Bayesian inference for model error quantification and propagation with UQTk

Khachik Sargsyan, Xun Huan, Habib Najm,
Cosmin Safta, Kenny Chowdhary, Bert Debusschere

Sandia National Laboratories, Livermore, CA

SIAM Annual Meeting
Portland, OR
July 9-13, 2018

Bayesian inference for model error quantification and propagation with UQTk

Khachik Sargsyan, Xun Huan, Habib Najm,
Cosmin Safta, Kenny Chowdhary, Bert Debusschere

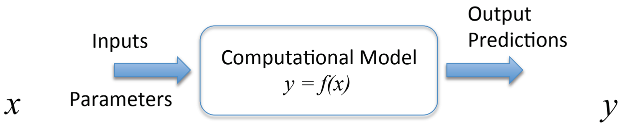
Sandia National Laboratories, Livermore, CA

SIAM Annual Meeting
Portland, OR
July 9-13, 2018

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell under International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration contract DE-NA-0003525.

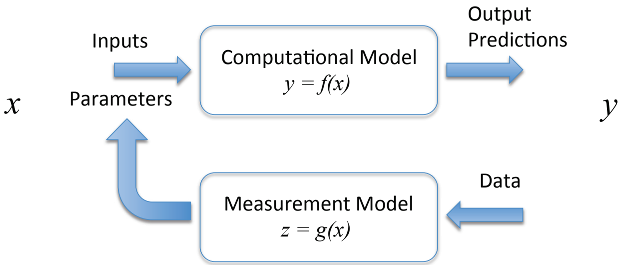
Outline

Uncertainty Quantification and Computational Science



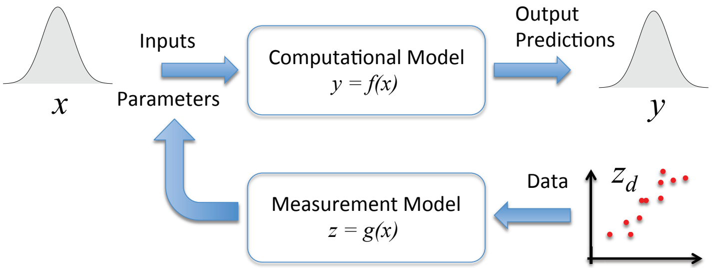
Forward problem

Uncertainty Quantification and Computational Science



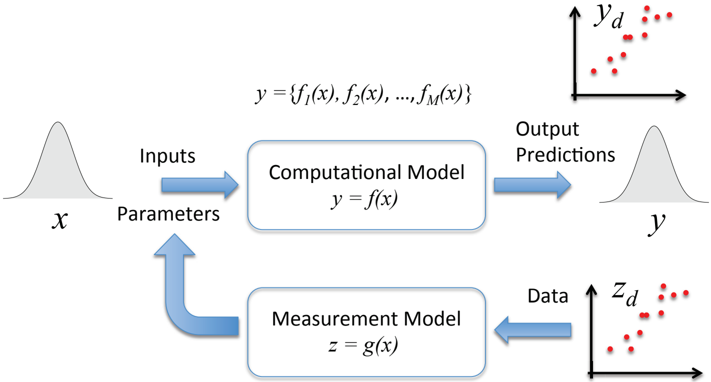
Inverse & Forward problems

Uncertainty Quantification and Computational Science



Inverse & Forward UQ

Uncertainty Quantification and Computational Science



Inverse & Forward UQ

Model validation & comparison, Hypothesis testing

The Case for Uncertainty Quantification

UQ needed for...

