SAND2013-1682P

Big Data Meets Big Models
— An Uncertainty Quantification Perspective —

Habib N. Najm

hnnajm@sandia.gov

Sandia National Laboratories,
Livermore, CA

Panel Discussion, SIAM CS&E,
Boston, MA,
Feb 27, 2013

Sandia National Laboratories is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under contract DE-AC04-94-AL85000.

Big Data, Models, & Uncertainty Quantification (UQ)

- Challenges with large data volume are broadly relevant
  - Biology – DNA sequencing
  - Astronomy – Sky surveys
  - Physics – Particle acclerators, detectors
  - Internet – Consumer data, social media
  - Climate – Satellite/monitoring, global data fields
- Challenges with increased model complexity
  - Multiphysics, multiscale, natural & engineered systems
  - Software complexity, verification, and maintenance
  - Computational cost
  - Analysis and visualization of computational data
- Data and models are of central relevance in UQ
  - UQ entails significant growth in the degree of information represented, in comparison to deterministic problems
  - UQ relies on data for parameter estimation; model calibration and validation

Inverse UQ: data, inference and model calibration

- Representation and estimation of random fields with large range of scales – *e.g.* Geophysics data
  - Heterogeneous material properties
  - Large range of correlation length scales
  - Large random field representation data volume
  - High-dimensional statistical inference
- Reconciling multiple data sources – disagreements
  - Different instruments, exptl. techniques, models
- Identifiability
  - Sensitivity of observables to any given parameter
  - Some parameters may not be discoverable from available observables
  - Data not equally informative on all parameters – no change from prior to posterior

Costs of Data Collection and Use

- Context: Estimate a (posterior) density on parameters by fitting model output observables to a data set
- Focusing data campaigns on specific regions/conditions can can lead to improved confidence in estimated parameters
  - Some of the data can be uninformative on the parameters
  - Use available budget to collect more data where it matters
  - Experimental design
- Using all available data for inference can require significant computational effort for likelihood evaluation
  - Not justifiable when large subsets of the data are uninformative

Forward UQ and Data Volume

- UQ increases the size of computational data bases
- Forward UQ context:
  - Large number of model runs
  - Representations of output random fields
- Need for enhanced frameworks for managing large volumes of computational UQ data
- Even embarrasingly parallel computations hit I/O and memory bandwidth bottlenecks
  - Multiple model realizations running in parallel
  - Similar I/O and memory access patterns/volume
  - Memory and I/O bus contention

Validation of Big Models – Predictive Skill

- Complex models have many observables
- Danger of overfitting when model complexity exceeds the degree of information in the data
- Computational challenges in model selection
  - estimation of marginal likelihood – integration
  - cross validation as a measure of robustness
- Challenge in proper assessment of predictive skill
  - Even when climate models are calibrated with available historical data, they still have poor predictive skill
  - They do not model the climate well enough to be “valid"

Keeping in mind:

All models are wrong, but some are useful.

George E.P. Box

Closure

- Large experimental data volumes, model complexity, and large computational data volumes can all lead to challenges for UQ
- Random field quantities raise significant challenges
- Experimental design, informative data
- Forward UQ samples
- Model complexity, overfitting
- Validation & predictive skill