Final report for award ER26160/DE-SC0010518

Extreme-Scale Bayesian Inference for Uncertainty Quantification of Complex Simulations

January 11, 2018

Lead Institution:
The University of Texas at Austin

Lead Principal Investigator:
Prof. George Biros
The University of Texas at Austin
Institute for Computational Engineering and Sciences
201 E. 24th St., Stop C0200
Austin, Texas 78712-0027
tel: 512-232-9566
email: `biros@ices.utexas.edu`

Administrative point of contact name, phone, email:
Courtney Frazier Swaney
tel: 512-471-6424; email: `c_frazier@austin.utexas.edu`

Funding Opportunity Announcement Number: DE-FOA-0000895
DOE/Office of Science Program Office: Office of Advanced Scientific Computing Research
DOE/Office of Science Program Office Technical Contact: Dr. Steven L. Lee
Research area or areas: Parallel algorithms for uncertainty quantification, Bayesian inference, sampling, experimental design, kernel density estimation, and Bayesian coarse graining

List of all Collaborating Institutions and their PIs/Co-PIs:

The University of Texas at Austin:	George Biros (Lead PI), Tan Bui-Thanh, Omar Ghattas, Robert Moser, Tinsley Oden, Todd Oliver, Georg Stadler
Massachusetts Institute of Technology:	Youssef Marzouk
Oak Ridge National Laboratory:	Jeffrey Vetter

REPORTS

year	1	report: see Section 2
year	2	report: see Section 3
year	3/4	report: see Section 4

There was a one-year, no-cost extension for this project.

1. Overview and objectives

Uncertainty quantification (UQ)—that is, quantifying uncertainties in complex mathematical models and their large-scale computational implementations—is widely viewed as one of the outstanding challenges facing the field of CS&E over the coming decade.

The EUREKA project set to address the most difficult class of UQ problems: those for which both the underlying PDE model as well as the uncertain parameters are of extreme scale. In the project we worked on these extreme-scale challenges in the following four areas:

1. Scalable parallel algorithms for sampling and characterizing the posterior distribution that exploit the structure of the underlying PDEs and parameter-to-observable map. These include structure-exploiting versions of the randomized maximum likelihood method, which aims to overcome the intractability of employing conventional MCMC methods for solving extreme-scale Bayesian inversion problems by appealing to and adapting ideas from large-scale PDE-constrained optimization, which have been very successful at exploring high-dimensional spaces.

2. Scalable parallel algorithms for construction of prior and likelihood functions based on learning methods and non-parametric density estimation. Constructing problem-specific priors remains a critical challenge in Bayesian inference, and more so in high dimensions. Another challenge is construction of likelihood functions that capture unmodeled couplings between observations and parameters. We will create parallel algorithms for non-parametric density estimation using high dimensional N-body methods and combine them with supervised learning techniques for the construction of priors and likelihood functions.

3. Bayesian inadequacy models, which augment physics models with stochastic models that represent their imperfections. The success of the Bayesian inference framework depends on the ability to represent the uncertainty due to imperfections of the mathematical model of the phenomena of interest. This is a central challenge in UQ, especially for large-scale models. We propose to develop the mathematical tools to address these challenges in the context of extreme-scale problems.

4. Parallel scalable algorithms for Bayesian optimal experimental design (OED). Bayesian inversion yields quantified uncertainties in the model parameters, which can be propagated forward through the model to yield uncertainty in outputs of interest. This opens the way for designing new experiments to reduce the uncertainties in the model parameters and model predictions. Such experimental design problems have been intractable for large-scale problems using conventional methods; we will create OED algorithms that exploit the structure of the PDE model and the parameter-to-output map to overcome these challenges.

Parallel algorithms for these four problems were created, analyzed, prototyped, implemented, tuned, and scaled up for leading-edge supercomputers, including UT-Austin's own 10 petaflops *Stampede* system, ANL's *Mira* system, and ORNL's *Titan* system. While our focus is on fundamental mathematical/computational methods and algorithms, we will assess our methods on model problems derived from several DOE mission applications, including multiscale mechanics and ice sheet dynamics.

Several software packages were produced and updated in the context of this project:

- https://hippylib.github.io/, `hIPPYlib - Inverse Problem PYthon library`: hIPPYlib implements state-of-the-art scalable adjoint-based algorithms for PDE-based deterministic and Bayesian inverse problems. It builds on FEniCS for the discretization of the PDE and on PETSc for scalable and efficient linear algebra operations and solvers

- http://muq.mit.edu/, `MUQ`: In a nutshell, MUQ is a collection of tools for constructing models and a collection of uncertainty quantification (UQ)focused algorithms for working on those models. Our goal is to provide an easy and clean way to set up and efficiently solve UQ problems.

- http://libqueso.com/, QUESO: A collection of algorithms and other functionalities aimed for the solution of statistical inverse problems, for the solution of statistical forward problems, for the validation of a model and for the prediction of quantities of interest from such model along with the quantification of their uncertainties.
- http://padas.ices.utexas.edu/libaskit/, LIBASKIT: This is the LIBASKIT set of scalable machine learning and data analysis tools. Currently we provide codes for kernel sums, nearest-neighbors, kmeans clustering, kernel regression, and multiclass kernel logistic regression. All codes use OpenMP and MPI for shared memory and distributed memory parallelism.

Several publications (see Section 5) and presentations (see Section 6) are also part of the outcomes.

2. Progress report, Aug 31, 2013–May 30, 2014

Our progress so far is categorized by our main research threads as outlined in §1

2.1. HPC methods for uncertainty quantification:

Papers have been submitted for publication but most of the threads below are ongoing work; participating researchers:Biros, Ghattas, Stadler, Vetter

The majority of work on scalable parallel algorithms in CS&E has focused on the forward problem: given parameter inputs, solve the governing equations to determine the outputs. We have been working on the broader question: given a model containing uncertain parameters, (possibly noisy) observational data, and a prediction quantity of interest, how do we construct scalable parallel algorithms to (1) infer the model parameters from the data via the model, (2) quantify the uncertainty in the inferred parameters, and (3) propagate the resulting uncertain parameters through the model to issue predictions with quantified uncertainties? And do all of this at extreme scales? In the past year we have developed scalable parallel algorithms for this data-to-prediction process based on the notion of linearizing the appropriate maps (parameter-to-observable, parameter-to-prediction) and exploiting low-dimensional structure of these maps via randomized methods. We have applied these parallel algorithms and assessed their performance on one of the our testbed problems: modeling the dynamics of the Antarctic ice sheet and its ultimate effect on sea level. We have shown that the work required to execute this data-to-prediction process, measured in number of forward (and adjoint) model solves, is independent of the state dimension, parameter dimension, data dimension, and number of processor cores.

In a second direction, we survey UQ methods and derive basic performance models that can be used as guidelines for system design. We focus on UQ methods for systems governed by partial differential equations, and on algorithms that can scale to high-dimensional parameter spaces. We review the complexity of the representative methods like gradient and Hessian-based schemes for sampling, data-driven priors, and stochastic collocation methods. We use two state-of-the-art UQ algorithms to demonstrate the ideas outlined in the performance models and we use Aspen to explore the design space for future systems with respect to these UQ algorithms. Specifically, we have created a preliminary model for the adjoint UQ method using the Aspen performance modeling language and tools. This model describes the iterated forward and backward steps used to calculated gradients for simulation parameters, and it can be applied to various application forward solutions to explore their predicted runtime and memory usage when extended to perform UQ. As memory constraints can be a bottleneck in UQ, the specific implementation of the adjoint method may vary depending on the quantity of memory and performance of the I/O subsystem. To explore these effects, we begun extending Aspen to support I/O components within a machine model and I/O resource usage within application models. We combine our analysis with experiments on the Stampede system at the Texas Advanced Computing Center and facilities at ORNL and demonstrate the feasibility of our approach. We find that even if the forward problem has low communication intensity, UQ methods require non-local computations and thus more intense communication (collective vs point-to-point) and more local storage.

2.2. Theoretical foundations:

Papers have been submitted for publication; participating researchers:Bui-Thanh, Ghattas, Stadler, Moser, Oliver, Biros, Marzouk

We have analyzed novel sampling schemes for highly non-Gaussian posteriors. based on the randomized maximum likelihood method (RML). First we construct an RML method in infinite dimensional setting and show that the RML correctly samples the infinite dimensional posterior. Second, we have speed up with RML computation using a trust region inexact Newton CG. Third, we develop a sensitivity analysis technique to compute good initial guess for RML optimization problem, and hence further speeding the optimization work. We have also derived an exact joint distribution of the RML samples and the randomized data. We then use it as a proposal for nonlinear Bayesian inverse problems. Our numerical results show that the acceptance rate is 50%.