- Model predictions
- Model validation and comparison
- Confidence assessment
- Reliability analysis
- Dimensionality reduction
- Optimal design
- Decision support
- (Noisy) data assimilation

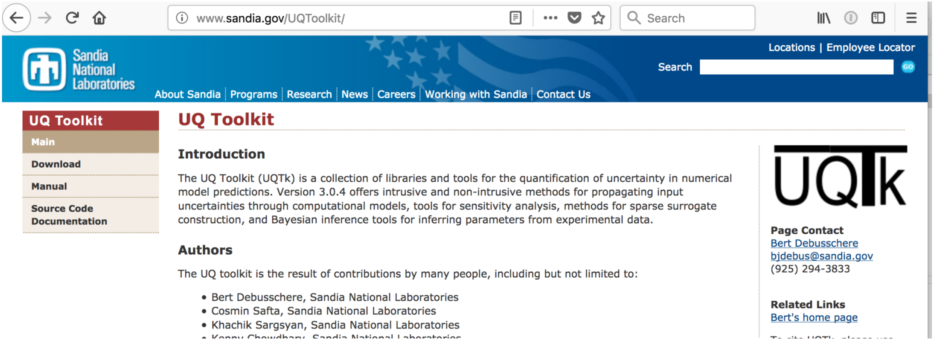
Uncertainty Sources

- Model parameters
- Initial/boundary conditions
- Model geometry/structure
- Lack of knowledge
- Data noise
- Intrinsic stochasticity
- Numerical errors, too

UQ Enables Predictive Simulations

- Locate all sources of (manageable) uncertainties
- Parameter selection/estimation
  - Auxilliary data collection
  - Expert opinion, physical bounds, maximum entropy
  - Submodel fitting/regression
- Forward propagation of uncertainties
  - Local SA (deterministic, error propagation)
  - Interval math, evidence theory
  - Global SA (stochastic, variance-based decomposition)
- Calibration with a separate set of data (inverse UQ)
- Model (in)validation
  - No model is perfect
  - Compare model prediction with uncertainties versus data on some QoI
  - Model comparison (Bayes Factors, Model Plausibility)

UQTk provides tools to build general UQ workflows



- Representation of random variables and stochastic processes
- Forward uncertainty propagation: Polynomial Chaos
- Inverse problems: Bayesian methods, MCMC
- Quadrature integration
- Sensitivity analysis: Sobol indices
- Dimensionality reduction: Karhunen-Loève expansions
- Gaussian Processes
- Sparse approximations: Bayesian Compressive Sensing
- Low Rank Tensors

UQTk is meant to be straightforward to download, install and use

UQTk

- Target usage:
  - Rapid prototyping of UQ workflows
  - Algorithmic research in UQ
  - Tutorials / educational
  - Expertise in UQ methods (or a desire to acquire it) helpful
- Released under the GNU Lesser General Public License
  - `http://www.sandia.gov/UQToolkit/`
  - Current version 3.0.4
  - Version 3.1.0 planned for later this year
- Download as tar file and configure with `CMake`
- No massive third party libraries to download, install, and configure

UQTk is used in a variety of applications

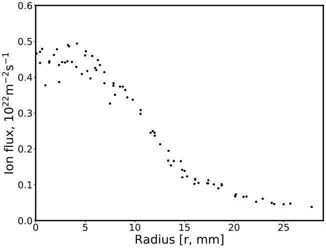
UQTk

- Direct collaborations
  - US DOE SciDAC FASTMath institute
    https://fastmath-scidac.llnl.gov/
  - Variety of US DOE SciDAC partnership projects
  - Part of US DOE BER E3SM climate model analysis tools
- Many other research groups at universities, National Labs, and industry
- Always welcome new applications / collaborations
- Mailing lists
  - uqtk-announce@software.sandia.gov
  - uqtk-users@software.sandia.gov
  - Join at http://www.sandia.gov/UQToolkit

Outline

Fitting parametric model to data: least squares

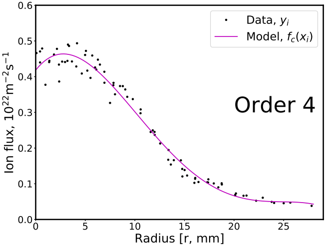
- Given data (x_i, y_i) for $i = 1, \ldots, N$



Fitting parametric model to data: least squares

- Given data (x_i, y_i) for $i = 1, \dots, N$
- Given parametrized model form $f_c(x)$
- Tune c, such that $y_i \approx f_c(x_i)$
- Least-squares

$$\underset{c}{\operatorname{argmin}} \sum_{i=1}^{N} (y_i - f_c(x_i))^2$$

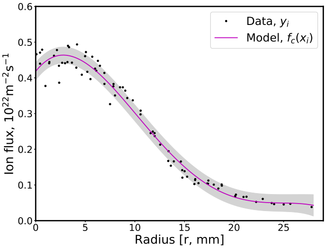


- Linear parametrization (basis expansion)... $$f_c(x) = \sum_{k=0}^{K} c_k \Psi_k(x)$$
- ... allows analytical answer $c = (P^T P)^{-1} P^T y$, where $P_{ik} = \Psi_k(x_i)$

