samsi

NSF·Duke·NCSU·UNC

# MUMS Program Opening Workshop
# August 20-24, 2018

## SPEAKER TITLES/ABSTRACTS

**Amy Braverman**
Jet Propulsion Laboratory, California Institute of Technology

“Model Uncertainty in Data Fusion for Remote Sensing”

Data fusion is the process of combining data from different sources to enhance the utility of the combined product. In remote sensing, input data sources are typically massive, noisy, and have different spatial supports and sampling characteristics. We take an inferential approach to this data fusion problem: we seek to infer a true but not directly observed spatial (or spatio-temporal) field from heterogeneous inputs. We use a statistical model to make these inferences, but like all models it is at least somewhat uncertain. In this talk, we will discuss our experiences with the impacts of these uncertainties and some potential ways addressing them.

**Jenný Brynjarsdóttir**
Case Western Reserve University

“Model Discrepancy and Physical Parameters in Calibration and Prediction of Computer Models”

The main goal of calibration is usually to improve the predictive performance of the simulator but the values of the parameters in the model may also be of intrinsic scientific interest in their own right. As an example of the latter we will discuss CO2 retrievals from the the Orbiting Carbon Observatory 2 (OCO-2). In order to make appropriate use of observations of the physical system it is important to recognize model discrepancy, the difference between reality and the simulator output. We illustrate through a simple example that an analysis that does not account for model discrepancy may lead to biased and over-confident parameter estimates and predictions. The challenge with incorporating model discrepancy in statistical inverse problems is being confounded with calibration parameters, which will only be resolved with meaningful priors. For our simple example, we model the model-discrepancy via a Gaussian process and demonstrate that through accounting for model discrepancy our prediction within the range of data is correct. We will then discuss the effect of model discrepancy in CO2 retrievals. This is joint work with Anthony O'Hagan, University of Sheffield, and Jonathan Hobbs and Amy Braverman at the Jet Propulsion Laboratory.

**Kevin Carlberg**
Sandia National Laboratories

“Machine-Learning Error Models for Quantifying the Epistemic Uncertainty in Low-Fidelity Models”

Uncertainty-quantification tasks are often ``many query" in nature, as they require repeated evaluations of a model that often corresponds to a parameterized system of nonlinear equations (e.g., arising from the spatial discretization of a PDE). To make this task tractable for large-scale models, low-fidelity models (e.g., reduced-order models, coarse-mesh solutions) must be employed. However, such approximations introduce additional error, which may be treated as a source of epistemic uncertainty that must be quantified to ensure rigor in the ultimate UQ result.

We present a new approach to quantify the error (i.e., epistemic uncertainty) introduced by these low-fidelity models approximations. The approach (1) engineers features that are informative of the error using concepts related to dual-weighted residuals and rigorous error bounds, and (2) applies machine learning regression techniques (e.g., artificial neural networks, random forests, support vector machines) to construct a statistical model of the error from these features. We consider both (signed) errors in quantities of interest, as well as global state-space error norms. We present several examples to demonstrate the effectiveness of the proposed approach compared to more conventional feature and regression choices. In each of the examples, the predicted errors have a coefficient of determination value of at least 0.998.

**Peter Challenor**
University of Exeter

“The Isaac Newton Institute Uncertainty Quantification Programme: A Personal Perspective”

Between January and June 2018 the Isaac Newton Institute for Mathematical Sciences in the UK ran a six month programme on uncertainty quantification. The aim of the programme was to bring together the applied mathematics/numerical analysis and the statistical communities, who have different approaches to the problem on quantifying uncertainty in complex numerical models. Despite joint initiatives from groups such as SIAM and the ASA on journals and conferences the to communities remain separate and there was little understanding from one group on what the other does. The programme was organised by Peter Challenor (University of Exeter), Max Gunzberger (Florida State University), Catherine Powell (University of Manchester) and Henry Wynn (London School of Economics). Our core theme were: surrogate models; multilevel, multi-scale, and multi-fidelity methods; dimension reduction methods; inverse UQ methods; and careful and fair comparisons. INI programme participants attend the programme for up six months and have opportunities to work together in a collaborative way. Most of our participants attended for between 2-4 weeks. In addition to the participants we had a number of workshops. Four one week workshops on:  key UQ methodologies and motivating applications (an introductory workshop to give introductions to UQ methodologies from both traditions);, surrogate models for UQ in complex systems; reducing dimensions and cost for UQ in complex systems; and UQ for inverse problems in complex systems; and two one day workshops aimed at industry and other stakeholders. Most of the talks from the workshops are available on line. I will outline what happened during the programme and give my personal views on the achievements f the programme and what is still left to do.

