SAND2012-2191C

Uncertainty Quantification in Computational Models

Habib N. Najm

hnnajm@sandia.gov

Sandia National Laboratories,
Livermore, CA

PNNL seminar, Mar. 2012

Acknowledgement

B.J. Debusschere, R.D. Berry, K. Sargsyan, C. Safta
— Sandia National Laboratories, Livermore, CA
R.G. Ghanem — U. South. California, Los Angeles, CA
O.M. Knio — Duke Univ., Durham, NC
O.P. Le Maître — CNRS, Paris, France
Y.M. Marzouk — Mass. Inst. of Tech., Cambridge, MA

This work was supported by:

- US Department of Energy (DOE), Office of Advanced Scientific Computing Research (ASCR), Scientific Discovery through Advanced Computing (SciDAC)
- DOE Office of Basic Energy Sciences (BES) Division of Chemical Sciences, Geosciences, and Biosciences.
- DOE ASCR Applied Mathematics program.
- DOE Office of Biological and Environmental Research, Climate and Environmental Sciences Division.

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

The Case for Uncertainty Quantification

UQ is needed in:

- Assessment of confidence in computational predictions
- Validation and comparison of scientific/engineering models
- Design optimization
- Use of computational predictions for decision-support
- Assimilation of observational data and model construction
- Multiscale and multiphysics model coupling

Overview of UQ Methods

Estimation of model/parametric uncertainty

- Expert opinion, data collection
- Regression analysis, fitting, parameter estimation
- Bayesian inference of uncertain models/parameters

Forward propagation of uncertainty in models

- Local sensitivity analysis (SA) and error propagation
- Fuzzy logic; Evidence theory — interval math
- Probabilistic framework — Global SA / stochastic UQ
  - Random sampling, statistical methods
  - Galerkin methods
    - Polynomial Chaos (PC) — intrusive/non-intrusive
  - Collocation, interpolants, regression, fitting ... PC/other

Polynomial Chaos Methods for UQ

- Model uncertain quantities as random variables (RVs)
- Given a *germ* $\boldsymbol{\xi}(\omega) = \{\xi_1, \cdots, \xi_n\}$ – a set of *i.i.d.* RVs
  - where density of $\boldsymbol{\xi}$ is uniquely determined by its moments
- Any RV in $L^2(\Omega, \mathfrak{S}(\boldsymbol{\xi}), P)$ can be written as a Polynomial Chaos Expansion (PCE), thus:

$$u(\mathbf{x}, t, \omega) = f(\mathbf{x}, t, \boldsymbol{\xi}) \simeq \sum_{k=0}^{P} u_k(\mathbf{x}, t) \Psi_k(\boldsymbol{\xi}(\omega))$$

– $u_k(\mathbf{x}, t)$ are mode strengths
– $\Psi_k()$ are functions orthogonal w.r.t. the density of $\boldsymbol{\xi}$

- with dimension n and order p:

$$P + 1 = \frac{(n+p)!}{n!p!}$$

Orthogonality

By construction, the functions $\Psi_k()$ are orthogonal with respect to the density of ξ

$$u_k(\mathbf{x}, t) = \frac{\langle u\Psi_k \rangle}{\langle \Psi_k^2 \rangle} = \frac{1}{\langle \Psi_k^2 \rangle} \int u(\mathbf{x}, t; \lambda(\boldsymbol{\xi}))\, \Psi_k(\boldsymbol{\xi}) p_{\boldsymbol{\xi}}(\boldsymbol{\xi}) d\boldsymbol{\xi}$$

Examples:

- Hermite polynomials with Gaussian basis
- Legendre polynomials with Uniform basis, ...
- Global versus Local PC methods
  - Adaptive domain decomposition of the support of ξ

Essential Use of PC in UQ

Strategy:

- Represent model parameters/solution as random variables
- Construct PCEs for uncertain parameters
- Evaluate PCEs for model outputs

Advantages:

- Computational efficiency
- Sensitivity information

Requirement:

- Random variables in L^2, i.e. with finite variance

Intrusive PC UQ: A direct *non-sampling* method

- Given model equations: $\mathcal{M}(u(\mathbf{x},t);\lambda)=0$
- Express uncertain parameters/variables using PCEs

$$u=\sum_{k=0}^{P} u_k\Psi_k;\quad \lambda=\sum_{k=0}^{P}\lambda_k\Psi_k$$

- Substitute in model equations; apply Galerkin projection
- New set of equations: $\mathcal{G}(U(\mathbf{x},t),\Lambda)=0$
  - with $U=[u_0,\dots,u_P]^T$, $\Lambda=[\lambda_0,\dots,\lambda_P]^T$
- Solving this system *once* provides the full specification of uncertain model ouputs

Intrusive Galerkin PC ODE System

$$\frac{du}{dt} = f(u;\lambda)$$

$$\lambda = \sum_{i=0}^{P} \lambda_i \Psi_i \qquad u(t) = \sum_{i=0}^{P} u_i(t)\Psi_i$$

$$\frac{du_i}{dt} = \frac{\langle f(u;\lambda)\Psi_i\rangle}{\langle \Psi_i^2 \rangle} \qquad i = 0,\ldots,P$$

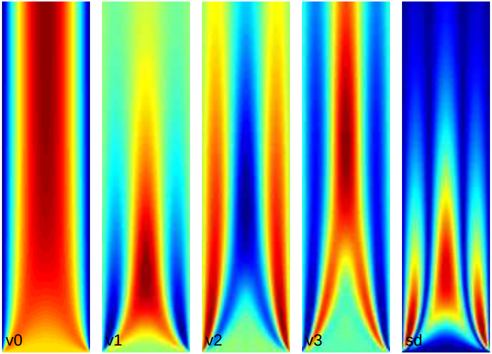
Say $f(u;\lambda) = \lambda u$, then

$$\frac{du_i}{dt} = \sum_{p=0}^{P}\sum_{q=0}^{P} \lambda_p u_q C_{pqi}, \quad i = 0,\cdots,P$$

where the tensor $C_{pqi} = \langle \Psi_p \Psi_q \Psi_i \rangle / \langle \Psi_i^2 \rangle$ is readily evaluated

Laminar 2D Channel Flow with Uncertain Viscosity

- Incompressible flow
- Viscosity PCE
  - $\nu = \nu_0 + \nu_1 \xi$
- Streamwise velocity
  - $\mathrm{v} = \sum_{i=0}^{P} \mathrm{v}_i \Psi_i$
  - v_0: mean
  - v_i: i-th order mode
  - $\sigma^2 = \sum_{i=1}^{P} \mathrm{v}_i^2 \langle \Psi_i^2 \rangle$



v_0 v_1 v_2 v_3 σ

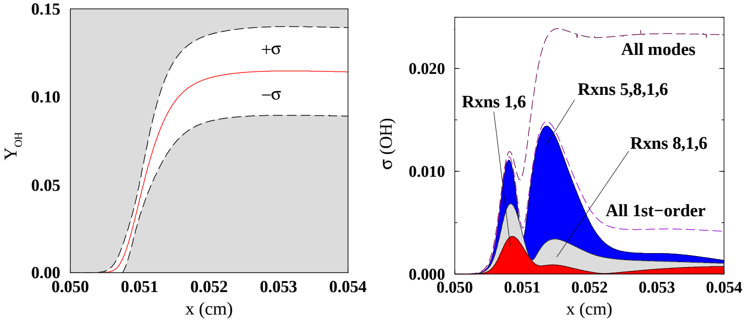
Non-intrusive Spectral Projection (NISP) PC UQ