Fitting parametric model to data: least squares

- Given data (x_i, y_i) for $i = 1, \dots, N$
- Given parametrized model form $f_c(x)$
- Tune c, such that $y_i \approx f_c(x_i)$
- Least-squares

$$\underset{c}{\operatorname{argmin}} \sum_{i=1}^{N} (y_i - f_c(x_i))^2$$



- Linear parametrization (basis expansion)... $$f_c(x) = \sum_{k=0}^{K} c_k \Psi_k(x)$$
- ... allows analytical answer $c = (P^T P)^{-1} P^T y$, where $P_{ik} = \Psi_k(x_i)$
- ... with covariance information $\Sigma_c \propto (P^T P)^{-1}$

Bayesian viewpoint of fitting $y_i \approx f_c(x_i)$

- Bayes' formula
$$p(\mathcal{M}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathcal{M})p(\mathcal{M})}{p(\mathcal{D})}$$

- Data $\mathcal{D} = \{x_i, y_i\}_{i=1}^N$
- Model $\mathcal{M} \equiv c$
- Rewrite Bayes' formula

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

- Prior $p(c)$: expert knowledge, or uninformative
- Posterior $p(c|y)$: updated 'knowledge' of c, given data y
- Likelihood $L(c) = p(y|c)$: key, noise/error model, encapsulates assumptions about data collection
- Evidence $p(y)$: not important for parameter (coeff. c) estimation; crucial for model selection (e.g. poly order)

Posterior sampling via Markov chain Monte Carlo (MCMC)

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

- In general, when model is not linear or noise is not Gaussian, there is little alternative to MCMC
- MCMC is a search procedure in parameter space leading to a stochastic process with a stationary distribution $p(c|y)$
- Given samples from posterior, one can interrogate it further
  - Estimate PDF with KDE
  - Compute moments
  - Build functional representation, such as PC
  - Pipe it to the next model as an input

Polynomial Chaos – functional representation for RVs

$$U \simeq \sum_{k=0}^{p} u_k \psi_k(\xi)$$

- First introduced by Wiener, 1938
- Revitalized by Ghanem and Spanos, 1991
- Convergent series if U has finite variance
- Selection of order p is a modeling choice
- Describes a r.v. U with a vector of *PC modes* $(u_0, u_1, \dots, u_p)$

- Standard r.v. ξ, standard orthogonal polynomials $\psi_k(\xi)$, *i.e.*

$$\int \psi_i(\xi)\psi_j(\xi)\pi_\xi(\xi)d\xi = \delta_{ij}||\psi_i||^2$$

PC Type	Domain	Density $\pi_\xi(\xi)$	Polynomial	Free parameters
Gauss-Hermite	$(-\infty, +\infty)$	$\frac{1}{\sqrt{2\pi}} e^{-\frac{\xi^2}{2}}$	Hermite	none
Legendre-Uniform	$[-1, 1]$	$\frac{1}{2}$	Legendre	none
Gamma-Laguerre	$[0, +\infty)$	$\frac{\xi^\alpha e^{-\xi}}{\Gamma(\alpha+1)}$	Laguerre	$\alpha > -1$
Beta-Jacobi	$[-1, 1]$	$\frac{(1+\xi)^\alpha (1-\xi)^\beta}{2^{\alpha+\beta+1} B(\alpha+1,\beta+1)}$	Jacobi	$\alpha > -1, \beta > -1$

[Wiener, 1938; Ghanem & Spanos, 1991; Xiu & Karniadakis, 2002; Le Maître & Knio, 2010]

Essential use of PC in UQ $U \simeq \sum_{k=0}^{K} u_k \Psi_k(\boldsymbol{\xi})$

Strategy:

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

Advantages:

