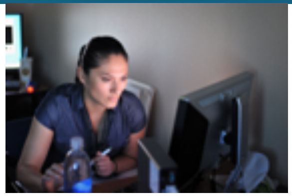
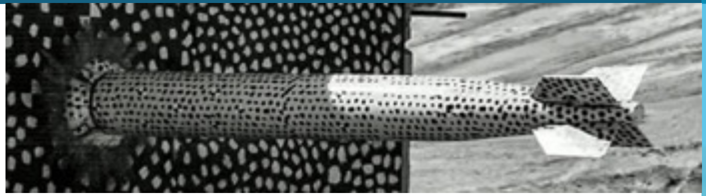
Statistical Learning and Model Error Estimation





PRESENTED BY

Khachik Sargsyan, 8351

April 13, 2021

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Outline

- UQ and statistical learning
- Model structural error
- Applications
- Summary/future

- Chemistry (BES)
- Fusion science (FES+ASCR)
- Turbulence modeling (DARPA)
- Climate land model (BER+ASCR)
- Thermodynamics (EERE)
- ….

Thanks to

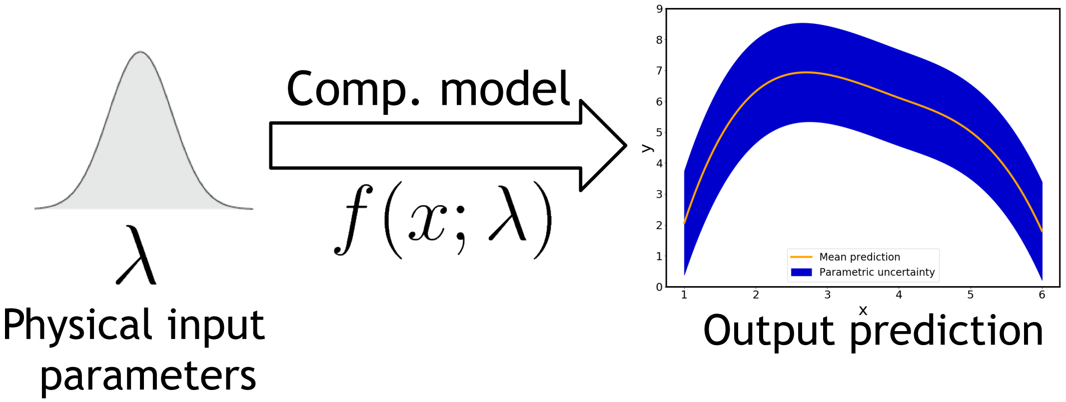
Habib Najm, Cosmin Safta, Tiernan Casey, James Oreluk, Bert Debusschere (SNL)

Daniel Ricciuto (ORNL), Jason Bender (LLNL)

Youssef Marzouk, Chi Feng (MIT), Roger Ghanem (USC), Xun Huan (UMichigan)

Forward Uncertainty Quantification (UQ):

not the focus in this talk



<u>*Forward predictions:*</u>
surrogate models,
sensitivity analysis,
parametric uncertainty

Combination of
supervised and unsupervised ML methods
to tackle
non-linearity, curse of dimensionality
and computational expense

Prediction variance = parametric uncertainty

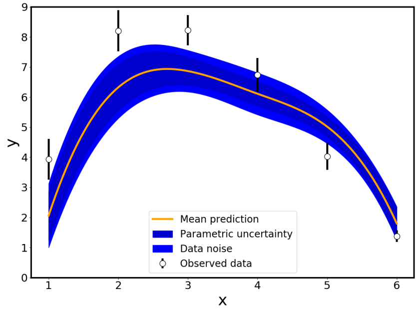
Inverse Uncertainty Quantification:
Statistical Learning from Data

λ

Physical input parameters

Comp. model

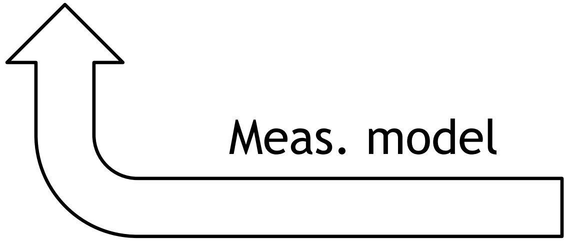
$f(x;\lambda)$



Output prediction

Forward predictions:

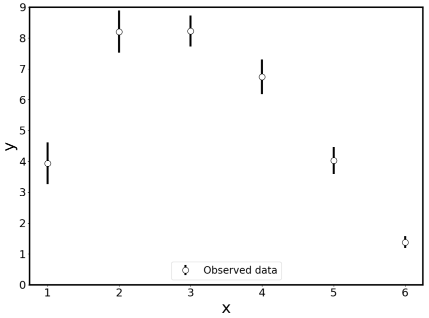
surrogate models,
sensitivity analysis,
parametric uncertainty



Meas. model

$y(x)$

Observed data



Inverse modeling:

parameter tuning,
calibration,
data noise

Prediction variance = parametric uncertainty + data noise

Bayesian inference for statistical learning of model parameters

- Collected data $\{(x_i, y_i)\}_{i=1}^{N}$
- Data model $y_i = f(x_i; \lambda) + \epsilon_i$

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

- Bayes formula

- Prior : knowledge of λ before seeing data (expert opinion, previous analysis, etc…)
- Likelihood : forward model and measurement noise
- Posterior : updated knowledge of λ, combining the prior and the likelihood
- Evidence : normalizing constant, useful for model selection, not for parameter estimation

Markov chain Monte Carlo is used to sample from posterior

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

Markov chain Monte Carlo (MCMC) samples from posterior by marching in the λ-space.