Finally, we have explored dimensionality reduction methods for Bayesian inference. The intrinsic dimensionality of an inverse problem is affected by prior information, the accuracy and number of observations, and the smoothing properties of the forward operator. From a Bayesian perspective, changes from the prior to the posterior may, in many problems, be confined to a relatively low-dimensional subspace of the parameter space. We have been developing dimension reduction techniques to identify these so-called "likelihood-informed" directions in nonlinear inverse problems. These techniques require the collection of Hessian information over the support of the posterior distribution, or a distribution that dominates the posterior distribution. In this context, we have developed methods that use the randomize-then-optimize (RTO) (or randomized maximum likelihood) algorithm to explore the variation of the Hessian over the parameter space. RTO is well suited to this problem because exploration of the Hessian does not require exact posterior sampling; as a result, our approach does not need to "correct" the RTO samples with importance weights or a Metropolis step, and is thus embarrassingly parallel and computationally efficient.

One of the challenges we face in the use of Bayesian inference for uncertainty quantification is the application of algorithms to compute or sample from the posterior distribution when the forward model is itself stochastic or uncertain. This occurs often when probabilistic representations for model inadequacy are employed. The difficulty is that the integral defining the likelihood becomes non-trivial in this case. To avoid this, we are exploring a reformulation of the problem that would increase the dimension of the inference problem by introducing auxiliary random variables and make the likelihood integral again trivial. The auxiliary variables can then be marginalized out in the posterior. This approach is being explored and tested in a model problem derived from model inadequacy representations in chemical kinetics.

In many science applications, the solution of a statistical inverse problem is a prelude to making a prediction—i.e., characterizing the posterior distribution of a quantity of interest (QoI), rather than the model parameters, conditioned on available data. The prediction may be described by a few scalar quantities, or perhaps even be as simple as the binary outcome of an event (e.g., collapse versus non-collapse of the ice sheet). Even if the QoI is quite simple to describe, however, inference methods that actually exploit this simplicity—in the nonlinear inverse problem setting—are sorely lacking. Avoiding a full characterization of the posterior remains an open challenge. To address this issue, we have been developing new Bayesian methods for event prediction, based on on pseudo-marginal Markov chain Monte Carlo. The ideas is to *approximately* marginalize away directions irrelevant to the prediction, while sampling from the exact posterior distribution on prediction-relevant directions. Identifying these directions again requires the exploration of principal Hessian directions, while to calculate approximate marginals we are exploring the use of Laplace approximations and of local surrogate models. A further challenge is *optimal experimental design* for event prediction—identifying observations that are most informative for the posterior probability of the event. This question focuses the optimal experimental design problem on few directions, and we have formulated it in a fully Bayesian manner. Making the computations tractable in large-scale problems is the subject of ongoing work.

Finally, we are designing a new approach to efficiently approximate kernel sums in high dimensions using randomized linear algebraic methods. We explore the following open question: can randomized algorithms serve as an efficient means of computing compact representations of subblocks of a kernel matrix? If the answer is yes, then these algorithms may open the door to the construction of fast, accurate kernel summation algorithms which do not scale exponentially with the ambient dimension of the input. We are working on a classification of randomized algorithms for approximating the evaluation of matrix-vector products for suitably structured matrices, such as those arising in kernel summations. j

2.3. Coarse-graining using Bayesian formulations

Papers have been submitted for publication; participating researchers: Biros, Oden

We have considered parametric models of coarse-grained representations of several types of atomistic systems and particulate flow systems. For atomistic systems we consider simulations with hexane and poly (PMMA) providing the test bed for numerical experiments. The molecular dynamics of the systems under study is modeling using the MD code LAMMPS and MCMC sampling of various distributions is implemented using QUESO. A Bayesian approach to model calibration and validation has been developed and applied to representative examples. The current work centers on the development of statistical calibration of models of molecular units that are the building blocks for representative polymer chains (RPCs) in a representative volume of the atomic system. Data is furnished by the All-Atom (first-principles) model defined using LAMMPS and OPLS. Priors are determined using maximum entropy principles, and posteriors of the calibration experiments are used as priors for a higher-level of statistical calibration embodied in a sequence of validation experiments with observables derived from appropriate projections of global quantities of interest. The challenge at this point is to select the model itself from a class of parametric models that correspond to possible interaction potentials for the system. We have introduced two significant tools : 1) the notion of posterior model plausibilities and 2) the idea of model sensitivity. We are currently investigating these approaches and we are extending them to complex fluid simulations.

2.4. Unexpended funds

So far we have spent a partial amount of our allocation. For the UT allocation, our total expenses are $130K, 26% of the allocated funds. This is due to ongoing hiring. Several students and postdoctoral students have recently jointed our group and additional ones will be joining our group this Fall. We anticipate that we will be able to spend those remaining Year 1 funds during budget Year 2.

3. Progress report, June 1, 2014–May 30, 2015

Our progress so far is categorized by our main research threads as outlined in section 1.

3.1. HPC methods for uncertainty quantification

Participating researchers:Biros, Ghattas, Stadler, Vetter

Scalable kernel methods for uncertainty quantification: We have designed a fast algorithm, ASKIT, for kernel summation problems in high dimensions. This has been a significant breakthrough since kernel sums are important and there are no algorithms or software that scales. Kernel sums, also known as N-body problems, appear in computational physics, numerical approximation, nonparametric statistics, and machine learning. In our context, the sums depend on a kernel function that is a pair potential defined on a dataset of points in a high-dimensional Euclidean space. A direct evaluation of the sum scales quadratically with the number of points. Fast kernel summation methods can reduce this cost to linear complexity, but the constants involved do not scale well with the dimensionality of the dataset. The main algorithmic components of fast kernel summation algorithms are the separation of the kernel sum between near and far field (which is the basis for pruning) and the efficient and accurate approximation of the far field. We introduce novel methods for pruning and for approximating the far field. Our far field approximation requires only kernel evaluations and does not use analytic expansions. Pruning is not done using bounding boxes but rather combinatorially using a sparsified nearest-neighbor graph of the input distribution. The time complexity of our algorithm depends linearly on the ambient dimension. The error in the algorithm depends on the low-rank approximability of the far field, which in turn depends on the kernel function and on the intrinsic dimensionality of the distribution of the points. The error of the far field approximation does not depend on the ambient dimension. We tested the new algorithm along on experimental results that demonstrate its performance. As a highlight, we solved problems with Gaussian kernel sums for 100 million points in 64 dimensions, for one million points in 1000 dimensions, and for problems in which the Gaussian kernel has a variable bandwidth. To the best of our knowledge, all of these experiments are prohibitively expensive with existing fast kernel summation methods.

In addition we parallelized this code. There is extensive work on treecodes and their parallelization for kernel summations in three dimensions, but there is little work on high-dimensional problems. Recently, we introduced a novel treecode, ASKIT, which resolves most of the shortcomings of existing methods. We introduced novel parallel algorithms for ASKIT, derive complexity estimates, and demonstrate scalability on synthetic, scientific, and image datasets. In particular, we introduced a local essential tree construction that extends to arbitrary dimensions in a scalable manner. We introduced data transformations for memory locality and use GPU acceleration. We conducted testson the Maverick and Stampede systems at the Texas Advanced Computing Center. Our largest computations involve two billion points in 64 dimensions on 32,768 x86 cores and 8 million points in 784 dimensions on 16,384 x86 cores.

Optimal experimenal design: This past year we have worked on developing scalable parallel algorithms for optimal experimental design (OED) problems. Bayesian inversion asks the question: how do we infer model parameters and their associated uncertainty, given uncertain observational data and a possibly uncertain model? The OED problem addresses the broader and more basic question: how do we decide what to observe in the first place, in order that the resulting model parameters are recovered with the least uncertainty? This is a notoriously difficult problem, particularly in the extreme-scale case (expensive models and high-dimensional parameter spaces), because the Bayesian inverse problem is *merely* an inner problem, with the outer problem represented as an optimization problem over the experimental design variables.

We have developed approximate optimization algorithms that overcome the challenges associated with extreme-scale OED. We measure uncertainty in the recovered inverse problem parameters by the *A-optimal design* criterion—the trace of the inverse of the Hessian evaluated at the maximum a posteriori point—which is effectively a Laplace approximation of the posterior. The basic ideas can be extended to other criteria (such as D-optimality). Thus we seek to design the observational system to minimize the average variance of the inferred parameters. The trace is approximated (to arbitrary precision) by a randomized trace estimator, which results in second-order adjoint (of the inverse problem) PDE constraints for each random trace vector (generally a small number are required). The (Laplace approximation of the) Bayesian inverse problem also appears as a constraint (in the form of first-order optimality conditions), whose solution defines the MAP point. We form a meta-Lagrangian for this OED problem, and compute gradients in the style of PDE-constrained optimization algorithms, that is, invoking adjoints of the OED problem. Algorithmic scalability tests for a subsurface flow Bayesian OED problem indicate that the cost, measured in number of forward or inverse PDE solves, is *independent of both the parameter and the data dimensions*.