- *Sampling*-based
- Relies on black-box utilization of the computational model
- Evaluate projection integrals *numerically*
- For any model output of interest $\phi(\boldsymbol{x}, t; \lambda)$:

$$\phi_k(\boldsymbol{x}, t) = \frac{1}{\langle \Psi_k^2 \rangle} \int \phi(\boldsymbol{x}, t; \lambda(\boldsymbol{\xi}))\, \Psi_k(\boldsymbol{\xi}) p_{\boldsymbol{\xi}}(\boldsymbol{\xi}) d\boldsymbol{\xi}, \quad k = 0, \ldots, P$$

- Integrals can be evaluated using
  - A variety of (Quasi) Monte Carlo methods
    - Slow convergence; $\sim$ indep. of dimensionality
  - Quadrature/Sparse-Quadrature methods
    - Fast convergence; depends on dimensionality

1D H_2-O_2 SCWO Flame NISP UQ/Chemkin-Premix



- Fast growth in OH uncertainty in the primary reaction zone
- Constant uncertainty and mean of OH in post-flame region
- Uncertainty in pre-exponential of Rxn.5 (H_2O_2+OH=H_2O+HO_2) has largest contribution to uncertainty in predicted OH

Other non-intrusive methods

- Response surface employing PC or other functional basis
- Collocation: Fit interpolant to samples
  - Oscillation concern
- Regression: Estimate best-fit response surface
  - Least-squares
  - Bayesian inference
- Useful when quadrature methods are infeasible, e.g. when
  - Can't choose sample locations; samples given *a priori*
  - Can't take enough samples
  - Forward model is noisy

PCE Construction for Noisy Functions

- Quadrature formulae presume a degree of smoothness
  - No convergence for a noisy function

$$u_k = \frac{1}{\langle \Psi_k^2 \rangle} \int u(\lambda(\xi))\, \Psi_k(\xi) p_\xi(\xi) d\xi, \quad k = 0, \ldots, P$$

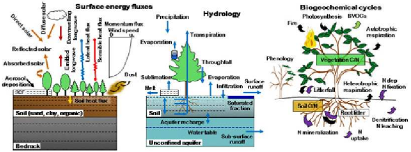
- Sparse-Quadrature formulae are *ill-conditioned* and highly-sensitive to noise
  - No convergence with order
  - Error grows with increased dimensionality
- Options in the presence of noise:
  - RMS fitting for PC coefficients
  - Bayesian inference of PC coefficients

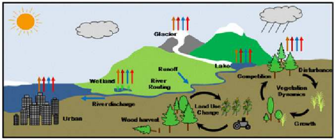
Challenges in PC UQ – High-Dimensionality

- Dimensionality n of the PC basis: $\boldsymbol{\xi} = \{\xi_1, \ldots, \xi_n\}$
  - number of degrees of freedom
  - $P + 1 = (n+p)!/n!p!$ grows fast with n
- Impacts:
  - Size of intrusive system
  - # non-intrusive (sparse) quadrature samples
- Generally $n \approx$ number of uncertain parameters
- Reduction of n:
  - Sensitivity analysis
  - Dependencies/correlations among parameters
  - Dominant eigenmodes of random fields
  - Manifold learning: Isomap, Diffusion maps
  - Sparsification: Compressed Sensing, LASSO

UQ in the Community Land Model (CLM)

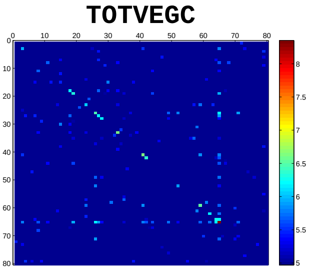
- Land component of CESM
- Spatial heterogeneity of the land surface
- A single-site, 1000-yr simulation:
  10 hr/1 CPU
- 80 input parameters
- Need to eliminate unimportant parameters