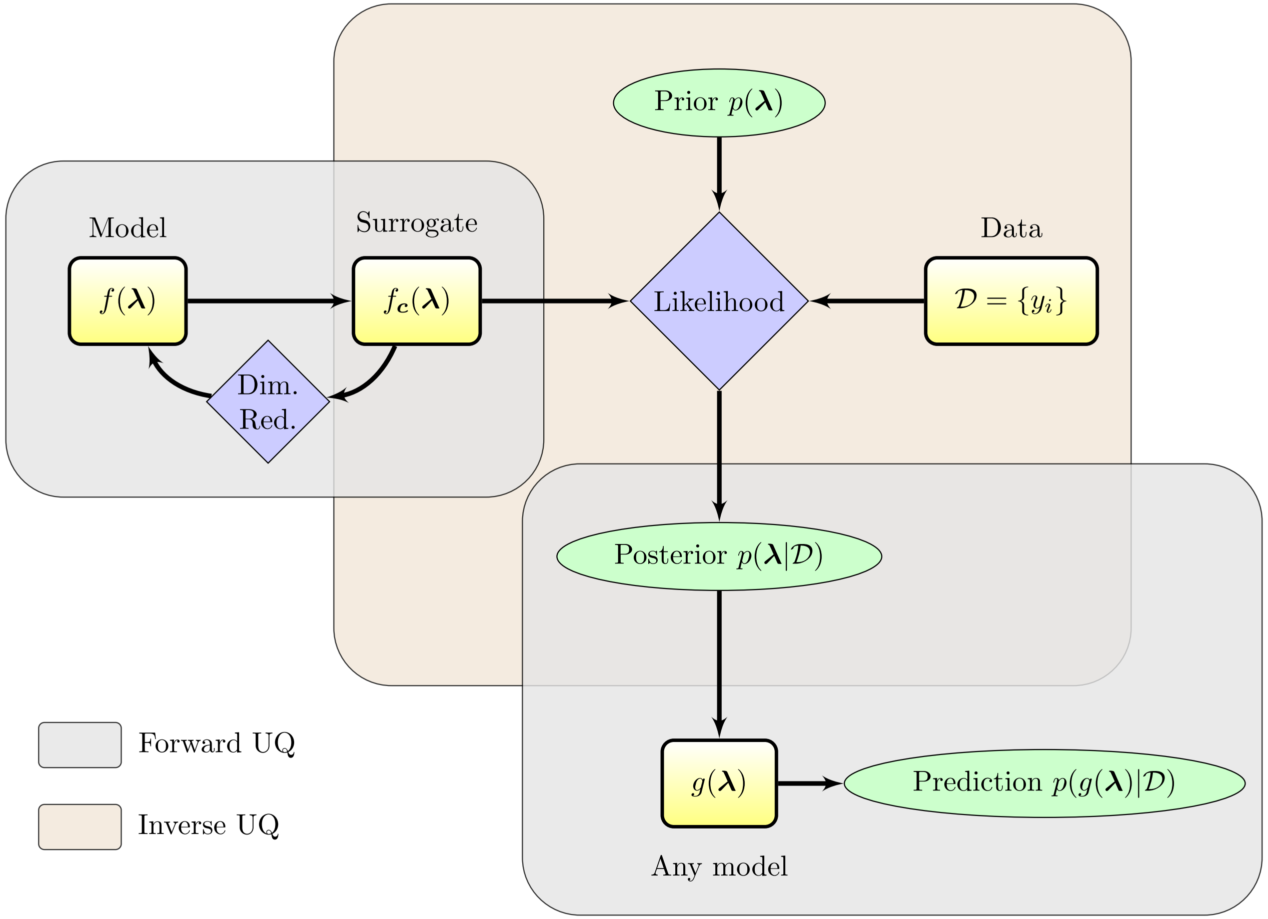
Likelihood is key:

- It incorporates statistical assumptions about the discrepancy between model and data.
- It requires model evaluation at a proposed parameter value λ.

… but it is often infeasible to use model online in an MCMC loop,

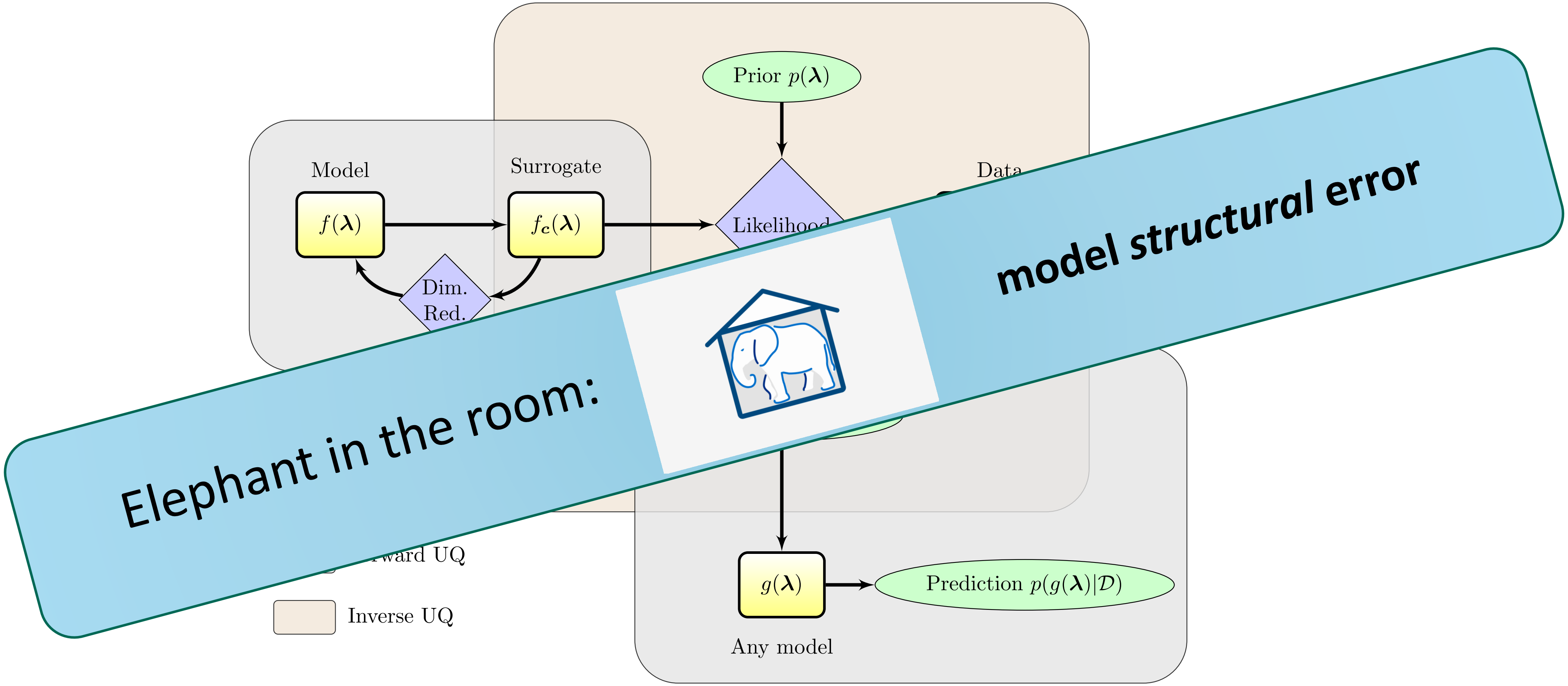
hence we pre-construct a model surrogate.