Development of domain-specific language for modeling UQ applications. Aspen is a domain-specific language used to model the organization and behavior of both applications and computing systems, and it includes both a library and a suite of tools to analyze these application and machine models. In the prior reporting period, we created a preliminary template model in Aspen to explore the design space for future systems with respect to UQ algorithms. As Aspen was initially designed to model compute-intensive scientific applications systems, we anticipated that expanded capabilities in Aspen would improve our analysis of UQ methods. For example, adding I/O components to a machine model and application resource usage would be necessary, but we realized that a more expansive way of describing computing systems and their infrastructure would be more effective, and so we have started to abstract I/O behavior into a more general set of machine characteristics; this will culminate in a new style of more flexible machine model capable of dealing with these abstractions and includes not just I/O, but multiple memory systems and networking layers. Another new feature we have added is a framework to study time-dependent behaviors in Aspen; this includes both resource usage which changes continuously over time as well as discrete changes and behaviors which occur at periodic intervals. We are investigating adding even more temporal support, including tracking more behaviors such as memory and file allocation as a run progresses. Combined, we expect these to lead to a richer set of analyses for UQ algorithms.

3.2. Theoretical foundations

Participating researchers:Bui-Thanh, Ghattas, Stadler, Moser, Oliver, Biros, Marzouk

Fast posterior sampling in Bayesian methods: We have been developing new Bayesian computational methods designed to characterize the marginal posterior probability distribution of a low dimensional quantity of interest (QoI), in situations where the QoI depends on high-dimensional inversion parameters and the associated joint distribution is non-Gaussian. These problems are large-scale in two respects: first, the parameter space, when not restricted to the QoI, is high dimensional; second, the forward model is computationally intensive. Our scheme is goal-oriented in the sense that characterizing uncertainty in the high-dimensional parameters is only an intermediate step towards characterizing uncertainty in the QoI. The aim of our algorithm is to explore the posterior distribution of the QoI in structure-exploiting manner—in particular, to approximate the expensive forward model and to sacrifice accuracy in parameters that are not of ultimate interest. For instance, high-dimensional inversion parameters may be inputs or boundary conditions of a PDE, while the QoI may be a scalar functional (deterministic or stochastic) of the PDE solution.

Our new sampling method targets the posterior distribution associated with the QoI directly, marginalizing over the other parameters. Outside of the Gaussian setting, however, marginal densities rarely have an analytic form. We thus compute stochastic estimates of the marginal density of the QoI, and use these estimates to construct local polynomial approximations of the target marginal over regions of high posterior probability. These polynomial approximations replace direct evaluation of the marginal in MCMC and are continually refined using noisy evaluations. Under suitable technical conditions, we show that these approximations converge to the true marginal density and hence that the MCMC scheme asymptotically samples from the true posterior distribution of the QoI. In numerical examples, we show that relatively crude pointwise estimates of the marginal density can yield accurate posteriors. In this sense, the scheme uses regularity of the marginal density and the decomposition of the posterior into dependent low-dimensional and high-dimensional parameters to achieve significant computational savings over standard pseudo-marginal MCMC methods or brute-force high-dimensional sampling. Demonstrations on a marine ice sheet problem, where the QoI represents projected ice sheet volume, are in progress.

There are many ways to explore the posterior distribution to estimate the inverse solution and its uncertainty. Though Markov chain Monte Carlo (MCMC) approaches are perhaps the most general UQ tool, they require millions of samples to converge, especially in high dimensions. As a result, millions of expensive forward simulations are necessaryan intractable proposition. More importantly, they discard costly work during the Metropolization if a sample is rejected. We have developed an ensemble-based approach in which the prior probability distribution are represented by empirical distribution with particles (or samples), which are then transformed, via an optimization problem, to particles representing the posterior probability distribution, and hence the statistics, of the inverse solution.

Our proposed approach induces several advantages over the contemporary counterparts. First, it is a consistent Monte Carlo approach that provably converges to the exact posterior distribution. Consequently, it does not require Metropolization and hence avoiding discarding useful expensive simulations. Second, since each particle requires one forward PDE solve, the method naturally accommodates computing resource constraints by matching the number of particles to that of allowable forward solves, and this task is embarrassingly parallel. Third, unlike existing MCMC approaches that couple the Markov chain and forward PDE solves, the proposed method decouples the process of constructing the posterior particles and PDE simulations. As such, it is well suited for current and future supercomputer infrastructures. Fourth, the beauty of this approach is that it casts the sampling challenge into a large-scale optimization problem that capitalizes on our decade of work on parallel large-scale optimization algorithms.

Efficient sampling using the Hamiltonian Monte Carlo Algorithm We are addressing the problem of developing algorithms for sampling probability density functions (pdfs) that arise as Bayesian solutions of statistical inverse problems. Our focus is specifically on cases where (i) the forward model is typically governed by computationally ex- pensive PDEs, which is a characteristic of large-scale statistical inverse problems, (ii) the uncertain parameters are field quantities, which consequently result in a very high-dimensional parameter space after discretization, (iii) the Bayesian problem could be hierarchical in na- ture, and (iv) the forward model is highly nonlinear, which results in a highly non-Gaussian posterior probability density. Our aim is to develop a sampling algorithm that can address all of the above challenges, while at the same time being computationally cheap and efficient when compared to the various existing sampling techniques. The need to explore highly non-Gaussian posterior distributions made the Hamiltonian Monte Carlo (HMC; sometimes also referred to as Hybrid Monte Carlo) sampling algorithm particularly attractive. The algorithm has been shown to have good mixing properties, as manifested by the fact that the sample autocor- relation drops to zero after just a few samples, through generating non-local (global) moves in state space with high acceptance probabilities. Namely, HMC generates proposals through first introducing an independent auxiliary momentum variable, p, associated with each of the model parameters, q, thus enlarging the state space to the (q,p) phase space.

3.3. Coarse-graining using Bayesian formulations

Papers have been submitted for publication; participating researchers: Biros, Oden

We have examined the development of coarse-grained models of atomistic systems for the purpose of predicting target quantities of interest in the presence of uncertainties. A framework has been developed that, enhanced with the concepts of information theory, sensitivity analysis, and Occams Razor, provides a systematic means of constructing coarse-grained models suitable for use in a prediction scenario. This new, adaptive algorithm, the Occam-Plausibility ALgorithm (OPAL), so named for its adherence to Occams Razor and the use of Bayesian model plausibilities, identifies, among a large set of models, the simplest model that passes the Bayesian validation tests, and may therefore be used to predict chosen quantities of interest. The novel application of a general framework of statistical calibration and validation to molecular systems has been applied to a system of polyethylene. The use of this coarse-grained model and OPAL will be extended to the development of a nano-scale continuum model.

We are inverstigating fast multiscale algorithms for particulate flows We developed adaptive high-order accurate time-stepping numerical scheme for the flow of vesicles suspended in Stokesian fluids, which we use as a testbed for complex nonlinear systems with multiple scales, but without scale separation. Our numerical scheme can be summarized as an approximate implicit spectral deferred correction (SDC) method. Applying a textbook fully-implicit SDC scheme to vesicle flows is prohibitively expensive. For this reason we introduce several approximations. Our scheme is based on a semi-implicit linearized low-order time stepping method. (Our discretization is spectrally accurate in space.) We expect that the accuracy can be arbitrary-order, but our examples suffer from order reduction which limits the observed accuracy to a little more than second-order. We also use invariant properties of vesicle flows, constant area and boundary length in two dimensions, to reduce the computational cost of error estima- tion for adaptive time stepping. We present results in two dimensions for single-vesicle flows, constricted geometry flows, converging flows, and flows in a Couette apparatus. We experimentally demonstrate that the proposed scheme enables automatic selection of the time step size and high-order accuracy.

4. Progress report, Jun1, 2015–August 31, 2017

Our progress so far is categorized by our main research threads as outlined in section 1.

4.1. HPC methods for uncertainty quantification

Participating researchers:Biros, Ghattas, Stadler, Vetter:

Scalable kernel methods for uncertainty quantification:

One of the greatest challenges in computational science and engineering today is how to combine complex data with complex models to create better predictions. This challenge cuts across every application area within CS&E, from geosciences, materials, chemical systems, biological systems, and astrophysics to engineered systems in aerospace, transportation, structures, electronics, biomedicine, and beyond. Many of these systems are characterized by complex nonlinear behavior coupling multiple physical processes over a wide range of length and time scales. Mathematical and computational models of these systems often contain numerous uncertain parameters, making high-reliability predictive modeling a challenge. Rapidly expanding volumes of observational dataalong with tremendous increases in HPC capabilitypresent opportunities to reduce these uncertainties via solution of large-scale inverse problems. In this work we have looked at a small set of these problems. One of our main efforts has been the development of scalable kernel methods, which are based on kernel matrices.

