SAND2015-6088C

Quantification of Uncertainty in Extreme Scale Computations

www.quest-scidac.org

Habib N. Najm

hnnajm@sandia.gov
Sandia National Laboratories
Livermore, CA

And the QUEST team at large

2015 SciDAC-3 PI Meeting
July 22 – 24, 2015
Washington, DC

Acknowledgement

QUEST Team:

SNL	M. Eldred, B. Debusschere, J. Jakeman, K. Chowdhary, C. Safta, K. Sargsyan, P. Rai
USC	R. Ghanem
Duke	O. Knio, O. Le Maître, J. Winokur, G. Li
UT	O. Ghattas, R. Moser, C. Simmons, A. Alexanderian
LANL	J. Gattiker, D. Higdon, E. Lawrence, S. Bhat
MIT	Y. Marzouk, D. Bigoni, T. Cui, M. Parno

This work was supported by:

- US Department of Energy (DOE), Office of Advanced Scientific Computing Research (ASCR), Scientific Discovery through Advanced Computing (SciDAC)

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

High Dimensionality and UQ

- Dimensionality of UQ problem is the number of degrees of freedom required to represent uncertain model inputs and/or parameters
  - Number of parameters
  - Karhunen-Loève expansion (KLE) for random fields
- Hi-D challenge in UQ: high-dimensional integration
- We discuss advances in
  - Local KLE
    - Reduced KLE dimensionality for random field in subdomains
  - Basis adaptation
    - Isometric transformations to low-dimensional subspaces
  - Low rank sparse tensors
    - Combinations of low-D integrals

Dimensionality Reduction via Local KLE

We wish to solve PDEs such as

$$\begin{cases} -\nabla \cdot (a(x,\omega)\nabla u(x,\omega)) = f(x), & x \in D, \\ u = g(x) & x \in \partial D. \end{cases}$$

Parameterize the random field $a(x,\omega)$ using KLE

$$a(x,\omega) \approx a(x,Z) = \mu_a(x) + \sum_{i=1}^{d} \sqrt{\lambda_i}\psi_i(x)Z_i(\omega).$$

Divide D into a set of non-overlapping subdomains $D^{(i)}$, $i = 1,\ldots,K$

The original problem on the full domain can be solved in each subdomain with proper coupling conditions at the interfaces.

The collection of the subdomain solutions is equivalent to that of the original problem in the global domain, i.e.,

$$u(x,\omega) = \sum_{i=1}^{K} u^{(i)}(x,\omega)\mathbb{I}_{D^{(i)}}(x),$$

Local KLE – Eigenstructure

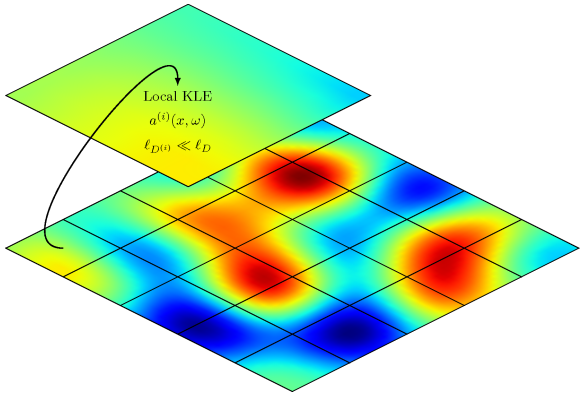
We represent the restriction of the input process $a(x,\omega)$ in the subdomain $D^{(i)}$ as

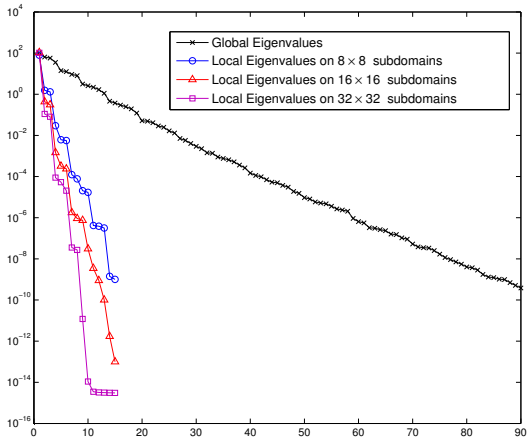
$$a^{(i)}(x,\omega) \approx \mu_a^{(i)}(x) + \sum_{j=1}^{d^{(i)}} \sqrt{\lambda_j^{(i)}}\psi_j^{(i)}(x)Z_j^{(i)}(\omega)$$

The decay rate of the eigenvalues depends critically on the *relative correlation length*.

The rel. correl. length on each subdomain is larger than that on the full domain.