Surrogate-enabled Bayesian inference



Prediction variance = parametric uncertainty + data noise + surrogate error

Surrogate-enabled Bayesian inference



Prediction variance = parametric uncertainty + data noise + surrogate error

Model error can be defined in a variety of ways

In this work, model error is the difference between our model and the 'truth' model behind noisy data

$$\underbrace{f(x_i;\lambda)}_{\text{model}} \quad \text{vs} \quad \underbrace{y(x_i)}_{\text{data}} = \underbrace{g(x_i)}_{\text{truth}} + \underbrace{\epsilon_i}_{\text{noise}}$$

Model error is associated with

- Simplifying assumptions, parameterizations
- Mathematical formulation, theoretical framework

Very loaded concept
... otherwise called
(with altered meanings)

- model discrepancy
- model structural error
- model inadequacy
- model misspecification
- model form uncertainty
- model uncertainty

Model exploration via embedded statistical representation of model error

Non-intrusive

$$y(x_i) = f(x_i; \lambda + \delta(x_i)) + \epsilon_i$$

- Allows meaningful extrapolation
- Respects physics
- Disambiguates model and data errors
- Predictive uncertainty attribution
  - surrogate errors
  - data noise
  - parametric uncertainty
  - structural errors

Intrusive

$$y(x_i) = \tilde{f}(x_i; \lambda, \delta(x_i)) + \epsilon_i$$

- For best impact, always look under the hood

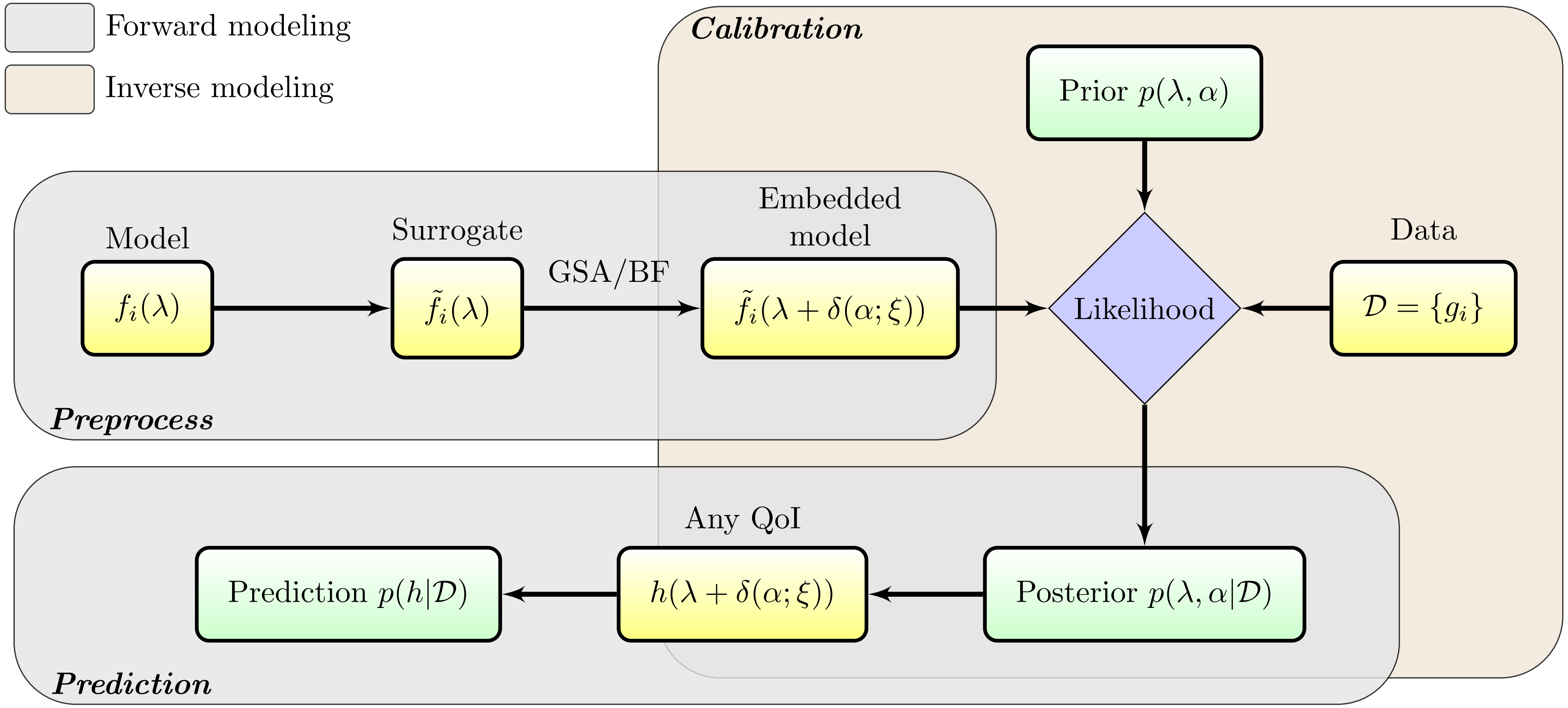
UQTk

- Code available via UQTk
  - (www.sandia.gov/uqtoolkit)
- Impacted many programs DOE/DOD/SNL
- Applied outside immediate group
- Provides alternative for the conventional external correction approaches

Method: Sargsyan, Najm, Ghanem, IJCK (2015); Sargsyan, Huan, Najm, IJUQ (2019).

Applications: Huan et. al, AIAA J (2018); Hakim et. al, CTM (2018); Cekmer et. al, IJUQ (2018); Rizzi et. al, CMAME (2019).

Embedded statistical representation of model error



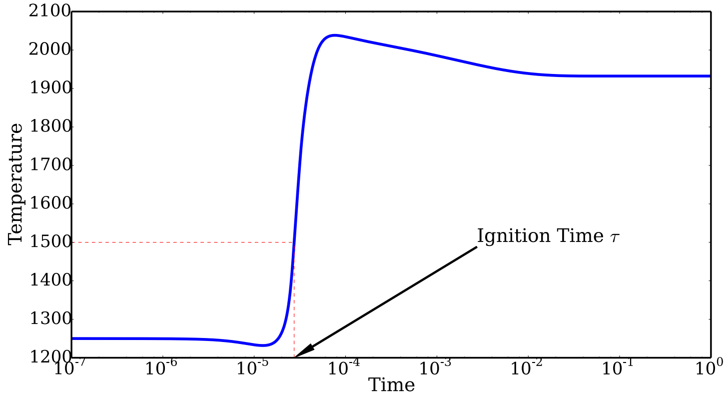
Prediction variance = parametric uncertainty + data noise + surrogate error + model error