- Computational efficiency
- Utility
  - Moments: $\mathbb{E}[u] = u_0, \ \mathbb{V}[u] = \sum_{k=1}^{K} u_k^2 ||\Psi_k||^2, \ \ldots$
  - Global Sensitivities – fractional variances, Sobol' indices
  - Uncertainty propagation
  - Surrogate for forward model

Requirements:

- Finite variances (not a handicap in practice)
- Smooth forward functions

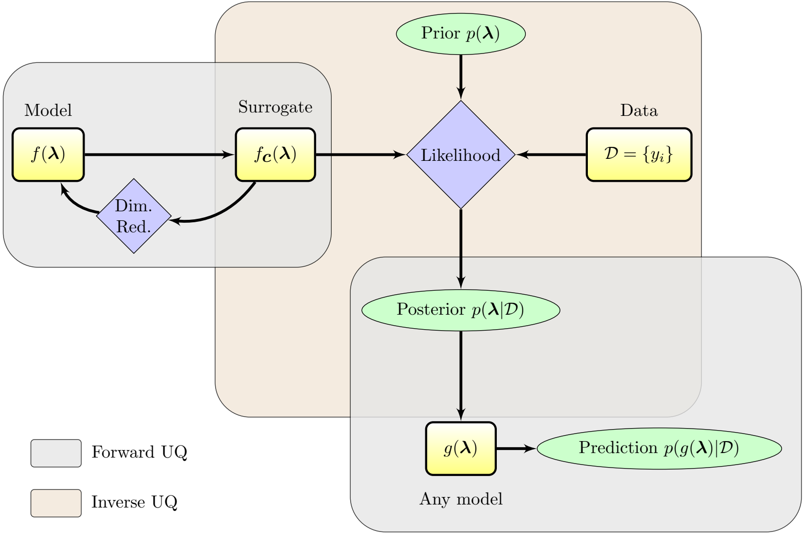
PC features: uncertainty propagation

$$U \simeq \sum_{k=0}^{K} u_k \Psi_k(\boldsymbol{\xi})$$

$$Z = f(U) \simeq \sum_{k=0}^{K} c_k \Psi_k(\boldsymbol{\xi})$$

- Basic task: given PC for inputs, find PC for outputs.
- Input-output map can also be defined implicitly, via governing equations $G(Z, U) = 0$.
- Two approaches
  - Intrusive: project governing equations
    - Results in set of equations for the PC modes
    - Requires redesign of computer code
    - PCEs for all uncertain variables in system
  - Non-intrusive: project outputs of interest
    - Sampling to evaluate projection operator
    - Can use existing code as black box
    - Only computes PCEs for quantities of interest

An example UQTk workflow (UQTk version 3.0)

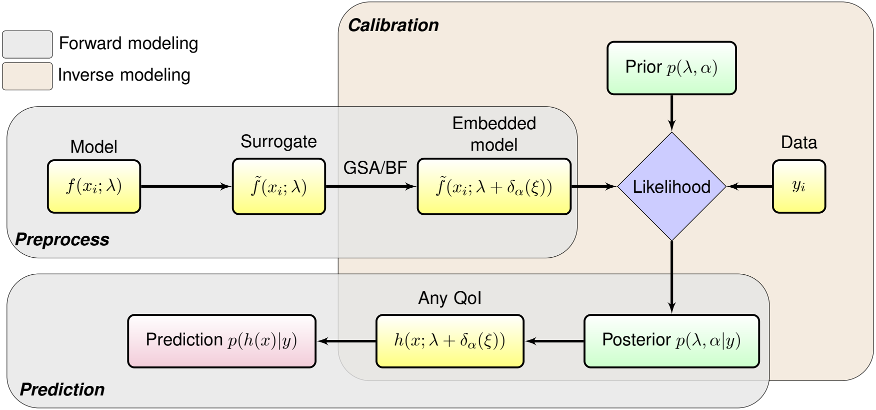


Model Error: Quantification of Structural Error

Model error = deviation from ‘truth’, or from a higher-fidelity model

- Represent and estimate the error associated with
  - Simplifying assumptions, parameterizations
  - Mathematical formulation, theoretical framework
  - Numerical discretization
- ...will be useful for
  - Model validation
  - Model comparison
  - Scientific discovery and model improvement
  - Reliable computational predictions
- Inverse modeling context
  - Given experimental or higher-fidelity model data, estimate the model error