**Merlise Clyde**
Duke University

“Model Uncertainty and Uncertainty Quantification”

The Bayesian paradigm provides a coherent approach for quantifying uncertainty given available data and prior information. Aspects of uncertainty that arise in practice include uncertainty regarding parameters within a model, the choice of model, and propagation of uncertainty in parameters and models for predictions. In this talk I will present Bayesian approaches for addressing model uncertainty given a collection of competing models including model averaging and ensemble methods that potentially use all available models and will highlight computational challenges that arise in implementation of the paradigm.

**Michael Frenklach**
University of California, Berkeley

"Bound-to-Bound-Data-Collaboration: Prediction on the Feasible Set"

The methodology of Bound-to-Bound-Data-Collaboration (abbreviated B2BDC) deploys semidefinite-programming algorithms, where the initial bounds on unknowns are combined with initial bounds of experimental data to produce new uncertainty bounds for the unknowns that are consistent with the data and, finally, deterministic uncertainty bounds for prediction in new settings. The presentation will review the current state of the B2BDC framework, emphasizing fundamental aspects of data-model analysis and interpolated prediction for physically-based models.

**Edward George**
Wharton, University of Pennsylvania

"Quantifying Nonparametric Modeling Uncertainty with BART"

For the canonical regression setup where one wants to discover the relationship between Y and a p-dimensional vector x, BART (Bayesian Additive Regression Trees) approximates the conditional mean E[Y|x] with a sum of regression trees model, where each tree is constrained by a regularization prior to be a weak learner. Fitting and inference are accomplished via a scalable iterative Bayesian backfitting MCMC algorithm that generates samples from a posterior. Effectively, BART is a nonparametric Bayesian regression approach which uses dimensionally adaptive random basis elements. Motivated by ensemble methods in general, and boosting algorithms in particular, BART is defined by a statistical model: a prior and a likelihood. This approach enables full posterior inference including point and interval estimates of the unknown regression function as well as the marginal effects of potential predictors. By keeping track of predictor inclusion frequencies, BART can also be used for model-free variable selection. To further illustrate the modeling flexibility of BART, we introduce two elaborations, MBART and HBART. Exploiting the potential monotonicity of E[Y|x] in components of x, MBART incorporates such monotonicity with a multivariate basis of monotone trees. To allow for the possibility of heteroscedasticity, HBART incorporates an additional product of regression trees model component for the conditional

**Roger Ghanem**
University of Southern California

"Modeling and Algorithmic Aspects of UQ for Material with Multiscale Behavior"

Increasingly, materials can be designed and built as a system, with constituents and components interacting with each other both locally and via longer range coupling. Examples of these materials include composites and multifunctional materials as well as materials synthesized through additive manufacturing processes. In all these instances, physical interactions are mediated through thin interfaces thus exacerbating the effect of compositional and functional gradients. In many such instances, small perturbations in physical specifications, be they geometrical mechanical or chemical, for any of the constituents, have tangible implications on system-level performance.

Accounting for many of the relevant uncertainties is hampered by the nature of physical interactions that magnify the interplay between modeling deviations and parametric deviations. It is also challenged by the computational complexity required for simultaneously resolving behaviors that are relevant to damage nucleation and propagation across constituents and components.

In this talk I will describe our experience in tackling some of these challenges in the context of NCF composites. Our efforts to link manufacturing processes with performance and cost analyses required the development of novel numerical algorithms that facilitated a more thorough statistical exploration of model validation issues. Our ability to iterate on model development and experiments permitted us to chart a path towards model improvement and ultimately validation.

**Dave Higdon**
Virginia Tech University

"Extrapolation: The Art of Connecting Model-Based Predictions to Reality"

In the presence of relevant physical observations, one can usually calibrate a computer model, and even estimate systematic discrepancies of the model from reality. Estimating and quantifying the uncertainty in this model discrepancy can lead to reliable predictions - so long as the prediction "is similar to" the available physical observations. Exactly how to define "similar" has proven difficult in many applications. Clearly it depends on how well the computational model captures the relevant physics in the system, as well as how portable the model discrepancy is in going from the available physical data to the prediction. This talk will discuss these concepts using computational models ranging from simple to very complex.