Kernel matrices are computational primitives in many of uncertainty quantification methods, e.g., regression, Gaussian processes, particle filters, manifold approximation, covariance approximation, prior construction and other. In our context, we use kernel matrices in the construction of radial basis approximation schemes for the numerical solution of PDEs that govern the propagation of probability distributions. Kernel matrices are large dense matrices and operations like matrix-vector multiplication, matrix inversion, and spectral decomposition are impossible using existing methodologies. In this task our goal is to develop scalable algorithms that resolves these issues. Our algorithm is based on our ASKIT suite of algorithms that use hierarchical approximations of the kernel matrix. Next, we give details on the kernel matrix approximation problem and our approach.

Let $\mathcal{X}$ be a set of N points $\in \mathbb{R}^d$ and let $\mathcal{K}(x_i, x_j) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a given kernel function. The *kernel matrix* is the $N \times N$ matrix whose entries are given by $K_{ij} = \mathcal{K}(x_i, x_j)$ for $i, j = 1, \ldots, N$, $x_i, x_j, \in \mathcal{X}$. In radial basis function approximation, we wish to solve a linear system with matrix $\lambda I + K$, where $\lambda > 0$ is a regularization parameter that controls generalization accuracy and I is the identity matrix. A key computational challenge in kernel methods is the acceleration of linear algebraic operations with kernel matrices and in particular, matrix-vector multiplication and solution of linear systems. This linear solve can be prohibitively expensive for large N because K is typically dense. A typical radial basis function is the Gaussian kernel,

$$\mathcal{K}(x_i, x_j) = \exp\left(-\frac{1}{2}\frac{\|x_i - x_j\|_2^2}{h^2}\right), \tag{1}$$

where h is the *bandwidth*. For small h, K approaches the identity matrix whereas for large h, K approaches the rank-one constant matrix. The first regime suggests sparse approximations while the second regime suggests global low-rank approximations. For the majority of h values, however, K is neither sparse nor globally low-rank. Direct factorization of $\lambda I + K$ requires $\mathcal{O}(N^3)$ work, whereas a Krylov iterative method costs $\mathcal{O}(N^2)$ work per iteration and may require 1000s of iterations. This **complexity barrier** has made the use of kernel methods for large-scale problems impossible. To address this challenge, we have developed a fast kernel approximation scheme, **ASKIT**.[1] ASKIT is the foundation

[1] *A*pproximate *S*keletonization *K*ernel *I*ndependent *T*reecode, which we have introduced. ASKIT approximates K in $\mathcal{O}(dN \log N)$ time.

on which the rest of the algorithms are built.

We completed the design and implementation of a fast direct solver (**INV-ASKIT**) for kernel matrices. We developed a parallel algorithm for computing the approximate factorization of an N-by-N kernel matrix. Once this factorization has been constructed with $O(N\log^2 N)$ work, we can solve linear systems with this matrix with $O(N\log N)$ work. Roughly speaking, ASKIT is based on the approximation of K as the sum of a block-diagonal matrix and a low-rank matrix followed by recursion for each diagonal block. We refer to this process as the construction of the *hierarchical representation* of K. Once we have this representation, one can factorize K by applying recursively the Sherman-Morrison-Woodbury (SMW) formula. The factorization however, can be quite expensive especially since in practice it has to be done for different values of λ during cross-validation studies. We developed two versions of the solver, the first one having $\mathcal{O}(N\log_2^N)$ complexity and the second solver having $\mathcal{O}(N\log_2 N)$ complexity. Since $N = 10^6$–10^9, the seemingly small $\log_2 N$ factor results in one order of magnitude speedups over our first implementation. In addition we have worked on two additional problems. The first one is spectral decomposition of kernel methods. The second one was improving accuracy and robustness of ASKIT so that it supports kernels over multiple bandwidths.

We completed our work on the $\mathcal{O}(N\log N)$ factorization of kernel matrices, including matrices that do not compress well in all spatial levels. Our algorithm only requires kernel evaluations and does not require that the kernel matrix admits an efficient global low rank approximation. Instead our factorization only assumes low-rank properties for the off-diagonal blocks under an appropriate row and column ordering. We designed a hybrid method that, when the factorization is prohibitively expensive, combines a partial factorization with iterative methods. As a highlight, we are able to approximately factorize a dense 11M × 11M kernel matrix in 2 minutes on 3,072 x86 Haswell cores and a 4.5M × 4.5M matrix in 1 minute using 4,352 Knights Landing cores. The paper appeared in IEEE IPDPS 2017. In particular, for the Knights-Landing architecture we also studied the performance of different algorithmic choices and different memory configurations of KNL, see Table 1.

Config	Haswell	Cache-Quad	F-Quad
Compute MatVec V with `GEMV`			
T	0.8	0.9	0.9
GF	14	12	12
Reevaluate V with `GEMM`			
T	4.4	7.4	7.5
GF	158	40	39
Compute MatVec V with `FUSED GEMM`			
T	1.1	1.1	1.2
GF	269	269	243

Table 1 *In this table we report comparison between Haswell (24 cores) and Intel Phi Knights Landing (68 cores) with different memory configurations.* **"T"** *is time (sec) and* **"GF"** *GFLOPS. Three different algorithmic variants for a GEMV with the kernel matrix. These schemes represent trade-offs between memory requirements an speed. The best variant is the FUSED GEMM one that although it is slightly slower (due to the increased number of flops) it delivers an order-of-magnitude savings in memory compared to the GEMV version for kernel matrices.*

Certain intermediate Schur complement matrices that appear in the factorization of a kernel matrix are hierarchical matrices and can be compressed, thus reducing the constants in the complexity estimate significantly. The challenge is that for those matrices we don't have a geometric structure, that is points and appropriate distances. We call algorithms for these types of problems as *"point-free kernel methods"*. To accomplish this, we used the fact that those matrices are positive definite (since they are related to Hessian and covariance operators), and as such, they define an *implicit* geometric structure.

Based on these observations, we designed `GOFMM` (geometry-oblivious FMM), a novel method that creates a hierarchical low-rank approximation, or *"compression,"* of an arbitrary dense symmetric positive definite (SPD) matrix. GOFMM enables an approximate matrix-vector multiplication in $N \log N$ or even N time, where N is the matrix size. Compression requires $N \log N$ storage and work. In general, our scheme belongs to the family of hierarchical matrix approximation methods. In particular, it generalizes the fast multipole method (FMM) to a purely algebraic setting by only requiring the ability to sample matrix entries. Neither geometric information (i.e., point coordinates) nor knowledge of how the matrix entries have been generated is required, thus the term *"geometry-oblivious."* Also, we introduced a shared-memory parallel scheme for hierarchical matrix computations that reduces synchronization barriers.

Dense SPD matrices appear in scientific computing, statistical inference, and data analytics. They appear in Cholesky and LU factorization, in Schur complement matrices for saddle point problems, in Hessian operators in optimization, in kernel methods for statistical learning, and in N-body methods and integral equations. In many applications, the entries of the input matrix K are given by $K_{ij} = \mathcal{K}(x_i, x_j) : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$, where $\mathcal{K}$ is a kernel function. Examples of kernel functions are radial basis functions, Green's functions, and angle similarity functions. For such *kernel matrices*, the input is not a matrix, but only the points $\{x_i\}_{i=1}^N$. The points are used to appropriately permute the matrix using spatial data structures. Furthermore, the construction of the sparse correction S uses nearest-neighbor structure of the input points. The low-rank matrices U, V can be either analytically computed using expansions of the kernel function, or semi-algebraically computed using fictitious points (or equivalent points), or using algebraic sampling-based methods that use geometric information. In a nutshell, geometric information is used in all aspects of an $\mathcal{H}$-matrix method.

In many cases however, such points and kernel functions are not available. For example, in dense graphs in data analysis (e.g., social networks, protein interactions). Related matrices include graph Laplacian operators and their inverses. Additional examples include frontal matrices and Schur complements in factorization of sparse matrices; Hessian operators in optimization; and kernel methods in machine learning without points (e.g., word sequences and diffusion on graphs).

GOFMM is inspired by the rich literature of algorithms for matrix sketching, hierarchical matrices, and fast multipole methods. Its unique feature is that by using only matrix evaluations it generalizes FMM ideas to compressing arbitrary SPD matrices. In more detail, our contributions are summarized below.

- A result from reproducing kernel Hilbert space theory is that any SPD matrix corresponds to a Gram matrix of vectors in some unknown Gram (or feature) space. Based on this result, the matrix entries are inner products, which we use to define distances. These distances allow us to design an efficient, purely algebraic FMM method.

- The key algorithmic components of GOFMM (and other hierarchical matrix and FMM codes) are tree traversals. We test parallel level-by-level traversals, *out-of-order* traversals using OpenMP's advanced task scheduling and an in-house tree-task scheduler. We found that scheduling significantly improves the performance when compared to level-by-level tree traversals. We also use this scheduling to support heterogeneous architectures.