Application: Chemistry

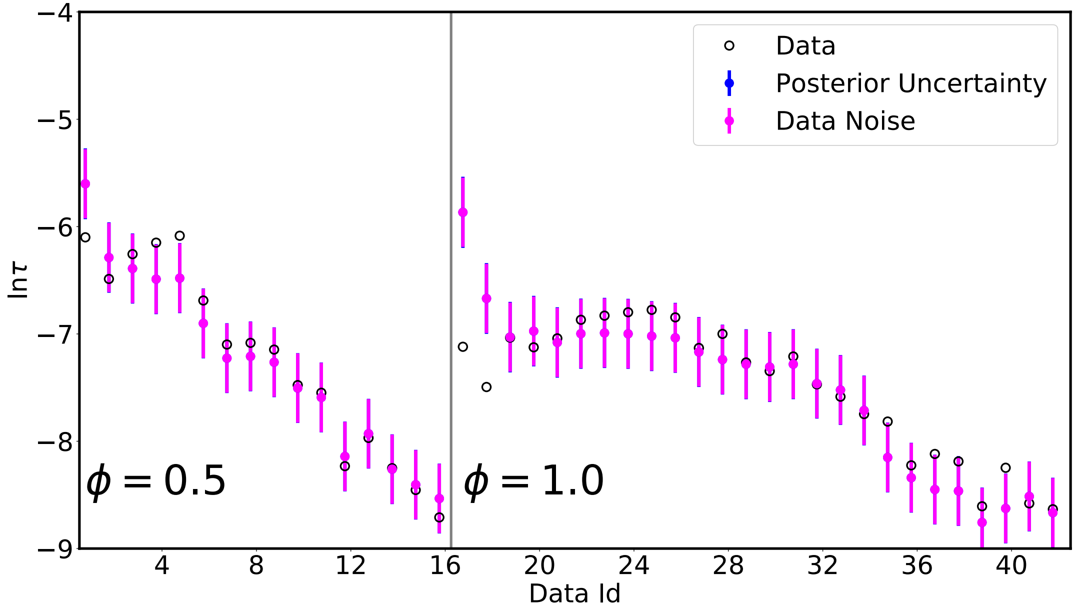
funded by **BES**

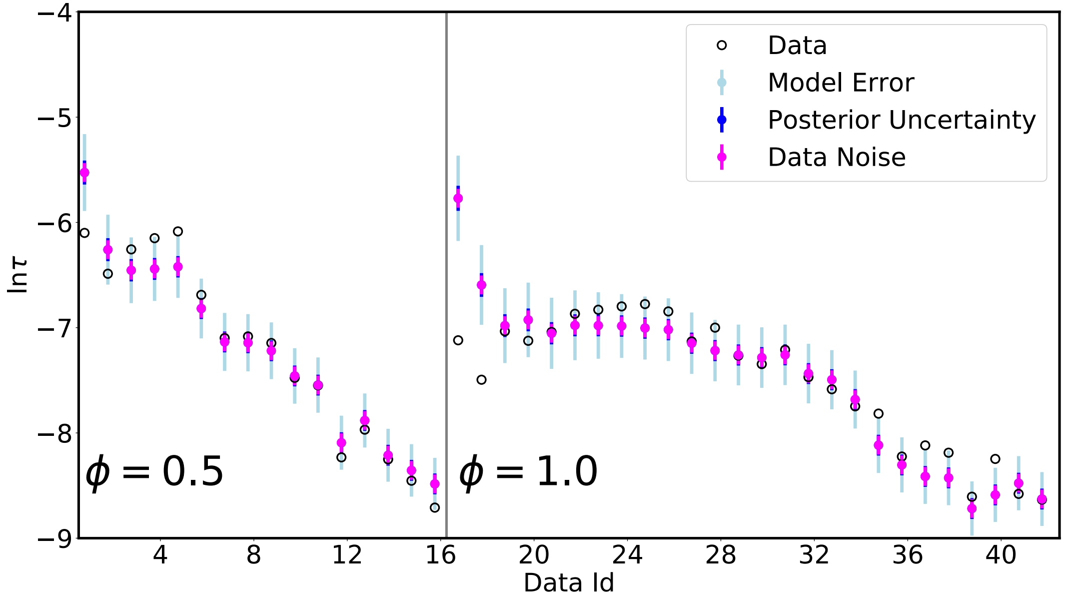
- A sandbox for developing the method
- Calibrating a simple 2-step reaction mechanism given high-fi model or experimental data

QoI: log-ignition time



Without model error, all the discrepancy is attributed to data noise





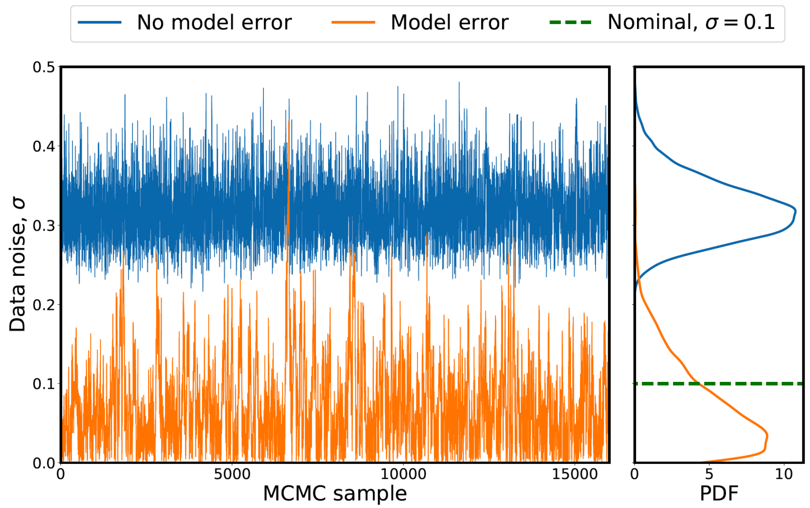
Sargsyan, Najm, Ghanem, IJCK (2015); Sargsyan, Huan, Najm, IJUQ (2019); Hakim et. al, CTM (2018)