- Local KL eigenvalues decay faster
- $a^{(i)}(x,\omega)$ parameterized w/ a smaller number of random variables, thus $d^{(i)} \ll d$





Decay of eigenvalue magnitudes

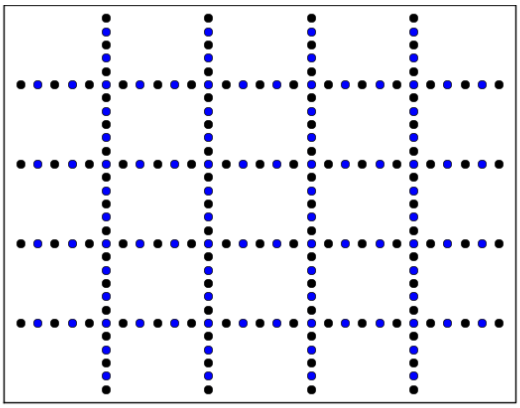
Local KLE – Algorithm

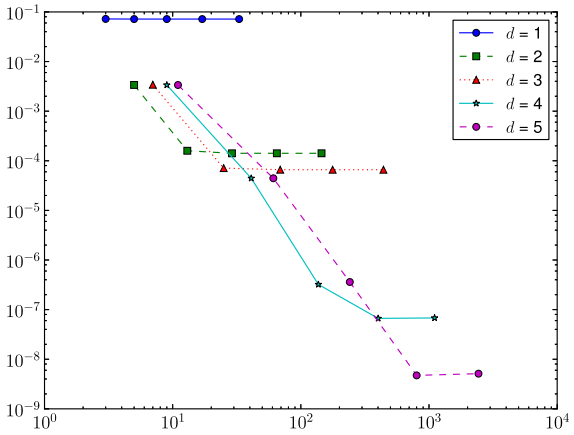
Off-line phase

- In parallel Build (independently) PCE surrogate on each subdomain $D^{(i)}$
- For linear PDE each random realization requires $n_{\mathrm{dof}:\partial D^{(i)}} + 1$

On-line phase

- Generate a realization of the radom field on the global domain
- Project global field onto each subdomain to obtain parameters $Z^{(i)}$ of local KLE $a^{(i)}(x, Z^{(i)})$
- Evaluate local PCE at local random parameters $Z^{(i)}$
- Generate solution on each subdomain by solving global interface problem





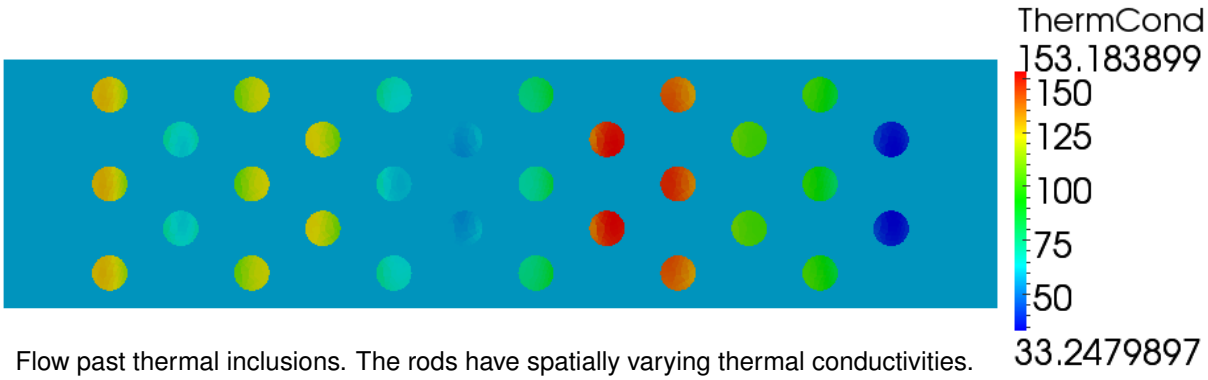
Error vs. # sub-domain solves

Basis Adaptation to Quantities of Interest – USC

- QoIs in hi-D systems are frequently low-dimensional
- We developed a procedure for basis discovery/adaptation
  - permits efficient and accurate approximation within a low-dimensional subspace in which the QoI is concentrated
- Using Gaussian parameterization of the uncertain inputs,
  - Isometry is first applied to induce a desired structure in the representation of the QoI
    - 1^{st}-order terms in one dimension
    - diag. quadratic form w/ 2^{nd}-order terms; match target CDF
  - Reduction is then achieved through projection of the resulting representation
- Reduced model captures:
  - the probabilistic content of the QoI
  - its functional dependence on the original parameterization