Model error embedding – workflow



- Predictive uncertainty decomposition: Total Variance =

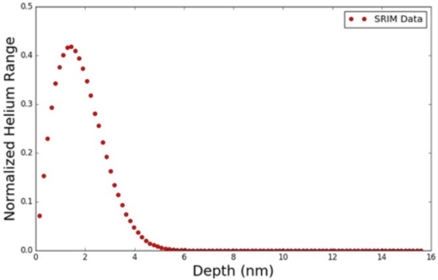
  Posterior uncertainty + Data noise + Model error + Surrogate error

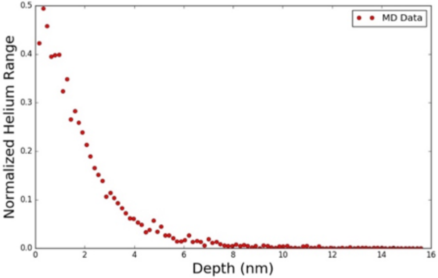
Outline

UQ in PSI – Xolotl: overview

Normalized helium range as a function of tungsten depth obtained by

- Stopping and Range of Ions in Matter (SRIM), or
- Molecular Dynamics (MD).



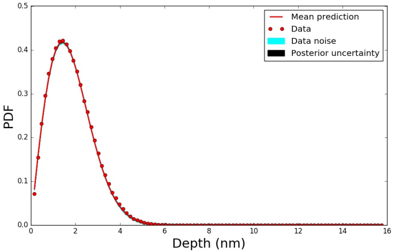


Main targets:

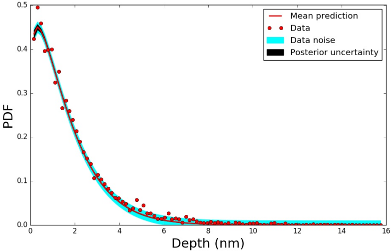
- fit a probability density function (PDF) to this data
- incorporate all uncertainty sources
- propagate through Xolotl (Plasma Surface Interactions Simulator)

UQ in PSI – Conventional calibration: no model error

SRIM only

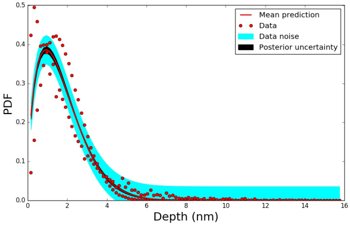


MD only



- Fit a Weibull PDF
  $$f(x;\lambda_1,\lambda_2) = \frac{\lambda_2}{\lambda_1}\left(\frac{x}{\lambda_1}\right)^{\lambda_2-1} e^{-\left(\frac{x}{\lambda_1}\right)^{\lambda_2}}$$
- Given data $y(x_i)$, apply Bayes' theorem
  $$p(\lambda|y) \propto p(y|\lambda)p(\lambda)$$
- Classical Bayesian Likelihood (generalized least-squares)
  $$\ln p(y|\lambda) = \text{const} - \sum_i \frac{(y(x_i) - f(x_i;\lambda))^2}{2\sigma^2}$$

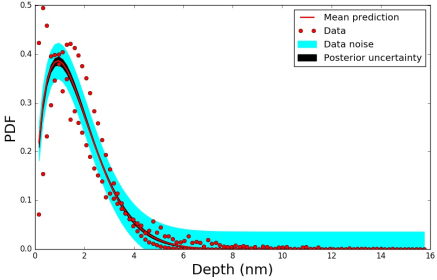
SRIM and MD



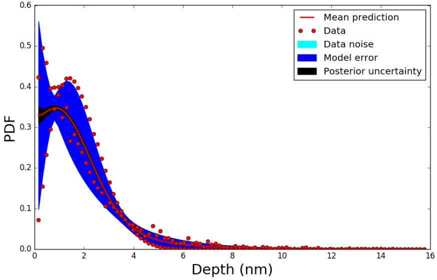
Embedded model error: captures modeling discrepancies