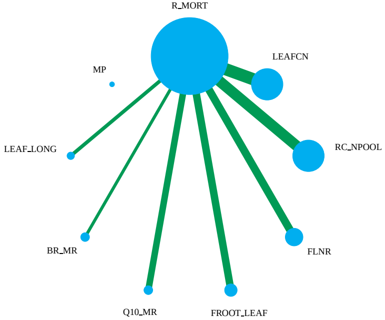




http://www.cesm.ucar.edu/models/clm/

Bayesian Compressed Sensing (BCS) CLM Analysis



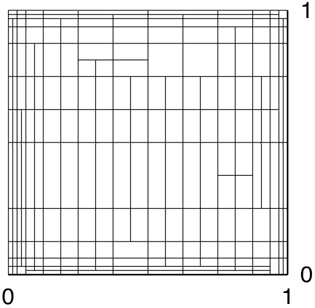


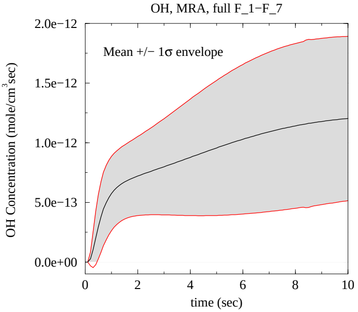
- 10^4 random sample model runs
- BCS fit of 80-D Legendre-Uniform PC
  - Laplace priors & evidence maximization (Babacan, 2010)
- Eliminate unimportant terms; discover sparse PCE
- Global sensitivity analysis

Challenges in PC UQ – Non-Linearity

- Bifurcative response at critical parameter values
  - Rayleigh-Bénard convection
  - Transition to turbulence
  - Chemical ignition
- Discontinuous $u(\lambda(\boldsymbol{\xi}))$
  - Failure of global PCEs in terms of smooth $\Psi_k()$
  - $\Leftrightarrow$ failure of Fourier series in representing a step function
- Local PC methods
  - Subdivide support of $\lambda(\boldsymbol{\xi})$ into regions of smooth $u \circ \lambda(\boldsymbol{\xi})$
  - Employ PC with compact support basis on each region
  - A spectral-element vs. spectral construction
  - Domain mapping

Multi-Block Multiwavelet PC UQ in Ignition

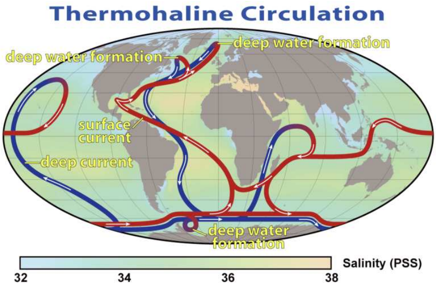


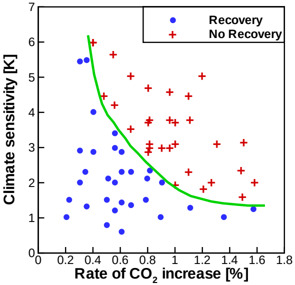


- H_2-O_2 supercritical water oxidation model
- Empirically-based uncertainty in all 7 reactions
- Adaptive refinement of MW block decomposition

(Le Maître, 2004, 2007)

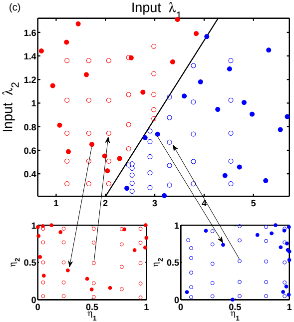
Uncertainty in Discontinuous Climate Response

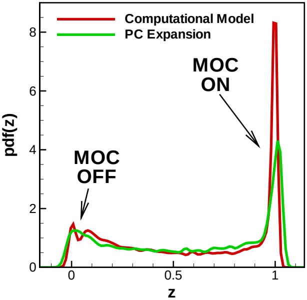




- Atlantic meridional ocean circulation (AMOC)
- Predicted response to increasing CO_2 (Webster, 2007)
- Circulation ON/OFF response over parameter space
  - Rate of CO_2 increase
  - Climate sensitivity

Domain Mapping for Discontinuous Response





- Initial set of computational samples
- Discover uncertain discontinuity with Bayesian inference
- Map sub-domains to unit hypercubes; Rosenblatt transform
- PC quadrature in mapped domains; map back
- Marginalize over uncertain curve (Sargsyan, 2012)