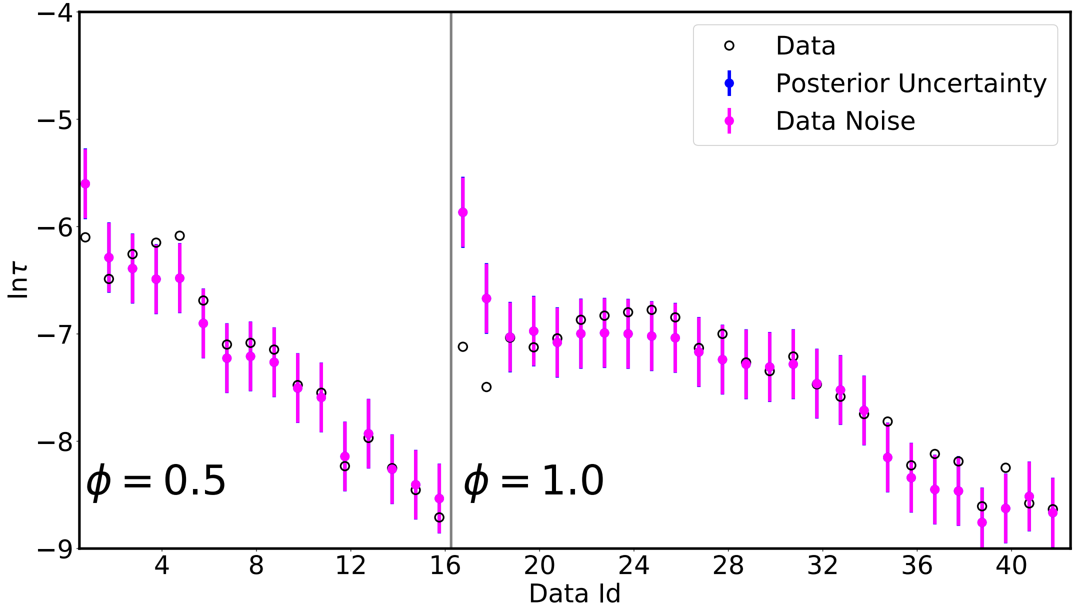
Application: Chemistry

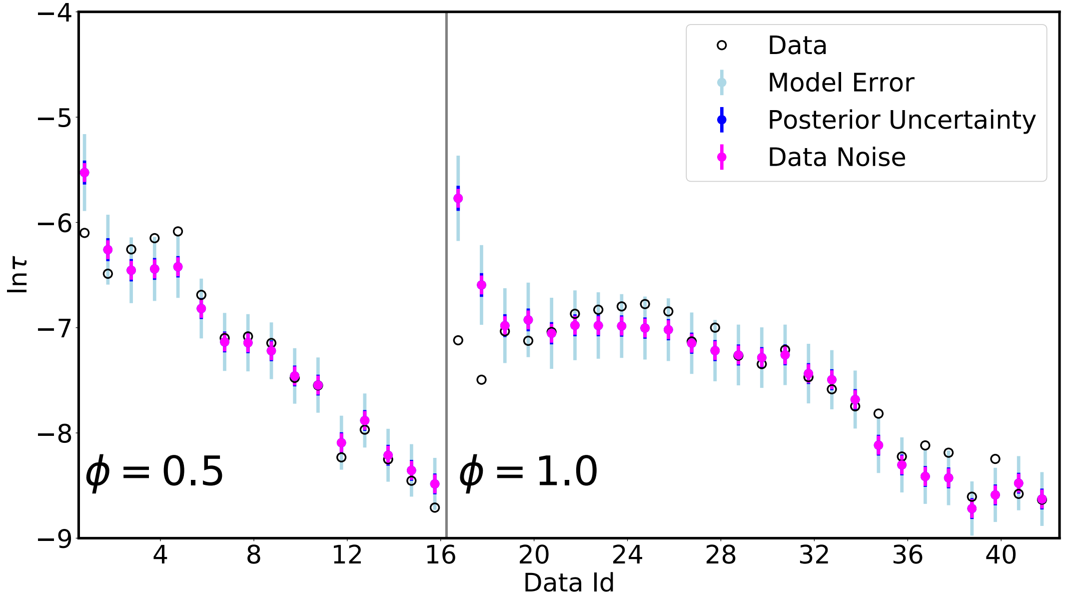
funded by **BES**

- A sandbox for developing the method
- Calibrating a simple 2-step reaction mechanism given high-fi model or experimental data