Basis Adaptation Demo – Convective Heat Transfer

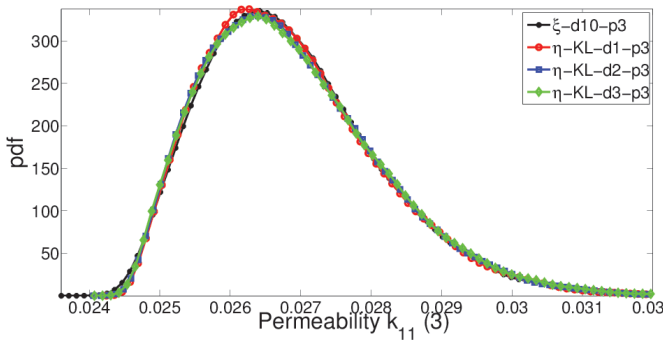
- Consider a 2D domain with flow past random heated inclusions, described in high stochastic dimension



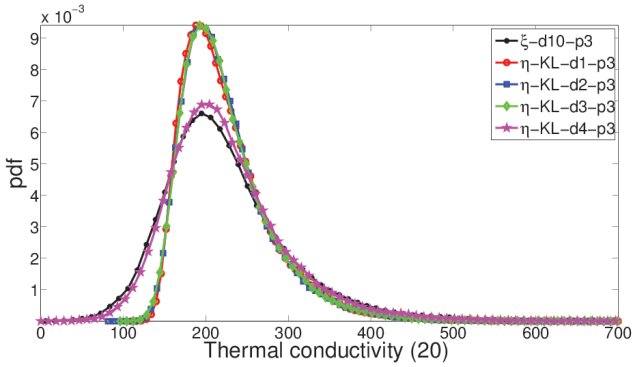
Flow past thermal inclusions. The rods have spatially varying thermal conductivities.

- An upscaled effective stochastic porous medium is computed. The QoI at every spatial point is the homogenized permeability and conductivity.
  - permeability and conductivity are statistically dependent.
  - can be evaluated as functions of the fine scale randomness.

Basis Adaptation to Quantities of Interest – Demo



Upscaled stochastic permeability verified at one spatial point.



Upscaled stochastic conductivity verified at the same spatial point.

- Basis adaptation makes it feasible to evaluate upscaled properties at each spatial location as function of fine scale uncertainty.
- Solve a number of low-stochastic-dimension UQ problems instead of one high-D problem

Fast Evaluation of MP Integrals (FEMPI) — Quantum Chemistry

SNL, UIUC ASCR-BES Partnership

Computational Challenge:

- Accurate computational prediction of key molecular properties requires *ab initio* all-electron theories.
- Initial focus on vibrational and electronic structure integrals
- Integrands involve series of tensor contractions and dense matrix manipulations — Nonscalable!
- Better scaling achieved via enhancements of Monte-Carlo.

QUEST:

Improve integration efficiency and scalability

- Advanced hi-D function representations used in UQ
  - Low rank sparse tensor representations
- Replace hi-D integral with a number of low-D integrals
- Evaluate using sparse quadrature

Vibrational Energy integral – Water Molecule – XVH2

Main approach: low-rank approximation of integrand:

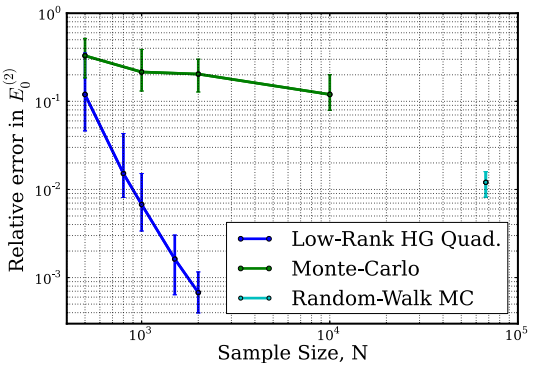
$$f(\boldsymbol{x}) \approx \sum_{i=1}^{R} \prod_{k=1}^{m} f_i^k(\tilde{\boldsymbol{x}}_k), \quad \tilde{\boldsymbol{x}}_k \in \boldsymbol{x} = (x_1, \ldots, x_d), \quad \cup_{k=1}^{m} \tilde{\boldsymbol{x}}_k = \boldsymbol{x}$$

High-dimensional integral is estimated via several low-d integrals

$$\int_{\Omega_{\boldsymbol{x}}} f(\boldsymbol{x}) d\boldsymbol{x} \overset{\text{low-rank}}{\approx} \sum_{i=1}^{R} \prod_{k=1}^{m} \int_{\Omega_{\tilde{\boldsymbol{x}}_k}} f_i^k(\tilde{\boldsymbol{x}}_k) d\tilde{\boldsymbol{x}}_k \overset{\text{quad.}}{\approx} \sum_{i=1}^{R} \prod_{k=1}^{m} \sum_{q=1}^{Q} w_q f_i^k(\tilde{\boldsymbol{x}}_k^q)$$