Challenges in PC UQ – Time Dynamics

- Systems with limit-cycle or chaotic dynamics
- Large amplification of phase errors over long time horizon
- PC order needs to be increased in time to retain accuracy
- Time shifting/scaling remedies

- Futile to attempt representation of detailed turbulent velocity field $\boldsymbol{v}(\boldsymbol{x}, t; \lambda(\boldsymbol{\xi}))$ as a PCE
  - Fast loss of correlation due to energy cascade
  - Problem studied in 60's and 70's
- Focus on flow statistics, *e.g.* Mean/RMS quantities
  - Well behaved
  - Argues for non-intrusive methods with DNS/LES of turbulent flow

Bayes formula for Parameter Inference

- Data Model (fit model + noise): $y = f(\lambda) + \epsilon$
- Bayes Formula:

$$p(\lambda, y) = p(\lambda|y)p(y) = p(y|\lambda)p(\lambda)$$

$$\underbrace{p(\lambda|y)}_{\text{Posterior}} = \frac{\overbrace{p(y|\lambda)}^{\text{Likelihood}}\ \overbrace{p(\lambda)}^{\text{Prior}}}{\underbrace{p(y)}_{\text{Evidence}}}$$

- Prior: knowledge of λ prior to data
- Likelihood: forward model and measurement noise
- Posterior: combines information from prior and data
- Evidence: normalizing constant for present context

Exploring the Posterior

- Given any sample λ, the un-normalized posterior probability can be easily computed

$$p(\lambda|y) \propto p(y|\lambda)p(\lambda)$$

- Explore posterior w/ Markov Chain Monte Carlo (MCMC)
  - Metropolis-Hastings algorithm:
    - Random walk with proposal PDF & rejection rules
  - Computationally intensive, $\mathcal{O}(10^5)$ samples
  - Each sample: evaluation of the forward model
    - Surrogate models
- Evaluate moments/marginals from the MCMC statistics

Surrogate Models for Bayesian Inference

- Need an inexpensive response surface for
  - Observables of interest y
  - as functions of parameters of interest x
- Gaussian Process (GP) surrogate
  - GP goes through all data points with probability 1.0
  - Uncertainty between the points
- Fit a convenient polynomial to $y = f(x)$
  - over the range of uncertainty in x

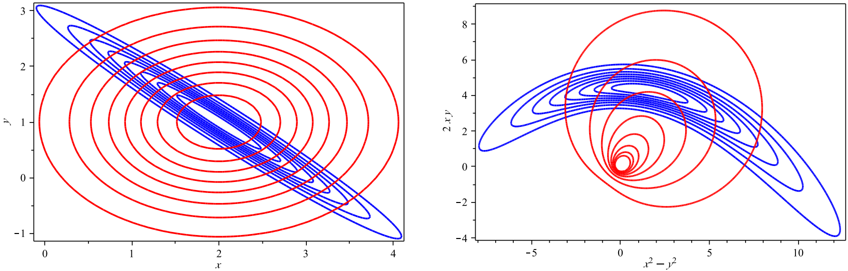
  - Employ a number of samples (x_i, y_i)
  - Fit with interpolants, regression, ... global/local
  - With uncertain x :
    - Construct Polynomial Chaos response surface

Marzouk *et al.* 2007; Marzouk & Najm, 2009

Uncertainty in Model Inputs

- Probabilistic UQ requires specification of uncertain inputs
- Require joint PDF on input space
- PDF can be found given data
- Typically such PDFs are not available from the literature
  - Summary information, e.g. nominals and bounds, is usually available
- Uncertainty in computational predictions can depend strongly on detailed structure of the missing parametric PDF
- Need a procedure to reconstruct a PDF consistent with available information in the absence of the raw data
  - “Data Free" Inference (DFI) (Berry *et al.*, JCP 2012)