- We conduct extensive experiments to demonstrate the feasibility of the proposed approach. We test our code on 22 different matrices related to machine learning, stencil PDEs, spectral PDEs, inverse problems, and graph Laplacian operators. We perform numerical experiments on Intel Haswell and KNL, Qualcomm ARM, and NVIDIA Pascal architectures. Finally, we compare with three state-of-the-art codes: HODLR, STRUMPACK, and ASKIT.

GOFMM also has several additional capabilities. If points and kernel functions (or Green's) function are available, they can be utilized in a similar way to our algebraic FMM code ASKIT. GOFMM currently

	HODLR			STRUMPACK			GOFMM		
case	ϵ_2	Comp	Eval	ϵ_2	Comp	Eval	ϵ_2	Comp	Eval
K02	6E−5	0.6	2.7	1E−4	9.2	0.6	2E−5	1.0	0.3
K04	6E−5	0.7	2.7	1E−4	507.8	7.8	2E−5	1.0	0.5
K07	7E−5	0.9	3.1	2E−4	528.4	8.2	4E−5	0.6	0.2
K12	6E−5	0.7	2.7	2E−4	18.8	0.8	1E−4	0.6	0.2
K17	1E−1	862.2	37.6	2E−1	663.4	8.2	9E−2	48.8	3.1
G03	3E−4	12.9	9.7	3E−2	29.8	1.3	8E−5	0.5	0.8

Figure 1 *In this figure, we compare GOFMM with two other open-source packages for six different SPD matrices. HODLR is a software library developed at Stanford. STRUMPACK is a software library developed at NERSC. "Case" indicates the matrix type. All these matrices are dense, symmetric, and positive definite. K02 is a the Hessian of an inverse source problem with the 5-point Laplacian, K04 is with the 5-point low-frequency Helmholtz, K07 is a covariance matrix on a domain with complicated geometry, K12 is a the Hessian of a source inversion problem with a spectral, variable coefficient Laplacian, K17 is the Hessian of a high-Peclet number advection-reaction-diffusion PDE with pseudo-spectral discretization. Finally, G03 is a dense Schur complement of a saddle point problem. The size of all these matrices are 65K. Also, ϵ_2 is the relative ℓ_2 norm between the exact matrix and its approximation; "Comp" is the time required to compress the matrix; and "Eval" is the time to perform a matrix-vector product with 256 right-hand sides. With the exception of K17, which is hard to compress, all the other matrices compress very well. Overall, the wall-clock time for GOFMM including the compression and the approximate matvec, 10× faster than an SGEMM. In addition, GOFMM (depending on the case) can be several orders of magnitude faster than STRUMPACK and HODLR.*

supports three different measures of distance: geometric point-based (if available), Gram-space ℓ^2 distance, and Gram-space angle distance. GOFMM has support for matvecs with multiple vectors, which is useful for Monte-Carlo sampling, optimization, and block Krylov methods. The overall complexity of both the compression and evaluation phases is $\mathcal{O}(N)$. This work appeared in the IEEE/ACM SC'17 conference.

Optimal experimental design: We have continued working on developing scalable parallel algorithms for optimal experimental design (OED) problems.

In the last period of the project, we addressed the problem of optimal experimental design (OED) for Bayesian nonlinear inverse problems governed by partial differential equations (PDEs). The inverse problem seeks to infer an infinite-dimensional parameter from experimental data observed at a set of sensor locations and from the governing PDEs. The goal of the OED problem is to find an optimal placement of sensors so as to minimize the uncertainty in the inferred parameter field.

Specifically, we seek an optimal subset of sensors from among a fixed set of candidate sensor locations. We formulate the OED objective function by generalizing the classical A-optimal experimental design criterion using the expected value of the trace of the posterior covariance. This expected value is computed through sample averaging over the set of likely experimental data. To cope with the infinite-dimensional character of the parameter field, we construct a Gaussian approximation to the posterior at the maximum a posteriori probability (MAP) point, and use the resulting covariance operator to define the OED objective function.

We used randomized trace estimation to compute the trace of this covariance operator, which is defined only implicitly. The resulting OED problem includes as constraints the system of PDEs characterizing the MAP point, and the PDEs describing the action of the covariance (of the Gaussian approximation to the posterior) to vectors. We control the sparsity of the sensor configurations using sparsifying penalty functions. Variational adjoint methods are used to efficiently compute the gradient of the PDE-constrained OED objective function. We elaborate our OED method for the problem of determining the optimal sensor configuration to best infer the coefficient of an elliptic PDE.

Furthermore, we conducted numerical experiments for inference of the log permeability field in a porous medium flow problem. Numerical results showed that the number of PDE solves required for the evaluation of the OED objective function and its gradient is essentially independent of both the parameter dimension and the sensor dimension (i.e., the number of candidate sensor locations). The number of quasi-Newton iterations for computing an OED also exhibits the same dimension invariance properties.

Development of domain-specific language for modeling UQ applications.

We have continued the development of Aspen domain language. With the anticipation of exascale architectures, energy consumption is becoming one of the critical design parameters, especially in light of the energy budget of 2030 Megawatts set by the U.S. Department of Energy. Understanding an applications execution pattern and its energy footprint is critical to improving application operation on a diverse heterogeneous architecture. Applying application-specific performance optimization can consequently improve energy consumption. However, this approach is only applicable to current systems. As we enter a new era of exascale architectures that is projected to contain more complex memory hierarchies, increased levels of parallelism, heterogeneity in hardware, and complex programming models and techniques, energy and performance management is getting more cumbersome. We therefore propose techniques that predict the energy consumption beforehand or at runtime to enable proactive tuning. Such energy prediction approaches must be generic and adapt themselves at runtime to changing application and hardware configurations. Most existing energy estimation and prediction approaches are empirical in nature and thus tied to current systems. To overcome this limitation, we propose two energy estimation techniques: ACEE (Algorithmic and Categorical Energy Estimation), which uses a combination of analytical and empirical modeling techniques; and AEEM (Aspens Embedded Energy Estimation), a system-level analytical energy estimation technique. Both of these models incorporate the Aspen domain specific language for performance modeling. We present the methodologies of these two models and test their accuracy using five proxy applications.

In addition, to DSLs, we also consider frameworks that allow use of additional memory resources, which are likely to be necessary in the context of uncertainty quantification problems. Indeed, a surprising development in recently announced HPC platforms is the addition of, sometimes massive amounts of, persistent (nonvolatile) memory (NVM) in order to increase memory capacity and compensate for plateauing I/O capabilities. However, there are no portable and scalable programming interfaces using aggregate NVM effectively. This paper introduces Papyrus: a new software system built to exploit emerging capability of NVM in HPC architectures. Papyrus (or Parallel Aggregate Persistent -YRU- Storage) is a novel programming system that provides features for scalable, aggregate, persistent memory in an extreme-scale system for typical HPC usage scenarios. Papyrus mainly consists of Papyrus Virtual File System (VFS) and Papyrus Template Container Library (TCL). Papyrus VFS provides a uniform aggregate NVM storage image across diverse NVM architectures. It enables Papyrus TCL to provide a portable and scalable high-level container programming interface whose data elements are distributed across multiple NVM nodes without requiring the user to handle complex communication, synchronization, replication, and consistency model. We evaluated Papyrus on two HPC systems, including UTK Beacon and NERSC Cori, using real NVM storage devices and the results are very encouraging.

4.2. Theoretical foundations

Participating researchers:Bui-Thanh, Ghattas, Stadler, Moser, Oliver, Biros, Marzouk

Prior distributions for Bayesian inference that rely on the l_1-norm of the parameters are of considerable interest, in part because they promote parameter fields with less regularity than Gaussian priors (e.g., discontinuities and blockiness). These l_1-type priors include the total variation (TV) prior and the

Besov space $B^s_{1,1}$ prior, and in general yield non-Gaussian posterior distributions. Sampling from these posteriors is challenging, particularly in the inverse problem setting where the parameter space is high-dimensional and the forward problem may be nonlinear. This paper extends the randomize-then-optimize (RTO) method, an optimization-based sampling algorithm developed for Bayesian inverse problems with Gaussian priors, to inverse problems with l_1-type priors. We use a variable transformation to convert an l_1-type prior to a standard Gaussian prior, such that the posterior distribution of the transformed parameters is amenable to Metropolized sampling via RTO. We demonstrate this approach on several deconvolution problems and an elliptic PDE inverse problem, using TV or Besov space $B^s_{1,1}$ priors. Our results show that the transformed RTO algorithm characterizes the correct posterior distribution and can be more efficient than other sampling algorithms. The variable transformation can also be extended to other non-Gaussian priors.