E.g., second-order correction to zero-point energy (6D):

$$E_0^{(2)} = \int e^{-||\boldsymbol{\omega}^T \boldsymbol{x}||^2} \Delta V(\boldsymbol{x}) H(\boldsymbol{x}, \boldsymbol{x}') e^{-||\boldsymbol{\omega}^T \boldsymbol{x}'||^2} \Delta V(\boldsymbol{x}') d\boldsymbol{x} d\boldsymbol{x}'$$



Ongoing work:

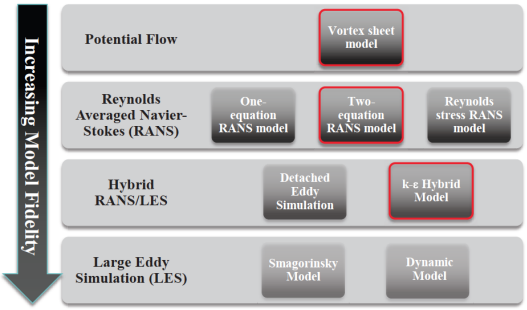
- Singular integrals from MP2 theory. Exponential sum apprx + low-rank.
- Automatic detection of groupings $\tilde{\boldsymbol{x}}_k$ and sparsity within them.
- Scale up to larger systems.

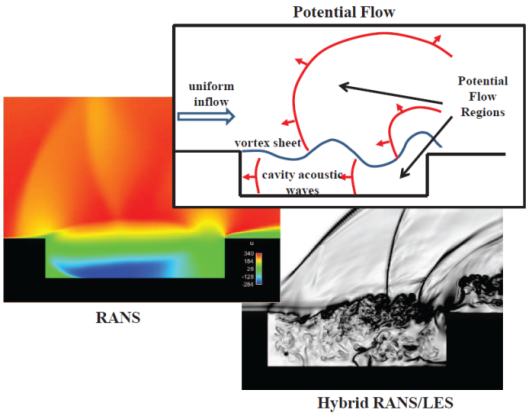
Model Complexity

- Large-scale models require substantial computational resources for solving the original deterministic problem
- This can lead to infeasible costs for either intrusive or non-intrusive/sampling-based UQ methods
- We discuss advances in
  - Multifidelity UQ methods
    - Use of predictions from models at different levels of fidelity
  - Hierarchical calibration and model discrepancy
    - Use of model bias and discrepancy in statistical calibration
  - Adaptive sparse quadrature
    - Selection of computational samples for hi-D integration

Multiple Model Forms in UQ – SNL

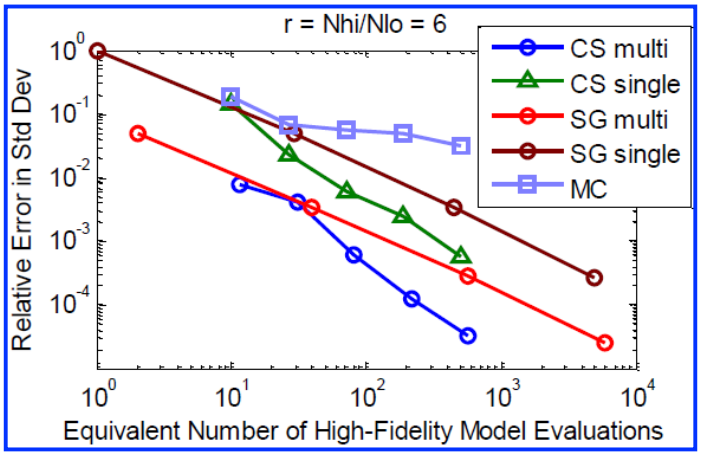
- Given a clear hierarchy of fidelity:
  - Multifidelity forward UQ:
    - UQ for hi-fi model leveraging cheaper low-fi models
  - Multifidelity inference:
    - Estimation of low-fi model discrepancies
- Given a non-hierarchical ensemble of credible models:
  - Model probability – prior info
  - Bayesian model selection
  - Model averaging
- Both hierarchy and peers
  - Leverage model selection and multifidelity inference





Multifidelity UQ Using Stochastic Expansions

- High-fidelity simulations can be prohibitive for use in UQ
- Low fidelity "design" codes often exist that are predictive of basic trends



- Leverage LF codes w/ HF UQ
  - Global approximations of model discrepancy
  - Adaptive sparse grids:
    - Gen. sparse grids for LF & discrepancy levels
    - Greedy selection from grids: max ΔQoI/ΔCost
    - Refine discrepancy where LF is less predictive
  - Compressed sensing:
    - Target sparsity within the model discrepancy

Hierarchical Calibration & Model Discrepancy – LANL

Hierarchical calibration addresses the relationship between model discrepancy and parameter bias