Without model error, all the discrepancy is attributed to data noise





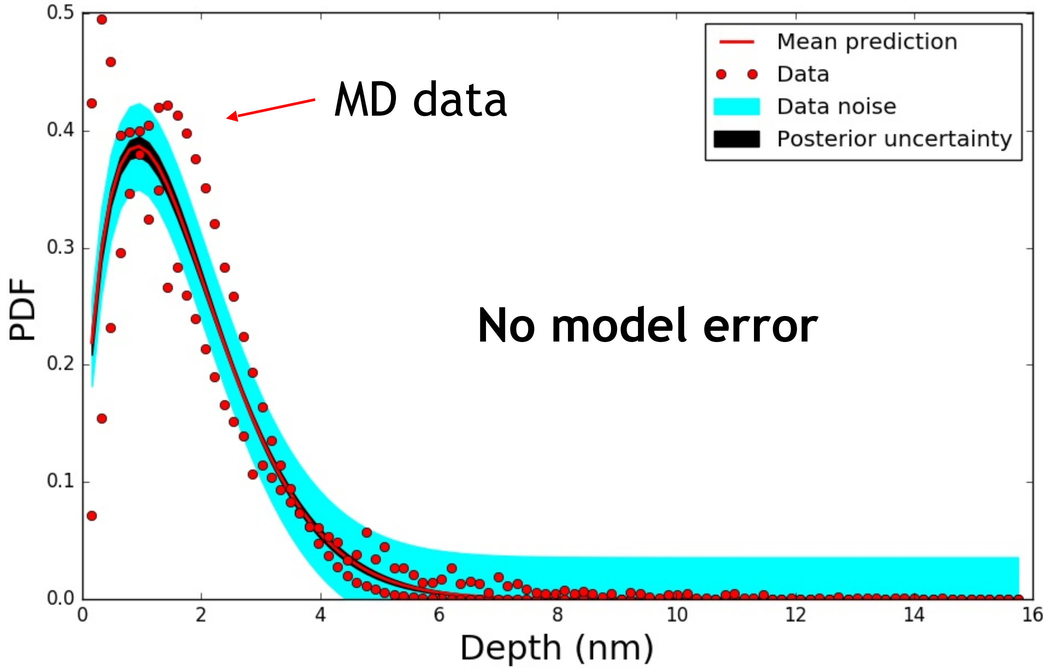


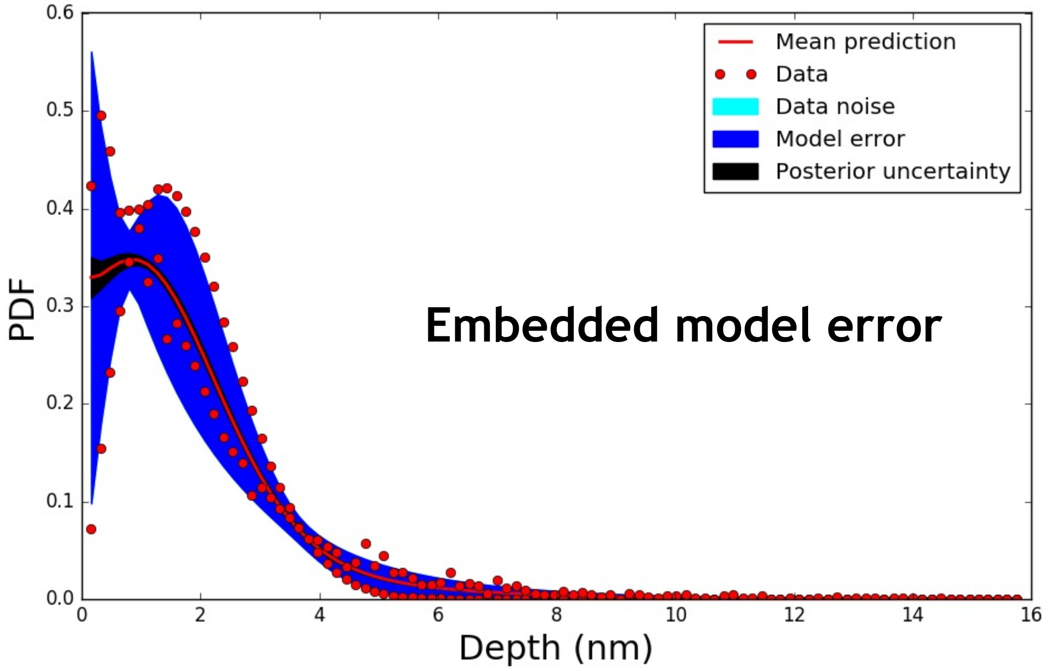
Sargsyan, Najm, Ghanem, IJCK (2015); Sargsyan, Huan, Najm, IJUQ (2019); Hakim et. al, CTM (2018)

Application: Plasma Surface Interaction funded by **FES+ASCR**

- Multi-institution partnership, but direct collaboration with ORNL and UTK
- Constructing uncertain input profiles for tungsten depth to propagate through Xolotl (PSI code)

Ignoring model error wrongly attributes uncertainty to data





O. Cekmer, K. Sargsyan, S. Blondel, H. Najm, D. Bernholdt,, B.D. Wirth, “Uncertainty quantification for incident helium flux in plasma-exposed tungsten”, *Int. J. Uncertainty Quantification*, Vol. 8, No. 5, p.429–446, 2018.

Summary

- Statistical learning of physical model parameters with Bayesian inference (inverse UQ)
- … accelerated by model surrogates (forward UQ)
- Embedded statistical model error representation allows:
  - respects physics; allows predictive variance attribution
  - stress-tested on a variety of applications
  - available via UQTk (www.sandia.gov/UQToolkit/)

Key References

- 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”, 9(4), *Int. J. Uncert. Quant.*, 365-394, 2019.

Additional slides

Model error is often the most dominant component of uncertainty

Ignoring model error leads to

- Biased parameter estimation
- Overconfident predictions

$$\underbrace{y(x_i)}_{\text{Data}} = \underbrace{f(x_i;\lambda)}_{\text{Model}} + \underbrace{\epsilon_i}_{\text{Data noise}}$$

Representing and estimating model error is useful for

- Reliable computational predictions
- Model comparison, selection

- Scientific discovery and model improvement:
  - *“is it worth resolving details, or just parameterize empirically?”*
- Optimal resource allocation:
  - *“do I improve my model (e.g. high-res), or run more simulations?”*

External correction is not satisfactory for physical models

$$y_i = \underbrace{f(x_i;\lambda) + \delta(x_i)}_{\text{truth } \; g(x_i)} + \epsilon_i$$

- Explicit additive statistical model for model error [KOH, 2001]
- Potential violation of physical constraints
- Disambiguation of model error $\delta(x_i)$ and data error ϵ_i
- Yes, priors help: [Brynjarsdottir and O'Hagan, 2014], [Plumlee, 2017]
- Calibration of model error on measured observable does not impact the quality of model predictions on other QoIs
- Physical scientists are unlikely to augment their model with a statistical model error term on select outputs
  - Calibrated predictive model: $f(x;\lambda) + \delta(x)$ or $f(x;\lambda)$?
- Problem is highlighted in model-to-model calibration ($\epsilon_i = 0$)
  - no a priori knowledge of the statistical structure of $\delta(x)$

Calibrate $f(x; \lambda)$, given data $g(x)$

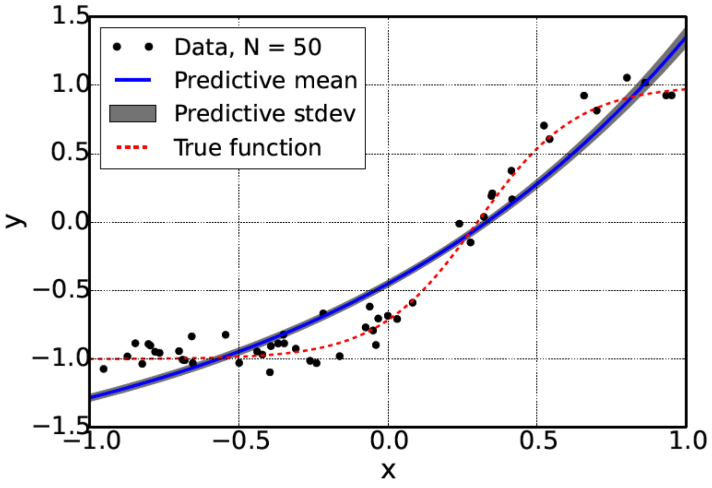
x are operating conditions, design parameters, various QoIs
λ are model parameters to be inferred/calibrated