In many inverse problems, model parameters cannot be precisely determined from observational data. Bayesian inference provides a mechanism for capturing the resulting parameter uncertainty, but typically at a high computational cost. This work introduces a multiscale decomposition that exploits conditional independence across scales, when present in certain classes of inverse problems, to decouple Bayesian inference into two stages: (1) a computationally tractable coarse-scale inference problem, and (2) a mapping of the low-dimensional coarse-scale posterior distribution into the original high-dimensional parameter space. This decomposition relies on a characterization of the non-Gaussian joint distribution of coarse- and fine-scale quantities via optimal transport maps. We demonstrate our approach on a sequence of inverse problems arising in subsurface flow, using the multiscale finite element method to discretize the steady state pressure equation. We compare the multiscale strategy with full-dimensional Markov chain Monte Carlo on a problem of moderate dimension (100 parameters) and then use it to infer a conductivity field described by over 10000 parameters.

4.3. Coarse-graining using Bayesian formulations

Papers have been submitted for publication; participating researchers: Biros, Oden

We have continued the development OPAL, the Occam-Plausibility Algorithm (Farrell et al. in J Comput Phys 295:189208, 2015) in which the use of Bayesian model plausibilities is replaced with information-theoretic methods, such as the Akaike information criterion and the Bayesian information criterion. Applications to complex systems of coarse-grained molecular models approximating atomistic models of polyethylene materials are described. All of these model selection methods take into account uncertainties in the model, the observational data, the model parameters, and the predicted quantities of interest. We compared all of the models chosen by Bayesian model selection criteria and those chosen by the information-theoretic criteria.

5. Publications acknowledging this award

- Ghattas, Omar and Isaac, Tobin and Petra, Noémi and Stadler, Georg, *Scalable Algorithms for Bayesian Inference of Large-Scale Models from Large-Scale Data*, International Conference on Vector and Parallel Processing, 2016
  http://dx.doi.org/10.1007/978-3-319-61982-8

- A. Alexanderian, N. Petra, G. Stadler, and O. Ghattas, *A fast and scalable method for A-optimal design of experiments for infinite-dimensional Bayesian nonlinear inverse problems*, SIAM Journal on Scientific Computing, 38(1):A243A272, 2016
  http://dx.doi.org/10.1137/140992564

- A. Alexanderian, N. Petra, G. Stadler, and O. Ghattas, *A fast and scalable method for A-optimal design of experiments for infinite-dimensional Bayesian nonlinear inverse problems*, SIAM Journal on Scientific Computing, 38(1):A243A272, 20
  http://dx.doi.org/10.1137/140992564

- A. Alexanderian, P. Gloor, and O. Ghattas, *On Bayesian A- and D-optimal experimental designs in infinite dimensions*, Bayesian Analysis, 11(3):671695, 2016.
  http://dx.doi.org/10.1214/15-BA969

- Wang, Z., Bardsley, J. M., Solonen, A., Cui, T., Marzouk, Y. M., *Bayesian Inverse Problems with l_1 Priors: A Randomize-Then-Optimize Approach*. SIAM Journal on Scientific Computing, 39(5), S140-S166, 2017
  https://doi.org/10.1137/16M1080938

- Parno M, Moselhy T, Marzouk Y., *A multiscale strategy for Bayesian inference using transport maps*, SIAM/ASA Journal on Uncertainty Quantification. 2016 Oct 11;4(1):1160-90
  https://doi.org/10.1137/15M1032478

- Farrell-Maupin K, Oden JT, *Adaptive selection and validation of models of complex systems in the presence of uncertainty*, Research in the Mathematical Sciences. 2017 Dec 1;4(1):14
  https://doi.org/10.1186/s40687-017-0104-2

- M Umar, SV Moore, JS Meredith, JS Vetter, KW Cameron, *Aspen-based performance and energy modeling frameworks*, Journal of Parallel and Distributed Computing, 2017
  https://doi.org/10.1016/j.jpdc.2017.11.005

- Kim J, Sajjapongse K, Lee S, Vetter JS, *Design and Implementation of Papyrus: Parallel Aggregate Persistent Storage*, 2017 IEEE International 2017 May 29 (pp. 1151-1162). IEEE
  https://doi.org/10.1109/IPDPS.2017.72

- Chenhan D. Yu, James Levitt, Severin Reiz, and George Biros, *Geometry-Oblivious FMM for Compressing Dense SPD Matrices*, Proceedings of the SC2017, The International Conference for High Performance Computing, Networking, Storage and Analysis, IEEE/ACM, Denver, Colorado, (18% acceptance rate), November 2017
  http://dx.doi.org/10.1145/3126908.3126921

- Amir Gholami. Klaudius Scheufele, Christos Davatzikos, Andreas Mang, Miriam Mehl, and George Biros, *Framework for Scalable Biophysics-based Image Analysis*, Proceedings of the SC2017, The International Conference for High Performance Computing, Networking, Storage and

Analysis, IEEE/ACM, Denver, Colorado, (18% acceptance rate), November 2017 (**Best student paper**)
http://dx.doi.org/10.1145/3126908.3126930

- Chenhan D. Yu and William B. March, and George Biros, *An $N \log N$ Parallel Fast Direct Solver for Kernel Matrices*, 31th IEEE International Parallel & Distributed Processing Symposium (IEEE IPDPS 2017), Orlando, USA, May 2017
https://dx.doi.org/10.1109/IPDPS.2017.10

- Andreas Mang, Amir Gholami and George Biros, *Distributed-Memory Large Deformation Diffeomorphic 3D Image Registration*, Proceedings of the SC2016, The International Conference for High Performance Computing, Networking, Storage and Analysis, IEEE/ACM, Salt Lake City, Utah, (19% acceptance rate), November 2016
http://dl.acm.org/citation.cfm?id=3014904.3015001

- Arash Bakhtiari, Dhairya Malhotra, Amir Raoofy, Miriam Mehl, Hans-Joachim Bungartz, and George Biros , *A Parallel Aribtrary-Order Accurate AMR Algorithm for the Scalar Advection-Diffusion Equation*, Proceedings of the SC2016, The International Conference for High Performance Computing, Networking, Storage and Analysis, IEEE/ACM, Salt Lake City, Utah, (19% acceptance rate), November 2016
http://dl.acm.org/citation.cfm?id=3014904.3014963

- Chenhan D. Yu, William March, and George Biros, *INV-ASKIT:A Parallel Fast Direct Solver for Kernel Matrices*, 30th IEEE International Parallel & Distributed Processing Symposium (IEEE IPDPS 2016), Chicago, IL, May 2016
http://dx.doi.org/10.1109/IPDPS.2016.12

- A. Mang and G. Biros, *A Semi-Lagrangian two-level preconditioned Newton-Krylov solver for constrained diffeomorphic image registration*, SIAM Journal on Scientific Computing, 39 (6), pp. B1064–B1101
http://dx.doi.org/10.1137/16M1070475

- G. Kabacaouglu, B. Quaife and G. Biros, *Quantification of mixing in vesicle suspensions using numerical simulations in two dimensions*, Physics of Fluids, 29 (2), 2017
http://dx.doi.org/10.1063/1.4975154

- H .Akbari et al, *Imaging surrogates of infiltration obtained via multiparametric imaging pattern analysis predict subsequent location of recurrence of glioblastoma*, Neurosurgery, 78 (4), pp. 572–580, 2016
https://dx.doi.org/10.1227/NEU.0000000000001202

- B. Xiao and G. Biros, *Parallel Algorithms for Nearest Neighbor Search Problems in High Dimensions*, SIAM Journal on Scientific Computing, 38 (5) pp. S667-S699, 2016
http://dx.doi.org/10.1137/15M1026377

- A. Mang and G. Biros, *Constrained H^1-regularization schemes for diffeomorphic image registration*, SIAM Journal on Imaging Sciences, 9 (3), pp. 1154–1194, 2016
http://dx.doi.org/10.1137/15M1010919

- A. Gholami, D. Malhotra, H. Sundar, and G. Biros, *FFT, FMM, or multigrid? A comparative study of state-of-the-art Poisson solvers*, SIAM Journal on Scientific Computing, 38 (3), pp. 280–306, 2016
  http://dx.doi.org/10.1137/15M1010798

- D. Malhotra and G. Biros, *Algorithm 967: A distributed memory fast multipole method for volume potentials*, ACM Transactions on Mathematical Software, 43 (2), pp. 1–27, 2016
  http://doi.acm.org/10.1145/2898349

- W. B. March, B. Xiao, C. D. Yu, and G. Biros, *ASKIT: An efficient, parallel library for high-dimensional kernel summations*, SIAM Journal on Scientific Computing, 38 (5) pp. S720-S749, 2016
  http://dx.doi.org/10.1137/15M1026468

- B. Quaife and G. Biros, *Adaptive Time Stepping for Vesicle Suspensions*, Journal on Computational Physics, pp. 478–499, 306 (1), 2016
  http://dx.doi.org/j.jcp.2015.11.050