**Surya Kalidindi**
Georgia Institute of Technology

"Materials Innovation Driven by Data and Knowledge Systems"

Current approaches to exploring materials and manufacturing (or processing) design spaces in pursuit of new/improved engineered structural materials continue to rely heavily on extensive experimentation, which typically demand inordinate investments in both time and effort. Although tremendous progress has been made in the development and validation of a wide range of simulation toolsets capturing the multiscale phenomena controlling the material properties and performance characteristics of interest to advanced technologies, their systematic insertion into the materials innovation efforts has encountered several hurdles. The most common of these are related to (i) the lack of a generalized (applicable to a wide variety of materials classes and phenomena) mathematical framework that allows objective extraction and synergistic integration of the high value materials knowledge (defined from the perspective of producing reliable process-structure-property (PSP) linkages) from all available datasets (including a variety of multiscale experiments and simulations), while accounting for the inherent uncertainty associated with each dataset, (ii) the lack of formal approaches that identify objectively where to invest the next effort (could be a new experiment or a new simulation) for maximizing the likelihood of success (i.e., meeting or exceeding the designer-specified combinations of materials properties) at any step of the innovation effort, and (iii) the lack of experimental techniques that are specifically designed to provide the quality and quantity of information needed to calibrate the large number of material parameters present in most multiscale materials models. This talk will describe ongoing efforts in my research group aimed at addressing the gaps identified above.

**J. Nathan Kutz**
University of Washington

"Data-Driven Discovery of Governing Physical Laws and their Parametric Dependencies in Engineering, Physics and Biology"

A major challenge in the study of dynamical systems is that of model discovery: turning data into models that are not just predictive, but provide insight into the nature of the underlying dynamical system that generated the data. This problem is made more difficult by the fact that many systems of interest exhibit diverse behaviors across multiple time scales. We introduce a number of data-driven strategies for discovering nonlinear multiscale dynamical systems and their embeddings from data. We consider two canonical cases: (i) systems for which we have full measurements of the governing variables, and (ii) systems for which we have incomplete measurements. For systems with full state measurements, we show that the recent sparse identification of nonlinear dynamical systems (SINDy) method can discover governing equations with relatively little data and introduce a sampling method that allows SINDy to scale efficiently to problems with multiple time scales. Specifically, we can discover distinct governing equations at slow and fast scales. For systems with incomplete observations, we show that the Hankel alternative view of Koopman (HAVOK) method, based on time-delay embedding coordinates, can be used to obtain a linear model and Koopman invariant measurement system that nearly perfectly captures the dynamics of nonlinear quasiperiodic systems. We introduce two strategies for using HAVOK on systems with multiple time scales. Together, our approaches provide a suite of mathematical strategies for reducing the data required to discover and model nonlinear multiscale systems.

**Bani Mallick**
Texas A&M University

"Hierarchical Bayesian Models for Inverse Problems and Uncertainty Quantification"

We consider a Bayesian approach to inverse problems with complex error structure. Hierarchical Bayesian models have been developed in this inverse problem setup. These Bayesian models contain a natural mechanism for regularization in the form of prior distributions. Different regularized prior distributions have been considered to induce sparseness. We propose MCMC as well as variational type algorithms for posterior inference. The proposed methods have been illustrated on several linear and nonlinear inverse problems.

**Akil Narayan**
University of Utah

"Emulators for models and Complexity Reduction"

We present an overview of mathematical tools for building model emulators. We will primarily discuss forward emulation, where one seeks to predict the output of a model given an input. We will emphasize methods that boast stability, accuracy, and computational efficiency, and will in particular discuss emulators built from non-adapted polynomials, and from adapted function spaces.

The talk will highlight some notable advances made in the field of building emulators, and will identify frontiers where mathematical or computational advances are needed.