- Given different calibration examples with different bias

$$y_i = \eta(x, \theta + b_i) + \delta_i(x, \theta + b_i) + \epsilon_i$$

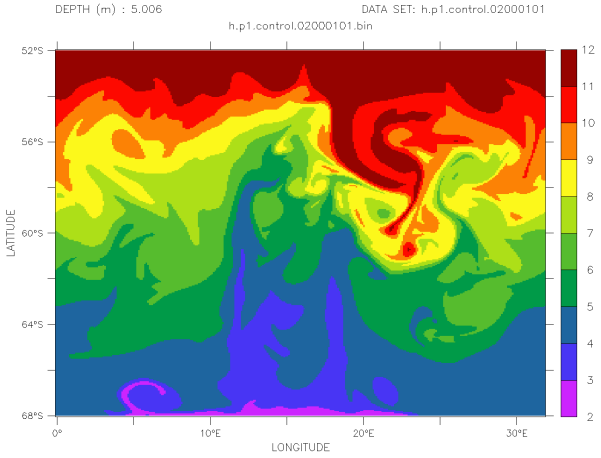
- e.g. climate model bias different at high vs. low latitudes

Employ a hierarchical model, reconciling the evidence of bias

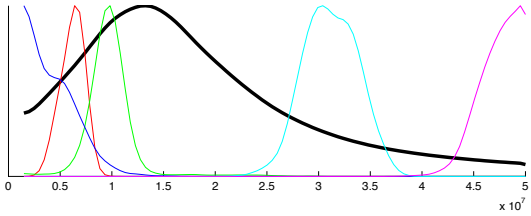
- Inferred discrepancy effects are better fit to problems
- Diagn. of relationship bet. parameter bias & discrepancy
- Additional source of uncertainty identifiable in UQ analyses
- Capability has been in GPMSA, developing the clarifying examples and framework of diagnostics for user adoption

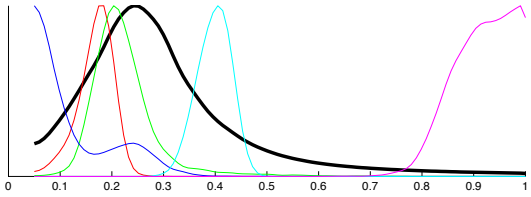
Hierarchical Calibration Demo – Southern Ocean

- Idealized southern ocean model with two parameters
- Calibration w.r.t. higher fidelity Parallel Ocean Program (POP) computations
  - LANL+NCAR
- A number of metrics
  - Temperature, salinity, density vs. depth
  - Vertical heat & salt transport
- Hierarchical distribution combines information from different metrics



Potential temperature (deg C)





Adaptive Sparse Quadrature (ASQ) – Duke/MIT

Non-Intrusive Pseudospectral projection

- Sparse tensorization of 1-D quadrature formulae
- Reduce number of simulations, improve accuracy

Adaptivity:

- Progressive construction with cost control
- Robust error indicator to guide adaptation
- Nested hierarchical approximation
- Sensitivity-based directional refinement

Application to forward UQ in Gulf of Mexico modeling

- Challenges with failed computational samples
  - both ASQ & MC-LHS
- Use L2-misfit constrained L1-norm minimization (BPDN)
  - Estimate QoIs: sparse learning from available samples