- W. B. March and G. Biros, *Far-Field Compression for Fast Kernel Summation Methods in High Dimensions*, Applied and Computational Harmonic Analysis, 43 pp. 39–75, 2017
  http://dx.doi.org/10.1016/j.acha.2015.09.007

- W. B. March, B. Xiao, and G. Biros, *ASKIT: Approximate Skeletonization Kernel-Independent Treecode in High Dimensions*, SIAM Journal on Scientific Computing, 37 (2), 2015, pp. A1089–A1110
  http://dx.doi.org/10.1137/140989546

- W. B. March and G. Biros, *Far-Field Compression for Fast Kernel Summation Methods in High Dimensions*, in review
  http://arxiv.org/abs/1409.2802

- A. Gholami, D. Malhotra, H. Sundar, and G. Biros, *FFT, FMM, or multigrid? A comparative study of state-of-the-art Poisson solvers*, SIAM Journal on Scientific Computing, in review
  http://arxiv.org/abs/1408.6497

- William B. March, Bo Xiao, Chenhan Yu, and George Biros, *An algebraic parallel treecode in arbitrary dimensions*, Proceedings of the 29th IEEE International Parallel and Distributed Computing Symposium (IPDPS 2015), IEEE, Hyderabad, India, May 2015

- Ahmed Khawaja, Jiajun Wang, Dhairya Malhotra, Andreas Gerstlauer, George Biros and Lizy John, *Performance Analysis of HPC Applications with Irregular Tree Data Structures*, Proceedings of the 20th IEEE International Conference on Parallel and Distributed Systems (ICPADS 2014), IEEE, Hsinchu, Taiwan, December 2014

- B. Quaife and G. Biros, *Adaptive Time Stepping for Vesicle Suspensions*, Journal on Computational Physics, pp. 1–28, to appear

- Dhairya Malhotra, Amir Gholami, and George Biros, *A volume integral equation Stokes solver for problems with variable coefficients*, Proceedings of SC2014, IEEE/ACM, New Orleans, LA, November 2014, (**Best Student Paper Finalist**)
  dx.doi.org/10.1109/SC.2014.13

- H. Sundar and O. Ghattas, *A Nested Partitioning Algorithm for Adaptive Meshes on Heterogeneous Clusters*, Proceedings of ICS'15: 2015 ACM International Conference on Supercomputing, to appear.

- T. Isaac, N. Petra, G. Stadler, O. Ghattas, *Scalable and efficient algorithms for the propagation of uncertainty from data through inference to prediction for large-scale problems, with application to flow of the Antarctic ice sheet*, Journal of Computational Physics, 296(1):348–368, 2015.

- T. Bui-Thanh and O. Ghattas, *A scalable MAP solver for Bayesian inverse problems with Besov priors*, Inverse Problems and Imaging, 9(1):27–54, 2015.

- T. A. Oliver, G. Terejanu, C. S. Simmons, R.D. Moser, *Validating predictions of unobserved quantities*, Computer Methods in Applied Mechanics and Engineering Volume 283, 1 January 2015, Pages 13101335

- K. Farrell, J.T. Oden, D. Faghihi, *A Bayesian framework for adaptive selection, calibration, and validation of coarse-grained models of atomistic systems*, Journal of Computational Physics, 295:189-208, 2015.

- J.T. Oden, K. Farrell, D. Faghihi, *Estimation of error in observables of coarse-grained models of atomic systems*, Advanced Modeling and Simulation in Engineering Sciences, 2:5, 2015

- K. Farrell and J.T. Oden, *Model misspecification and plausibility*, ICES Report 14-21. August 2015

- K. Farrell, *Selection, Calibration, and Validation of Coarse-Grained Models of Atomistic Systems*, PhD thesis, University of Texas at Austin, May 2015.

6. Presentations acknowledging this award

- U. Villa, N. Petra, O. Ghattas, *hIPPYlib: An Extensible Software Framework for Large-Scale Deterministic and Linearized Bayesian Inverse*, Texas Applied Mathematics and Engineering Symposium, Austin, TX, US, September 2017

- U. Villa, O. Ghattas, *Derivative-informed MCMC for Bayesian Calibration of Stochastic PDE Models*, SIAM Annual Meeting, July 10-14, 2017, Pittsburgh, PA, US

- U. Villa, O. Ghattas, *Hessian-based Sampling Techniques for Bayesian Inverse Problems with Stochastic PDE Forward Model*, Applied Inverse Problems, May 29-Jun 2, 2017, Hangzhou, China

- U. Villa, T. Oliver, B. Moser, O. Ghattas, *Bayesian Calibration of Inadequate Stochastic PDE Models*, SIAM Conference on Computer Science and Engineering, Feb 27-March 3, 2017, Atlanta, GA, US

- O. Ghattas, *Big data meets big models: Towards exascale Bayesian inverse problems*, 24th International Congress of Theoretical and Applied Mechanics (ICTAM 2016), Montreal, Quebec, Canada, August 2026, 2016

- T.A. Oliver, M. Lee, C. Simmons, D. Sondak, R.D. Moser, *Towards Model Inadequacy Repre- sentations for Flamelet-Based RANS Combustion Simulations*, 69th Annual Meeting of the APS Division of Fluid Dynamics, Portland, Oregon, November 20-22, 2016.

- C. Borges and G. Biros, *Reconstruction of a Compactly Supported Sound Profile in the Presence of Noisy Background Random Medium*, 14th Annual Conference on Frontiers in Applied and Computational Mathematics (FACM '17), NJIT, New Jersey NJ, July 2017
- A. Mang and G. Biros, *Preconditioners for the reduced space Hessian in hyperbolic optimal control problems*, Conference on Preconditioning Techniques for Scientific and Industrial Applications Vancouver, CA, July 2017
- A. Gholami and G. Biros, *Fast algorithms for inverse problems with parabolic PDE constraints with application to biophysics-based image analysis*, Institute for Computational and Mathematical Engineering, Stanford University, Palo Alto CA, May 2017
- G. Biros, *Exascale N-body algorithms for data analysis and simulation*, 7th AICS International Symposium on emerging numerical techniques for exascale and post-Moore era, Riken Advanced Institute for Computational Science, Kobe, Japan, February 2017
- G. Biros, *Scalable algorithms for kernel matrix approximation*, Big data meets computation workshop, Institute for Pure and Applied Mathematics, UCLA, Los Angeles CA, January 2017
- G. Biros, *Fast algorithms for hierarchical matrices*, Applied Mathematics Colloquium, University of North Carolina at Chapel Hill, Chapel Hill NC, October 2016
- G. Biros, *Parallel Hierarchical Matrix Algorithms for Kernel Methods*, Invited session on Algorithms for Extreme Scale in Practice, ISC High Performance Conference, June 19–23, 2016 Frankfurt, Germany
- G. Biros, *Extreme Scale Algorithms For The Inexact Factorization Of Kernel Matrices*, KAUST Workshop on Scalable Hierarchical Algorithms for Extreme Computing, 9–11 May 2016, KAUST, Saudi Arabia
- G. Biros, *Exascale N-body Algorithms for Data Analysis and Simulation*, **Plenary**, German Priority Program "Software for Exascale Computing" (SPPEXA) Symposium, Leibniz Supercomputing Centre, Garching, Germany, January 2016
- G. Biros, *Bayesian inverse problems for spatial random fields*, D^3: Deformation, Defects, Diagnosis Symposium, University of Pennsylvania, Penn Institute for Computational Science, Philadelphia, Pennsylvania, May 2015
- G. Biros, *Scalable N-body methods for kernel sums in high dimensions*, International Symposium on Big Data and Predictive Computational Modeling, Technical University of Munich, Institute for Advanced Study, Munich, Germany, May 2015
- G. Biros, *An algebraic treecode for fast kernel sums in high dimensions*, Computational and Applied Mathematics Colloquium, Rice University, Houston, Texas, April 2015
- G. Biros, *N-body methods in computational science and engineering*, in Minisymposium: Fast Multipole Methods Maturing at 30 years, SIAM Conference on Computational Science and Engineering, Salt Lake City, UT, March 2015
- G. Biros, *A fast N-body algorithm for kernel sums in high dimensions*, in Minisymposium: UQ in Large Scale Computing, SIAM Conference on Computational Science and Engineering, Salt Lake City, UT, March 2015

- G. Biros, *N-body algorithms in computational science and engineering*, Institute for Advanced Studies Colloquium, Technical University of Munich, Munich, Germany, June 2014

- G. Biros, Fast algorithms for the evaluation for volume integral equations on hybrid architectures, **Keynote** Workshop: Exploiting Different Levels of Parallelism for Exascale Computing, ACM International Conference on Supercomputing, Munich, Germany, June 2014

- A. Davis, P. Heimbach, Y. Marzouk, *Predicting the probability of West Antarctic collapse: a Bayesian approach*, International Society for Bayesian Analysis World Meeting, Cancun, Mexico. July 2014.