**J. Tinsley Oden**
University of Texas at Austin

“Principles of Predictive Computational Science: Predictive Models of Random Heterogeneous Materials and Tumor Growth”

This presentation begins with a survey of the principles of predictive computational science: the discipline concerned with assessing the predictability of mathematical and computational models of events that occur in the physical universe in the presence of uncertainties. We then focus on key aspects of predictability: modeling error and how to estimate it, and model selection. The idea of optimal control of modeling error in which a sequence of models is generated so as control error relative to a high-fidelity ground truth model is discussed, as well as the related problem in which, given noisy data , we wish to select, calibrate, and validate a model’s ability to predict quantities of interest with a preset level of accuracy. As applications, we consider the analysis of random heterogeneous media and the construction and selection of mathematical models of tumor growth. We discuss OPAL, the Occam Plausibility Algorithm, as a framework for systematic model selection and validation. Examples of applications of these methodologies are given.

**Bruno Sanso**
University of California, Santa Cruz

“Inferring Release Characteristics from an Atmospheric Dispersion Model using Bayesian Adaptive Splines”

Atmospheric particle dispersion simulators are developed to predict the path of a plume of material released accidentally or intentionally, based on characteristics of the release (location, amount and duration) and meteorological condition. Since release characteristics and meteorological conditions are often unknown, the inverse problem is of great interest that is, based on all the observations of the plume so far, what can be inferred about the release characteristics? This is the question we seek to answer using plume observations from a controlled release at the Diablo Canyon Nuclear Power Plant in Central California. With access to a large number of evaluations of an expensive particle dispersion simulator that includes continuous and categorical inputs and spatio-temporal output, building a fast statistical surrogate model presents many statistical challenges, but is an essential tool for inverse modeling and sensitivity analysis. We achieve accurate emulation using Bayesian adaptive splines to model weights on empirical orthogonal functions. We use this emulator as well as appropriately identifiable simulator discrepancy and observational error models to calibrate the simulator, thus finding a posterior distribution of the release characteristics. The assessment of the calibration is performed using the predictive distributions of the particle concentrations,that blend information from both the observations and simulations, for a comparison with the observed particle counts. In addition, as the release was controlled, its characteristics are known, making it possible to compare our findings to the truth.

**Leonard Smith**
London School of Economics, Pembroke College, Oxford

“On the Impact(s) of Structural Model Error on Simulation Modelling”

Modern, large-scale simulation models are often based on the "laws of physics". Interpreting the outputs of these models nevertheless introduces a host of new challenges for uncertainty quantification and decision making. Unlike the traditional low-dimensional models in statistics or in applied maths, the state space of these simulation models is often large (10^7 D) and the components of the model state vector do not correspond to real-world observables. Nevertheless, in contexts like weather and climate, for example, such models sometimes provide significantly more information than traditional empirical models. Structural Model Error (SME) is the difference between the mathematical structure of the simulation model and the system that generates the observations (assuming that the system has a nontrivial mathematical description). In the absence of SME, reducing imprecision in parameter values and in the current state of the system is a daunting but tractable task, and forecasting deterministic systems takes on a probabilistic This presentation illustrates how SME, and the (almost certain) lack of topological conjugacy it implies, has significant impacts on what can be expected from simulation modelling. Challenges to various approaches of Uncertainty Quantification currently used in practice ("ensemble forecasting") and statistical methods exploiting a discrepancy function are demonstrated.

**Elaine Spiller**
Marquette University

“An Overview of Reduced-Order Models and Emulators”

Typically quantifying uncertainty requires many evaluations of a computational model or simulator. If a simulator is computationally expensive and/or high-dimensional, working directly with a simulator often proves intractable. Surrogates of expensive simulators are popular and powerful tools for overcoming these challenges. I will give an overview of surrogate approaches from an applied math perspective and from a statistics perspective with the goal of setting the stage for the "other" community.

**Robert Wolpert**
Duke University

“UQ Data Fusion: An Introduction and Case Study”

Data Fusion is the process of integrating multiple data sources to produce more consistent, accurate, and useful information than that provided by any individual data source. We show how this may be accomplished in the Bayesian paradigm by constructing non-exchangeable hierarchical models with submodels for each of the several data sources.

In the UQ setting, where we wish to synthesize evidence from large and slow Simulation models and possibly other data sources, it can be much more efficient to construct Gaussian Process Emulators of the Simulation models, and perform Data Fusion in the Emulators rather than the Simulators.

We introduce an abstract model sitting for Fusion, and illustrate several examples from a single case study: the forecasting of hazard from Pyroclastic Density Currents (PDCs) near an active volcano.