- Allow variability in λ: cast it as a random variable $\lambda(\alpha)$
- Deterministic model $f(x;\lambda) \Rightarrow$ Stochastic model $f(x;\lambda)$ with mean $\mu(x;\alpha)$ and variance $\sigma^2(x;\alpha)$
- Infer α instead of λ
- Likelihood function $\ln p(y|\alpha) = \mathrm{const} - \sum_i \frac{(y(x_i)-\mu(x_i;\alpha))^2}{2\sigma^2(x_i;\alpha)}$

No model error



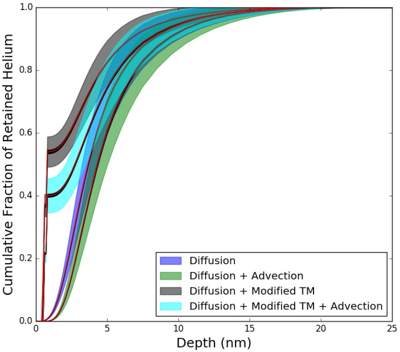
Model error



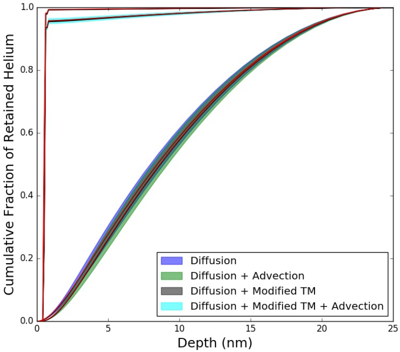
Propagate all sources of uncertainty through Xolotl

- Propagate further in the pipeline with uncertainties
- Two values of nominal flux F_N

$F_N = 4 \times 10^{25}$ He m^{-2}s^{-1}



$F_N = 4 \times 10^{22}$ He m^{-2}s^{-1}

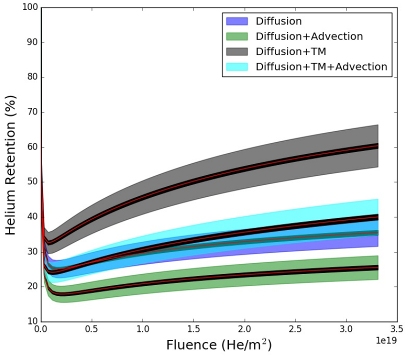


Cumulative Fraction of Helium Retention vs. Depth

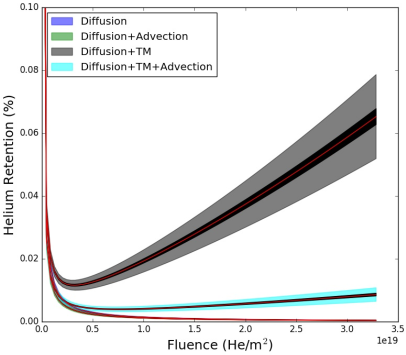
Propagate all sources of uncertainty through Xolotl

- Propagate further in the pipeline with uncertainties
- Two values of nominal flux F_N

$F_N = 4 \times 10^{25}$ He m^{-2}s^{-1}



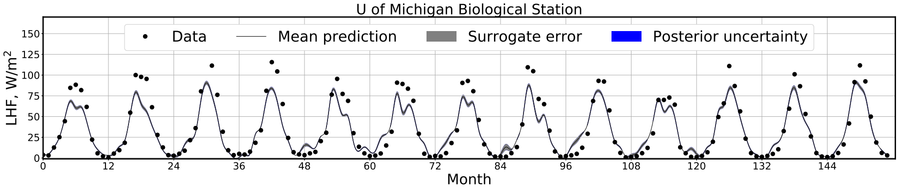
$F_N = 4 \times 10^{22}$ He m^{-2}s^{-1}



Helium Retention vs. Fluence

E3SM Land Model (ELM)

