) \longrightarrow c_{ik}(i = \overline{1, L}, k = \overline{0, P})$$

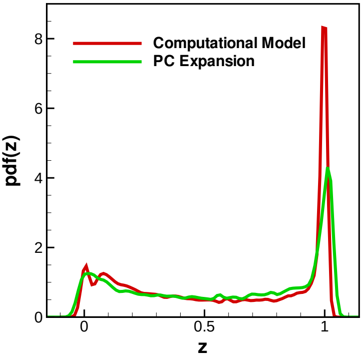
$$X(t, \theta) - \bar{X}(t) \simeq \sum_{i=1}^{L} \xi_i(\theta) \sqrt{\lambda_i} f_i(t) \simeq \sum_{i=1}^{L} \left(\sum_{k=0}^{P} c_{ik} \Psi_k(\boldsymbol{\eta}) \right) \sqrt{\lambda_i} f_i(t)$$

SSA $\longrightarrow$ KL : Karhunen-Loève (KL) decomposition
of the stochastic process

KL $\longrightarrow$ PCE: Polynomial Chaos expansion
of each KL random variable

Estimate Climate Model PDF





- Samples from the joint pdf(r,λ) are used to estimate the density of the surrogate climate model output (z).