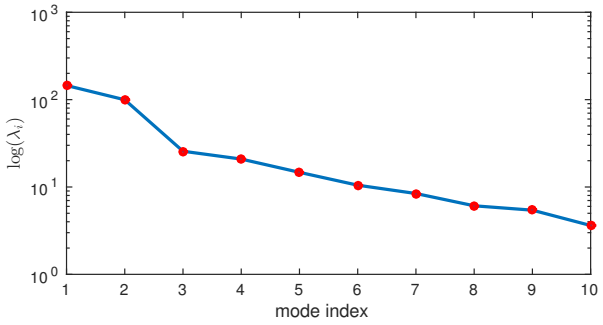
UQ for Circulation in Gulf of Mexico

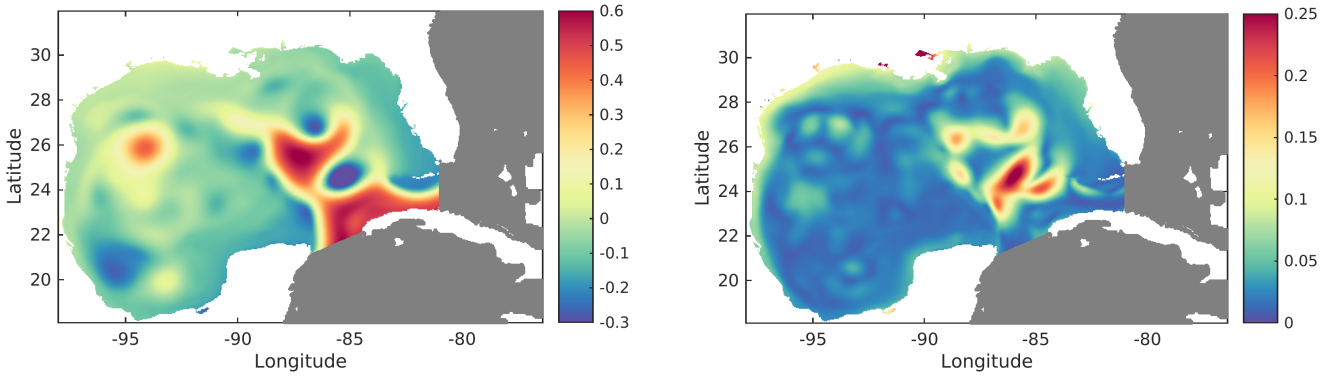
Impact of uncertainty in:

- Initial conditions (4 dims)
- Wind stress (4-dims)
  - Time-dependent EOFs

on circulation in Gulf of Mexico

Eigenvalue decay – SSH





Mean (left) and STD (right) of sea surface height (SSH) at day 30

Statistical Inverse Problems

- Statistical inversion is used for data-based estimation of model parameters/inputs with quantified uncertainty
- Inverse problems are hard!
  - Typically ill-posed; ill-conditioned
  - High-dimensionality
  - Forward model complexity
- We discuss select recent advances in Bayesian inversion
  - Optimal experimental design (OED)
    - Identify optimal sensor placement for geophysical inversion
  - Surrogates & Markov chain Monte Carlo methods
    - Adaptive local surrogates
    - Parallel MCMC methods

Scalable algorithms for optimal exptl design (OED)

Large-scale Bayesian inverse problems

UT

Context: Inference of parameter fields w/ quantified uncertainty

- OED asks the "outer loop" question:
  - How to choose sensor locations so that the inferred parameter field uncertainty is minimized?
- In its full generality, this is intractable:
  - Inner problem alone is an infinite dimensional Bayesian inverse problem
- Approach:
  - Represent covariance by inverse Hessian of negative log posterior (Laplace approximation)
  - Invoke fast randomized trace estimators
  - Employ techniques from PDE-constrained optimization

Result: *OED method whose cost–measured in forward PDE solves–scales independent of parameter/sensor dimension*

Formulation of OED for Bayesian inversion

Hessian/PDE-constrained optimization problem

- Seek an experimental design $\boldsymbol{w}$ (e.g., sensor locations) to collect data $\boldsymbol{d}$ to minimize average posterior variance
- OED problem:
  - Minimize average variance given by trace of inverse Hessian, evaluated at maximum a posteriori solution of inverse problem $\boldsymbol{m}^*$:

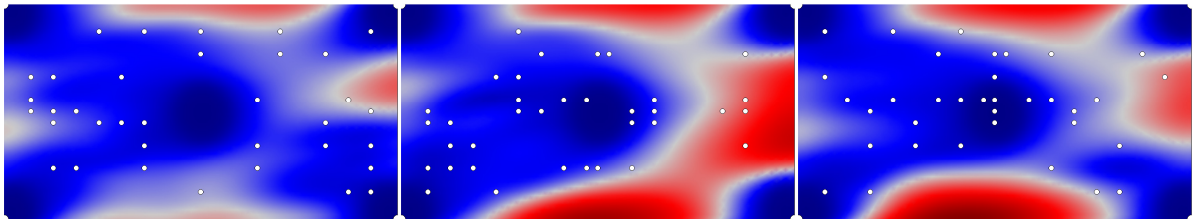
$$\min_{\boldsymbol{w}} \ \mathbb{E}_{\boldsymbol{d}}\Big\{ \text{trace}\left[\mathcal{H}^{-1}\big(\boldsymbol{m}^*(\boldsymbol{w}), \boldsymbol{w}; \boldsymbol{d}\big)\right] \Big\}$$

- Sample averaging to approximate expectation over $\boldsymbol{d}$
- Randomized trace estimation of $\mathcal{H}^{-1}$

A-optimal sensor placement for inferring log-permeability in subsurface flow (SPE model)

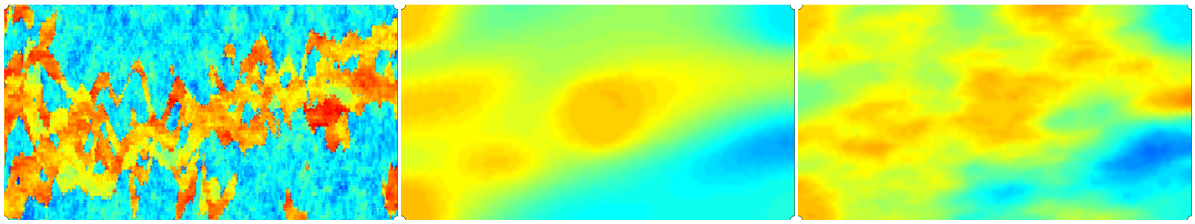
Posterior variance with various sensor placements