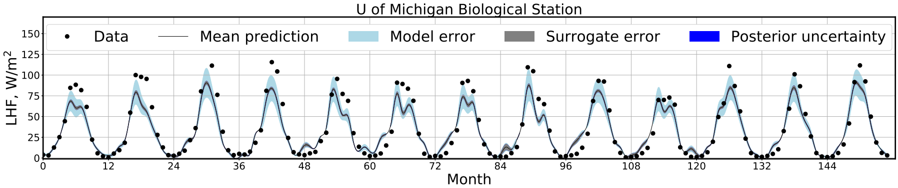
- US Department of Energy (DOE) sponsored Earth system model
- Land, atmosphere, ocean, ice, human system components
- High-resolution, employ DOE leadership-class computing facilities



- Conventional calibration without model error

E3SM Land Model (ELM)

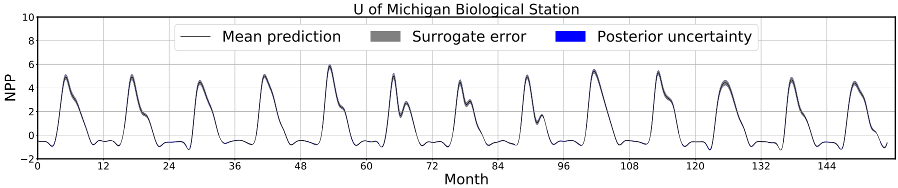
- US Department of Energy (DOE) sponsored Earth system model
- Land, atmosphere, ocean, ice, human system components
- High-resolution, employ DOE leadership-class computing facilities



- Predictive variance decomposition with model-error component
- ... with predictive uncertainty that captures model error

E3SM Land Model (ELM)

- US Department of Energy (DOE) sponsored Earth system model
- Land, atmosphere, ocean, ice, human system components
- High-resolution, employ DOE leadership-class computing facilities

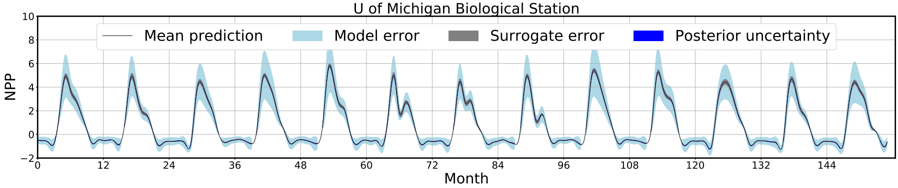


- Predictive variance decomposition with model-error component
- Allows meaningful prediction of other QoIs
  (e.g. no data/observable)

E3SM Land Model (ELM)

E3SM Energy Exascale Earth System Model

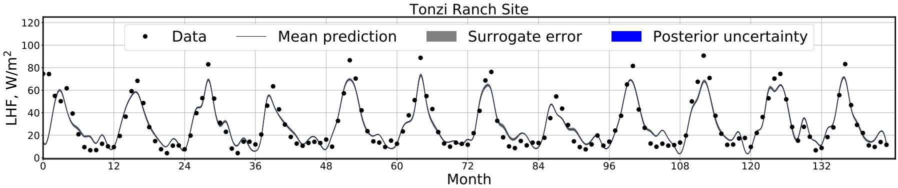
- US Department of Energy (DOE) sponsored Earth system model
- Land, atmosphere, ocean, ice, human system components
- High-resolution, employ DOE leadership-class computing facilities



- Predictive variance decomposition with model-error component
- Allows meaningful prediction of other QoIs (e.g. no data/observable)
- ... with predictive uncertainty that captures model error

E3SM Land Model (ELM)

- US Department of Energy (DOE) sponsored Earth system model
- Land, atmosphere, ocean, ice, human system components
- High-resolution, employ DOE leadership-class computing facilities

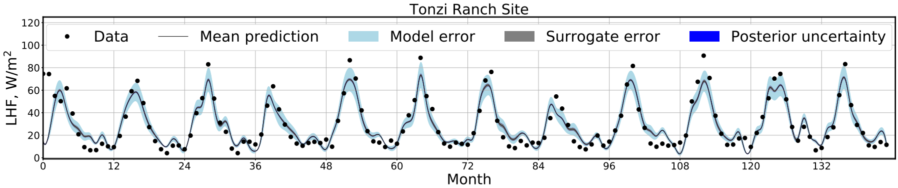


- Predictive variance decomposition with model-error component
- Allows (a more dangerous) extrapolation to other sites