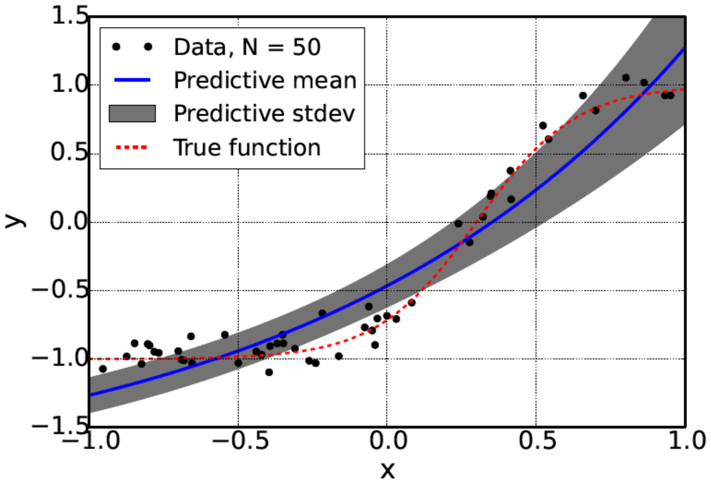
- Default: Ignore model errors: $g(x) = f(x; \lambda) + \epsilon$
  - Biased or overconfident physical parameters
  - Wrong model predictions

- Conventional: Correct for model errors: $g(x) = f(x; \lambda) + \delta(x) + \epsilon$
  - Physical parameters are ok
  - Wrong model predictions (data-specific corrections)
  - Model and data errors mixed up

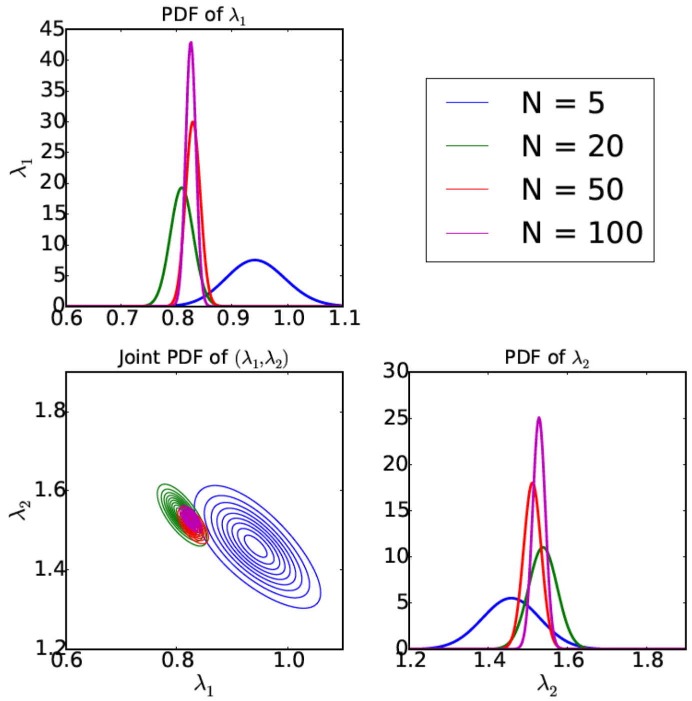
- What we do: Correct *inside* the model: $g(x) = f(x; \lambda + \delta(x)) + \epsilon$
  - Embedded model error
  - Preserves model structure and physical constraints
  - Disambiguates model and data errors
  - Allows meaningful extrapolation