The strong role of detailed input PDF structure

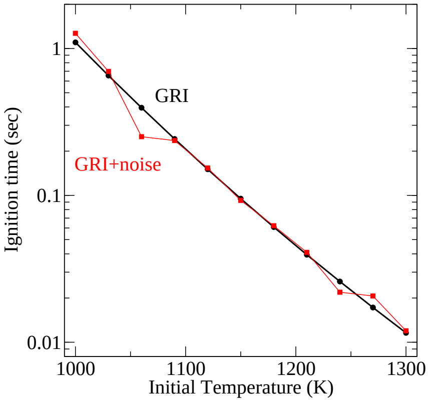


- Simple nonlinear algebraic model $(u,v) = (x^2 - y^2, 2xy)$
- Two input PDFs, $p(x,y)$
  - same nominals/bounds
  - different correlation structure
- Drastically different output PDFs
  - different nominals and bounds

Generate ignition "data" using a detailed model+noise

- Ignition using a detailed chemical model for methane-air chemistry
- Ignition time versus Initial Temperature
- Multiplicative noise error model
- 11 data points:

$$
\begin{aligned}
d_i &= t_{\text{ig},i}^{\text{GRI}}(1 + \sigma \epsilon_i) \\
\epsilon &\sim N(0, 1)
\end{aligned}
$$



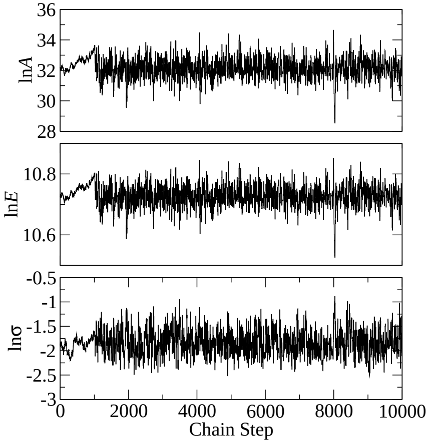
Fitting with a simple chemical model

- Fit a global single-step irreversible chemical model

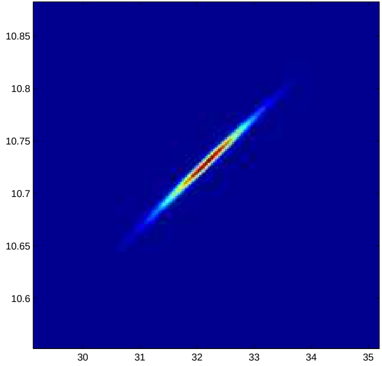
$$CH_4 + 2O_2 \rightarrow CO_2 + 2H_2O$$

$$\begin{aligned} \mathfrak{R} &= [CH_4][O_2]k_f \\ k_f &= A\exp(-E/R^oT) \end{aligned}$$

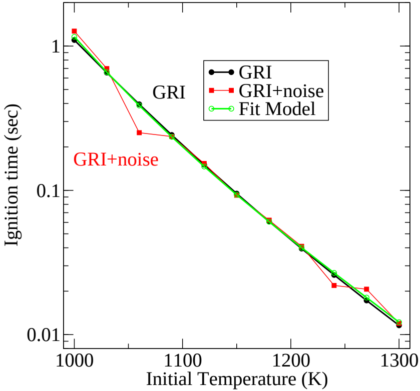
- Infer 3-D parameter vector $(\ln A, \ln E, \ln\sigma)$
- Good mixing with adaptive MCMC when start at MLE



Bayesian Inference Posterior and Nominal Prediction



Marginal joint posterior on $(\ln A, \ln E)$ exhibits strong correlation



Nominal fit model is consistent with the true model

Central Challenge for UQ in Chemical Kinetic Models

- Need joint PDF on model parameters for forward UQ
- Joint PDF structure is crucial
- Joint PDF not available for chemical kinetic parameters
- At best, have
  - Nominal parameter values
  - Bounds, *e.g.* marginal 5%, 95% quantiles
- PDF ***can*** be constructed by repeating experiments or access to original raw data
  - Neither is feasible
- Is there a way to construct an approximate PDF ***without*** access to raw data?
  - Yes!