E3SM Land Model (ELM)

E3SM Energy Exascale Earth System Model

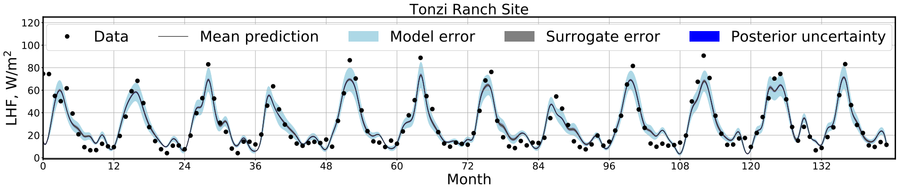
- US Department of Energy (DOE) sponsored Earth system model
- Land, atmosphere, ocean, ice, human system components
- High-resolution, employ DOE leadership-class computing facilities



- Predictive variance decomposition with model-error component
- Allows (a more dangerous) extrapolation to other sites
- ... with predictive uncertainty that captures model error

E3SM Land Model (ELM)

- US Department of Energy (DOE) sponsored Earth system model
- Land, atmosphere, ocean, ice, human system components
- High-resolution, employ DOE leadership-class computing facilities



Outline

1. Uncertainty Quantification
   - Intro
   - UQ Toolkit
2. Workflows
   - Inverse UQ: Bayesian Inference, Model Calibration
   - Forward UQ: Surrogate Modeling, Uncertainty Propagation
   - Model Structural Error
3. Applications in SciDAC
   - Plasma Surface Interactions
   - Land Model
4. Summary

Wrap-up

- UQTk provides a powerful set of tools for building general UQ workflows
- Multiple ways to access functionality
  - Direct linking of C++ code
  - Standalone apps
  - Python interface based on Swig
- Available at `http://www.sandia.gov/UQToolkit`
- Suggestions for improvements welcome!
- Do not hesitate to contact us
  `uqtk-users@software.sandia.gov`

Literature

Thank you!

General UQ

R. Ghanem, P. Spanos, “Stochastic Finite Elements: A Spectral Approach”, Springer Verlag, 1991.

D. Xiu, G. Karniadakis, “The Wiener-Askey Polynomial Chaos for Stocahstic Differential Equations”, *SIAM J. Sci. Comp.*, 24(2), 619-644, 2002.

O. Le Maître, O. Knio, “Spectral Methods for Uncertainty Quantification: With Applications to Computational Fluid Dynamics”, Springer-Verlag, 2010.

H. Najm, “Uncertainty Quantification and Polynomial Chaos Techniques in Computational Fluid Dynamics”, *Ann. Rev. Fluid Mech.*, 41(1):35-52, 2009.

D. Xiu, “Numerical Methods for Stochastic Computations: A Spectral Method Approach”, Princeton U. Press, 2010.

Marzouk, Y., Najm, H., “Dimensionality Reduction and Polynomial Chaos Acceleration of Bayesian Inference in Inverse Problems”, *J. Comp. Phys.*, 228(6):1862-1902, (2009).

B. Debusschere, H. Najm, P. Pébay, O. Knio, R. Ghanem and O. Le Maître, “Numerical Challenges in the Use of Polynomial Chaos Representations for Stochastic Processes”, *SIAM J. Sci. Comp.*, 26:2, 2004.

UQTk chapter in “Handbook of Uncertainty Quantification”, R. Ghanem, D. Higdon, H. Owhadi (Eds.), Springer, 2016, `http://www.springer.com/us/book/9783319123844`

UQ Tutorials: `http://www.quest-scidac.org/outreach/tutorials/`

Model error

M. Kennedy and A. O’Hagan, “Bayesian calibration of computer models”, *Journal of the Royal Statistical Society*, Series B. 63, 425-464, (2001).

K. Sargsyan, H. Najm, R. Ghanem, “On the Statistical Calibration of Physical Models”, *Int. J. Chem. Kinetics*, 47(4), 246-276, (2015).

K. Sargsyan, X. Huan, H. Najm. “Embedded Model Error Representation for Bayesian Model Calibration”. ArXiv version, arXiv:1801.06768, (2018).