- A. Davis, M. Parno, P. Conrad, Y. Marzouk, *MUQ (MIT Uncertainty Quantification): Flexible software for connecting algorithms and applications.*, SIAM CSE 2015. Salt Lake City, UT. March 2014.

- A. Davis, *Predicting marine ice sheet dynamics and volume loss under uncertainty*, MIT EAPS seminar. April 2015

- A. Davis, P. Heimbach, Y. Marzouk, *Uncertain prediction of marine ice sheet dynamics and volume loss*, SIAM GS15, Stanford, CA. June 2015

- O. Ghattas, *Bayesian inversion for large scale Antarctic ice sheet flow*, 52nd Meeting of the Society for Natural Philosophy, Rio de Janeiro, Brazil, October 22–24, 2014.

- O. Ghattas, *Mathematical and computational challenges in large-scale Bayesian inverse problems arising in the flow of the Antarctic ice sheet*,Prospects in Applied Mathematics, Chicago, IL, October 19–20, 2014.

- O. Ghattas, *Optimal Experimental Design for Large-Scale Bayesian Nonlinear Inverse Problems*, Applied Inverse Problems 2015, Helsinki, Finland, May 25–29, 2015.

- O. Ghattas, *Hessian-based Implicit Dimension Reduction for Large-scale Bayesian Inverse Problems*, Big Data and Predictive Computational Modeling, Technical University of Munich, Munich, Germany, May 18-21, 2015.

- O. Ghattas, *Scalable and Efficient Algorithms for the Propagation of Uncertainty from Data through Inference to Prediction for Flow of the Antarctic Ice Sheet*, Pan-American Congress on Computational Mechanics (PANACM 2015), Buenos Aires, Argentina, April 27–29, 2015.

- O. Ghattas, *From data to prediction via reduced parameter-to-observable maps: Applications to Antarctic ice sheet dynamics*, SIAM Conference on Computational Science and Engineering, Salt Lake City, UT, March 11-15, 2015.

- O. Ghattas, *Bayesian Inversion for Large Scale Antarctic Ice Sheet Flow*, American Geophysical Union Fall Meeting, San Francisco, CA, December 15–19, 2014.

- O. Ghattas, *Perspective on Co-Design: Uncertainty Quantification Challenges at Extreme Scale*, International Workshop on Co-Design, Guangzhou, China, November 6–10, 2014.

- O. Ghattas, *Sparse structure-exploiting methods for large-scale Bayesian inverse problems*, Workshop on Approximation, Integration, and Optimization, Institute for Computational and Experimental Research in Mathematics (ICERM), Brown University, September 29–October 3, 2014.

- O. Ghattas, *Big data meets big models: Large-scale Bayesian inference, with application to inverse modeling of Antarctic ice sheet dynamics*, 11th World Congress on Computational Mechanics, Barcelona, Spain, July 20–25, 2014.

- O. Ghattas, *Bayesian Inversion for Large Scale Antarctic Ice Sheet Flow*, Caltech Computing + Mathematical Sciences Colloquium Series, Caltech, Pasadena, CA, May 18, 2015.

- O. Ghattas, *Bayesian Inversion for Large Scale Antarctic Ice Sheet Flow*, Distinguished Speaker Series in Scientific Computing, Centre for Scientific Computing and PIMS, Simon Fraser University, Burnaby, BC, Canada, April 10, 2015.

- O. Ghattas, *Bayesian Inversion for Large Scale Antarctic Ice Sheet Flow*, ICES Babuska Forum Seminar, Austin, TX, February 13, 2015.

- J. S. Meredith and J. Vetter, *Exploring Emerging Technologies in the Extreme Scale HPC Co-Design Space with Holistic Performance Modeling*, Exascale Applications and Software Conference (EASC), Edinburgh, UK, April 22, 2015

- J. S. Meredith and J. Vetter, *Aspen Hackathon and Tutorial*, web-based and in person at Oak Ridge, TN, April 18, 2015

- R.E. Morrison and R.D. Moser, *Representing Model Inadequacy in Combustion Kinetics*, Bulletin of the American Physical Society, 67th Annual Meeting of the APS Division of Fluid Dynamics, November 2325, 2014; San Francisco, California

- September 19, 2013. "The Emergence of Predictive Computational Science: Validation and Verification of Computational Models of Complex Physical Systems," Presented by J.T. Oden, Florida International University, Miami, FL.

- October 11, 2013. "The Emergence of Predictive Computational Science: Validation and Verification of Computational Models of Complex Physical Systems," Presented by J.T. Oden, Distinguished Lecture, Scientific Computing and Imagine Institute, The University of Utah, Salt Lake City, UT.

- November 6, 2013, "The Emergence of Predictive Computational Science: Validation and Verification of Computational Models of Complex Physical Systems," Presented by J.T. Oden, 44th Henry M. Shaw Lecture, North Carolina State University, Raleigh, NC.

- February 18–21, 2014, Parallel Algorithms for Prior Functions in Bayesian Inference, George Biros, SIAM Conference on Parallel Processing, Portland, OR

- February 18–21, 2014, Reexamining Algorithm-Based Fault Tolerance for Exascale Architectures, Dong Li and Jeff Vetter, SIAM Conference on Parallel Processing, Portland, OR

- February 18–21, 2014, SIAM Conference on Parallel Processing, Portland, OR

- March 31-April 3, 2014. "Calibration, Validation, and Model Uncertainty of Coarse- Grained Models of Atomic Systems," Presented by K. Farrell, SIAM Conference on Uncertainty Quantification, Savannah, GA.

- March 31-April 3, 2014, Statistical Inversion for Basal Parameters for the Antarctic Ice Sheet Tobin Isaac, Noemi Petra, Georg Stadler, and Omar Ghattas, SIAM Conference on Uncertainty Quantification, Savannah, GA.

- March 31-April 3, 2014A Scalable MAP-Based Algorithm for Optimal Experimental Design for Large-Scale Bayesian Inverse Problems Alen Alexanderian, Noemi Petra, Georg Stadler, and Omar Ghattas, SIAM Conference on Uncertainty Quantification, Savannah, GA.

- March 31-April 3, Scalable Algorithms for Bayesian Inverse Problems and Optimal Experimental Design with Applications to Large-scale Complex Systems Alen Alexanderian, Omar Ghattas, Tobin Isaac, James R. Martin, Noemi Petra, and Georg Stadler, SIAM Conference on Uncertainty Quantification, Savannah, GA.

- March 31-April 3, Scalable Nearest Neighbor Search Algorithms in High Dimensions George Biros and Bo Xiao, SIAM Conference on Uncertainty Quantification, Savannah, GA.

- March 31-April 3, Parallel Multiscale Algorithms for Construction of Likelihood and Prior Densities for Bayesian Inverse Problems, George Biros, Bill March, Bo Xiao, SIAM Conference on Uncertainty Quantification, Savannah, GA.

- March 31-April 3, Optimal Bayesian Experimental Design in the Presence of Model Error abstract Youssef M. Marzouk, SIAM Conference on Uncertainty Quantification, Savannah, GA.

- March 31-April 3, Dimension-independent Likelihood-informed MCMC Samplers, Tiangang Cui, Youssef M. Marzouk, SIAM Conference on Uncertainty Quantification, Savannah, GA.

- March 31-April 3,Sequential Experimental Design Using Dynamic Programming and Optimal Maps, Xun Huan and Youssef M. Marzouk, SIAM Conference on Uncertainty Quantification, Savannah, GA.

- March 31-April 3, Probabilistic Representations of Model Inadequacy for RANS Turbulence Models, Todd Oliver and Robert D. Moser, SIAM Conference on Uncertainty Quantification, Savannah, GA.

- May 7-9, 2014. "Validation and Selection of Multiscale Models of Complex Systems," Presented by J.T. Oden, ASME 2014 Verification and Validation Symposium, Las Vegas, NV.

- May 79, 2014, Big data, sparse information: Bayesian inference for large-scale models, with application to inverse

  modeling of Antarctic ice sheet dynamics, International Conference on Scientic Computing at Extreme Scale, Shanghai Jiao Tong University, Shanghai, China,

- July 714, 2014, Big data, sparse information: Bayesian inference for large-scale models, with application to inverse

  modeling of Antarctic ice sheet dynamics,. To be presented by O. Ghattas, SIAM Annual Meeting, Chicago, IL,

- June 23, 2014, Big data, sparse information: Bayesian inference for large-scale models, with application to inverse

  modeling of Antarctic ice sheet dynamics, To be presented by O. Ghattas, PASC14 Conference, Zurich, Switzerland

Besides the presentations, let us remark that we have organized several minisymposia in UQ and HPC mainly in the SIAM conferences but elsewhere as well.