Data Free Inference (DFI) (Berry *et al.*, JCP 2012)

- Intuition: In the absence of data, the structure of the fit model, combined with the nominals and bounds, implicitly inform the correlation between the parameters
- Goal: Make this information *explicit* in the joint PDF
- DFI: discover a consensus joint PDF on the parameters consistent with <u>given</u> information
- Method construction is closely related to
  - Maximum entropy
  - Imputation and Bayesian missing data problems

Data Free Inference Challenge

Discarding initial data, reconstruct marginal $(\ln A, \ln E)$ posterior using the following information

- Form of fit model
- Range of initial temperature
- Nominal fit parameter values of $\ln A$ and $\ln E$
- Marginal 5% and 95% quantiles on $\ln A$ and $\ln E$

Further, for now, presume

- Multiplicative Gaussian errors
- $N = 8$ data points

DFI Algorithm Structure

Basic idea:

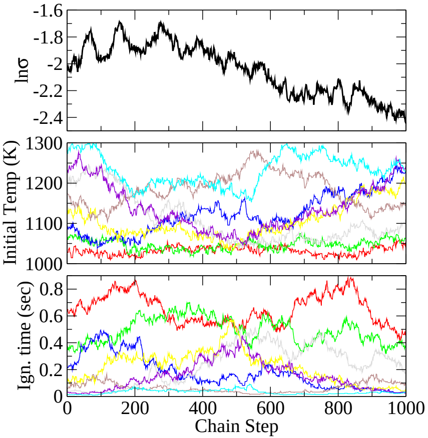
- Explore the space of hypothetical data sets
  - MCMC chain on the data
  - Each state defines a data set
- For each data set:
  - MCMC chain on the parameters
  - Evaluate statistics on resulting posterior
  - Accept data set if posterior is consistent with given information
- Evaluate pooled posterior from all acceptable posteriors
  Logarithmic pooling:

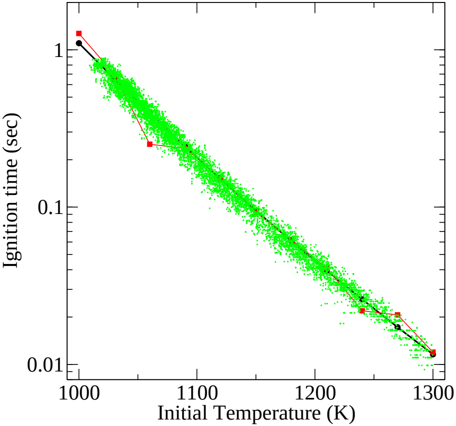
$$p(\lambda|y) = \left[\prod_{i=1}^{K} p(\lambda|y_i)\right]^{1/K}$$

DFI Uses two nested MCMC chains

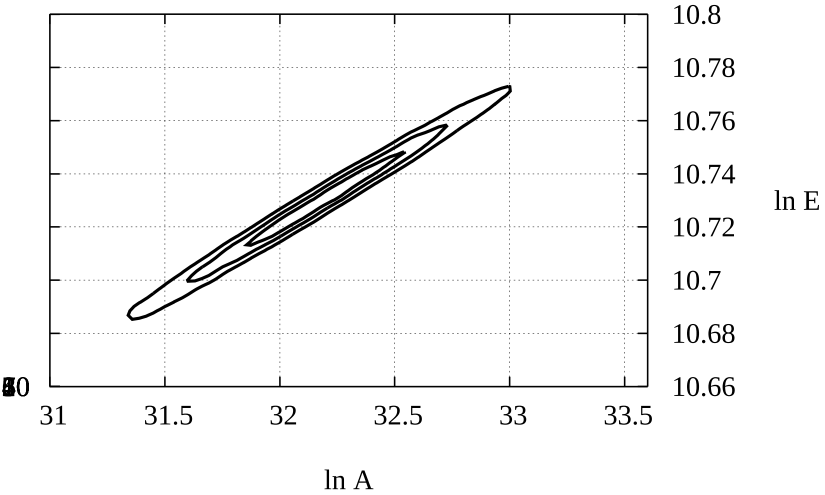
- An outer chain on the data, $(2N+1)$–dimensional
  - N data points (x_i, y_i) + σ
  - Likelihood function captures constraints on parameter nominals+bounds
- An inner chain on the model parameters
  - Conventional MCMC for parameter estimation
  - Likelihood based on fit-model
- Computationally challenging
  - Single-site update on outer chain
  - Adaptive MCMC on inner chain
  - Run multiple outer chains in parallel, and aggregate resulting acceptable data sets