Optimal | Sub-optimal | Sub-optimal



Inference with the optimal design (parameter dim $\sim 10^4$)

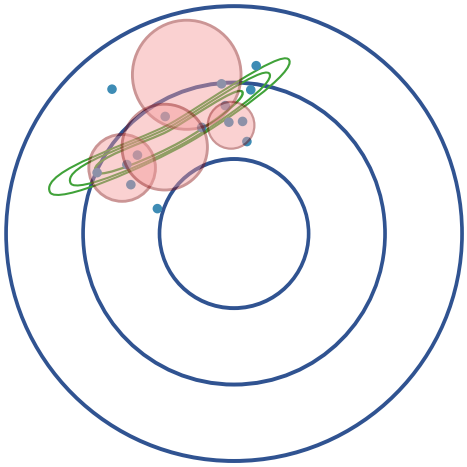
True parameter | Posterior mean | Posterior sample



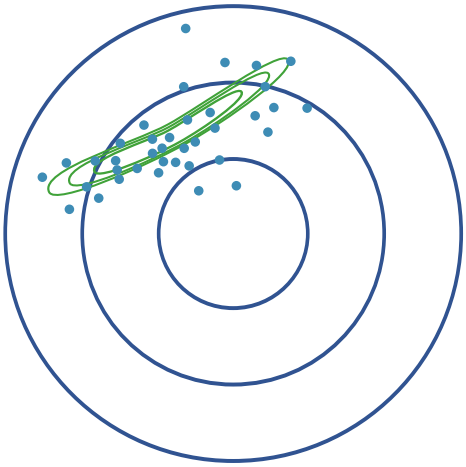
Asymptotically Exact MCMC

MIT

Earlier times



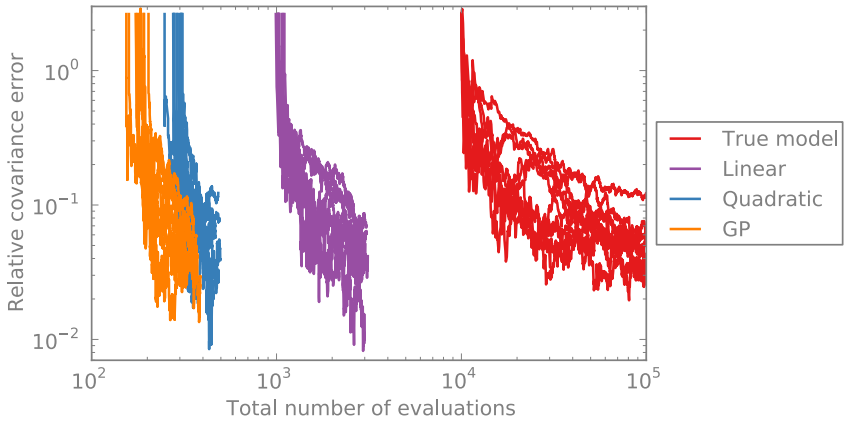
Later times



- **Inference** in computationally intensive models is essentially intractable without **surrogates**
- Key questions: *Where should a surrogate be accurate? How to construct it? Should it depend on the data? How does error in the surrogate corrupt inference?*
- Our approach: **incremental** and **asymptotically exact** construction of **posterior-focused** model approximations

Asymptotically Exact MCMC – Surrogates MIT

- Framework includes several different local approximation schemes: linear, quadratic, Gaussian process
- Accuracy versus cost *(below)*; orders of magnitude speedups



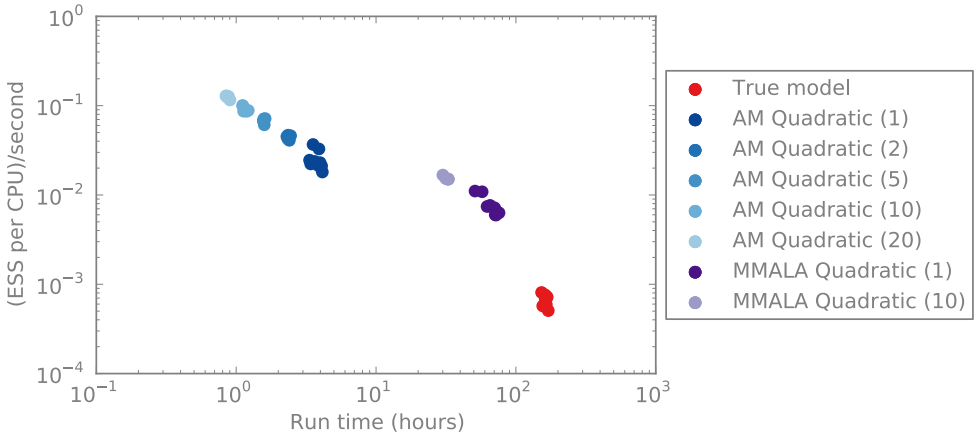
Recent developments:

- Surrogates coupled with more sophisticated (gradient and Hessian-exploiting) MCMC proposals
- **Parallel MCMC chains**, sharing a common pool of model evaluations

Parallel MCMC with Surrogates

MIT

- Build a common pool of model runs across parallel workers
- Approximation guaranteed to target the correct distribution; use *effective sample size (ESS)* to measure efficiency
- **ESS per CPU-sec** usually constant with simple parallel MCMC
- Instead, it *increases* dramatically: chains "borrow strength"
- Cost of a computationally intensive contaminant transport inverse problem reduced **from 200 hours to 30 minutes**



Architecture Awareness

- UQ on massively-parallel heterogeneous architectures
  - Scalability – Load balancing, communication, synchrony
  - MPI, OpenMP, GPU/...
  - Memory utilization
  - Parallel I/O – data
  - Fault tolerance and resilience
- Consequences:
  - UQ problem formulation, algorithms, software
- We discuss our practice & vision in architecture-aware UQ
  - UQ libraries
  - Non-intrusive/sampling-based UQ
  - Intrusive stochastic-Galerkin UQ

Architecture Awareness – Current Practice – SNL

- We lower the bar for UQ on advanced architectures
  - Large compute ensembles on leadership-class machines
    - Multilevel optimized partitioning & scheduling
  - Relax reqmt to converge all simulations for all partitions
    - Fault tolerance, failure mitigation & restart
- We ease UQ adoption via usability features/enhancements
  - Library embedding of UQ services in applications
    - Embed in familiar apps; eliminate custom interface code
    - Simplify parallel execution; *e.g.* FELIX/Dakota (PISCEES)
  - Rapid prototyping and integration with scripting languages
- We invest in emerging capabilities, directly or leveraged
  - Advanced fault tolerance (ASCR UQ)
  - Advanced UQ workflows (QUEST/SUPER, MUQ DAGs)
  - Centralized accessibility (e.g., Github) to maximize community adoption and involvement

Forward Vision – Non-intrusive UQ

SNL

- Leverage emerging runtime systems for task-based parallelism management within QUEST tools
  - e.g. Legion, Charm++, HPX, Uintah
  - Migrate from imperative hybrid-MPI scheduling to declarative parallelism models
- Aggregate UQ and simulation workflow tasks within the same runtime system, exposing new opportunities for streamlining, asynchrony, etc.
  - Move toward loop reordering / embedded ensembles
  - Mappers control task delegation to hybrid hardware
    - CPU, GPU, MIC

Forward Vision – Intrusive UQ

SNL

- Identify application partners for intrusive UQ methods
  - Additional dedicated investments for selected applications
  - Stochastic Galerkin methods
  - Hybrid Galerkin-Collocation methods
- Different levels and types of intrusion, in terms of
  - Software (library linking)
  - Coupling strategy (multiphysics/multiscale UQ)
  - Parallel task scheduling (aggregation of runtime workflows)
  - The actual simulation/solver
- Available intrusive and linear algebra libraries in Trilinos
  - Stokhos: stochastic Galerkin systems
  - Tpetra: serial and distributed parallel linear algebra
  - Kokkos: manycore performance portability

Sparse Linear Stochastic Galerkin Solvers – SNL

Explore algorithmic constructions that show potential to keep future hierarchy of HPC cores busy

- Additive-multigrid/multilevel (physical/stochastic)
  - Phys. mesh coarse/fine on communication/compute cores
  - Decoupled stochastic prolongation/restriction operators
  - Higher-order stoch. levels $\Longrightarrow$ compute intensive cores
- Recursive hierarchical matrix preconditioned inversion
  - Break up matrix hierarchically into smaller nested blocks
  - Each of which can be solved more easily and independently
- Hybrid stochastic Galerkin/collocation approaches
  - Coupled intrusive/non-intrusive strategy
  - Target optimal use of computational architecture
  - Tradeoff solution samples of deterministic problem for reduced-size/better-conditioned stochastic Galerkin system

Closure

- Presented select highlights of recent progress
  - High dimensionality
  - Model complexity
  - Statistical inversion
  - Architecture
- We continue to
  - Refine and robustify QUEST algorithms and software to address UQ challenges in large-scale problems
  - Address UQ needs of SciDAC application partnerships

quest-scidac.org