Short sample from outer/data chain

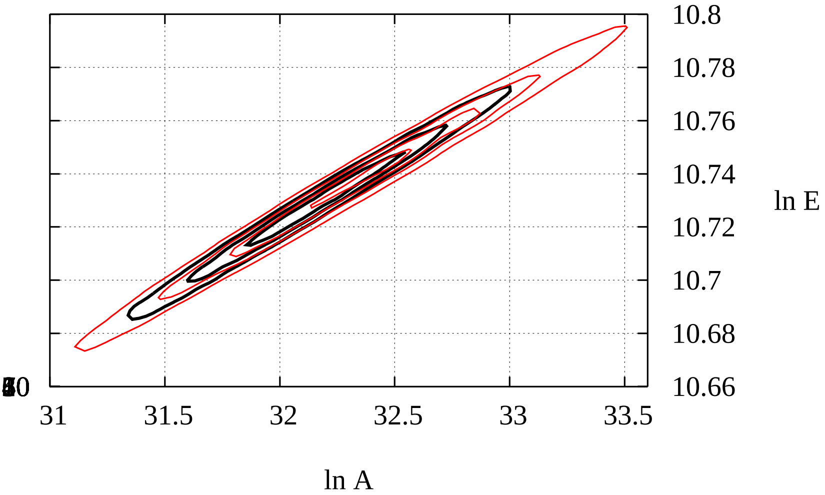




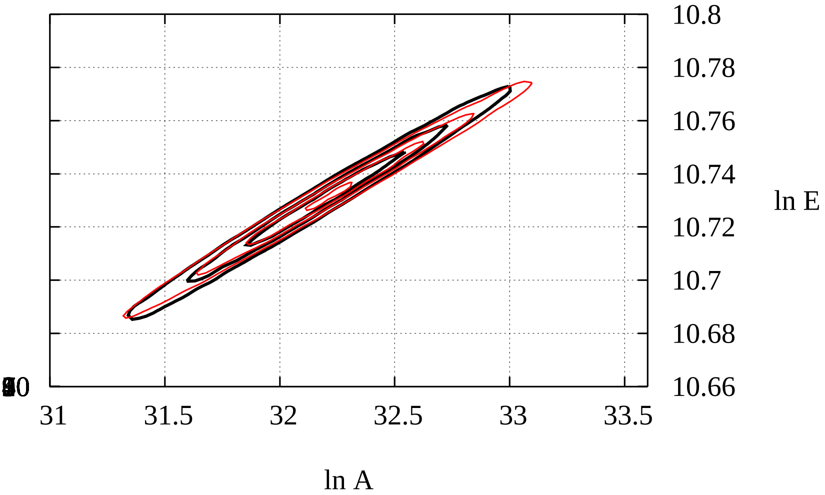
Reference Posterior – based on actual data



Ref + DFI posterior based on a 1000-long data chain



Ref + DFI posterior based on a 5000-long data chain



Closure

- Probabilistic UQ framework
  - PC representation of random variables
  - Utility in forward UQ
    - Intrusive PC methods
    - Non-intrusive methods
- Challenges
  - High Dimensionality
  - Non-linearity
  - Long term dynamics
- Need for probabilistic characterization of uncertain inputs
  - Correlations important for uncertainty in predictions
  - Discover joint PDF consistent with available information