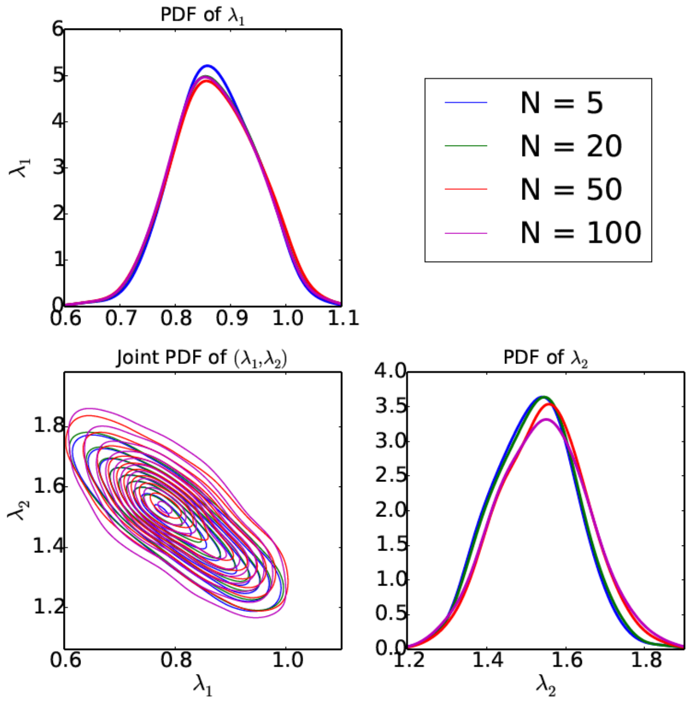
Back to toy example



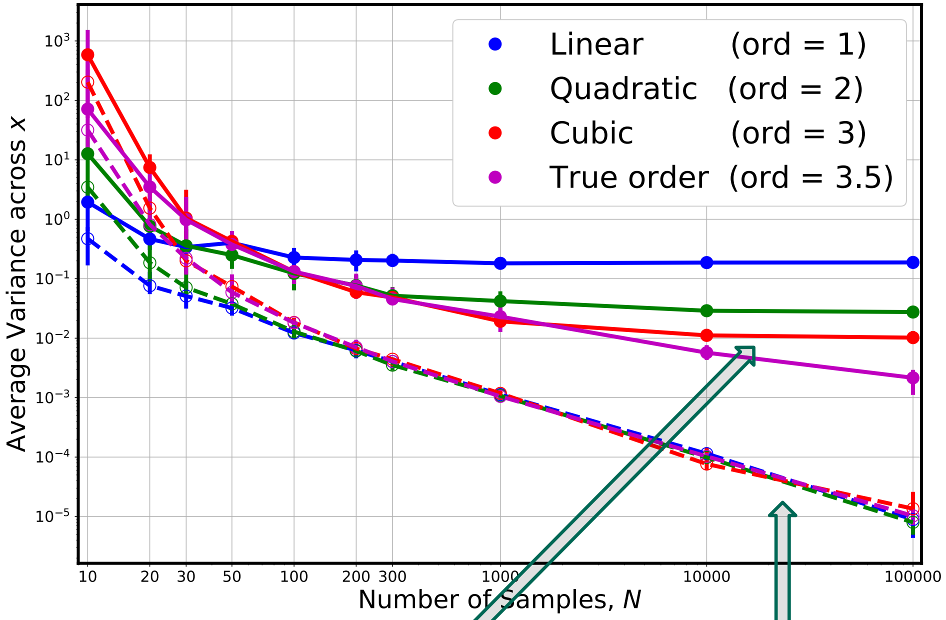


Predictive uncertainty captures model error





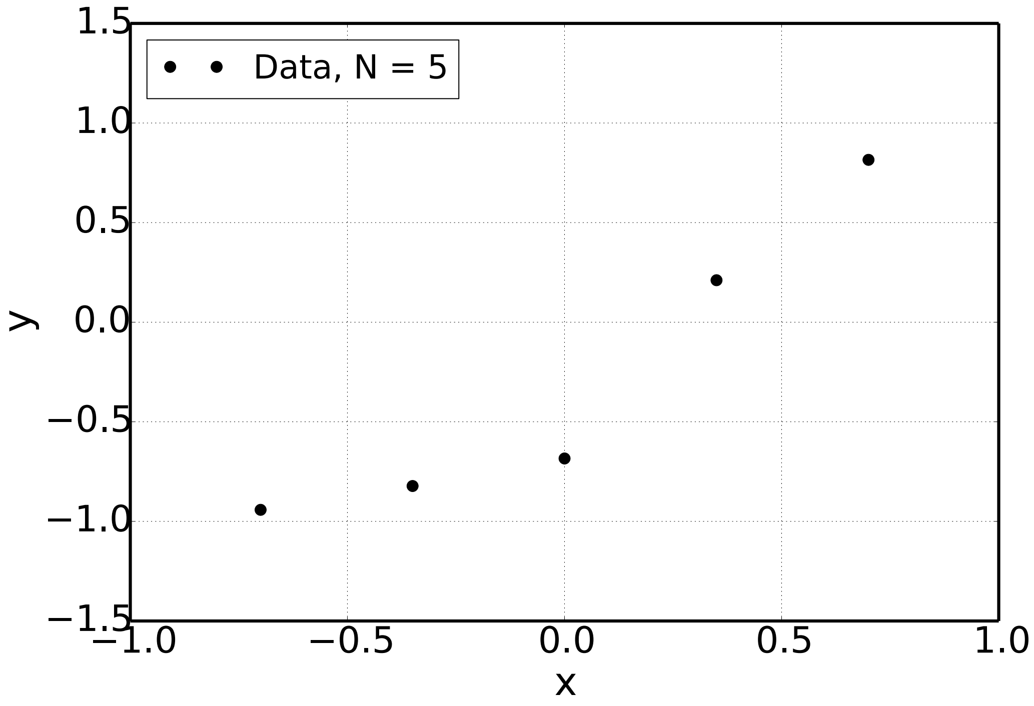
Stable prediction of "physical" parameters of the exponential function



Model error (ME)

Post. uncertainty (PU)

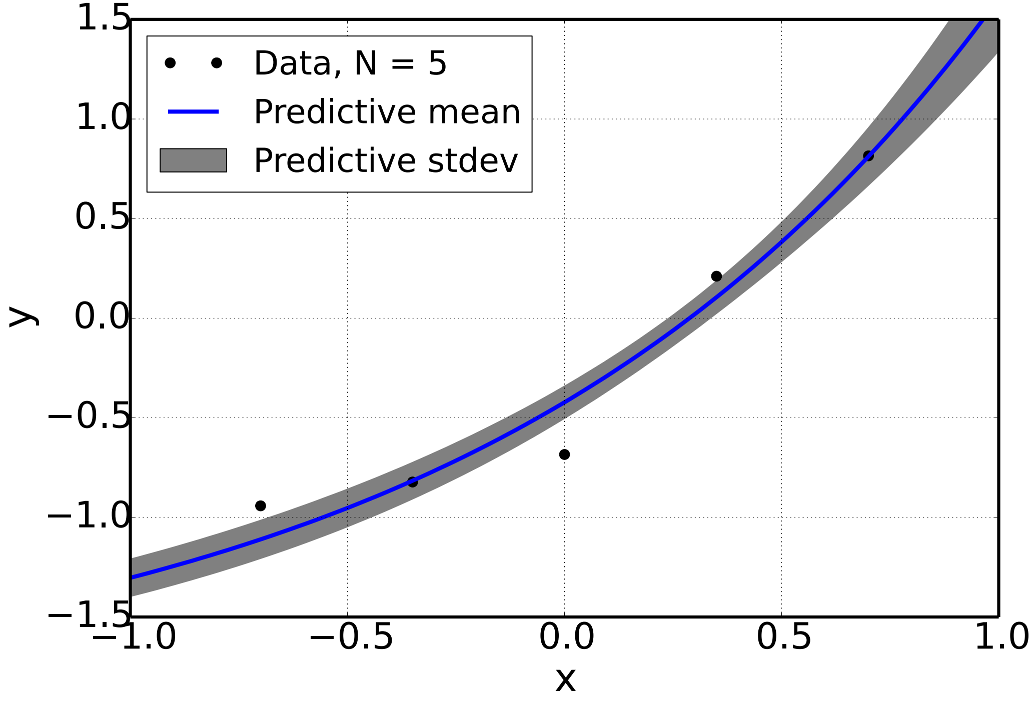
Wrong model leads to biased estimation



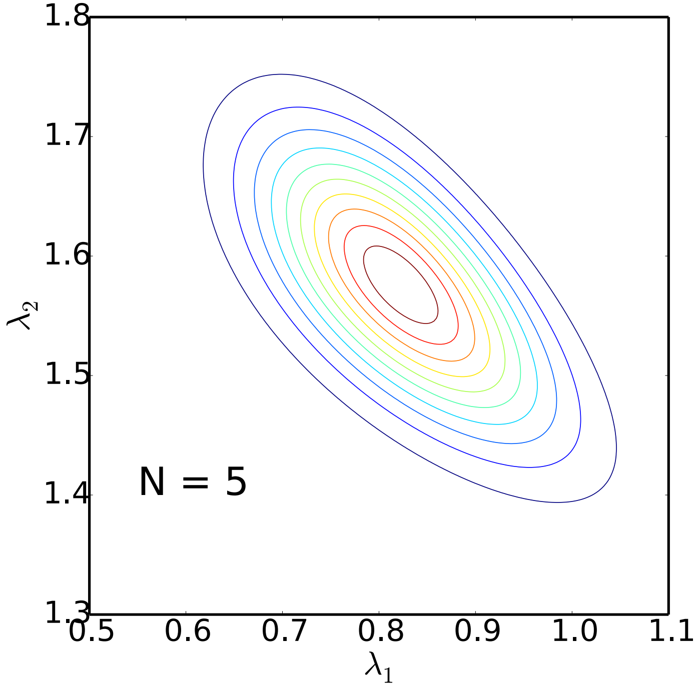
Given noisy data

Wrong model leads to biased estimation

Model prediction vs data

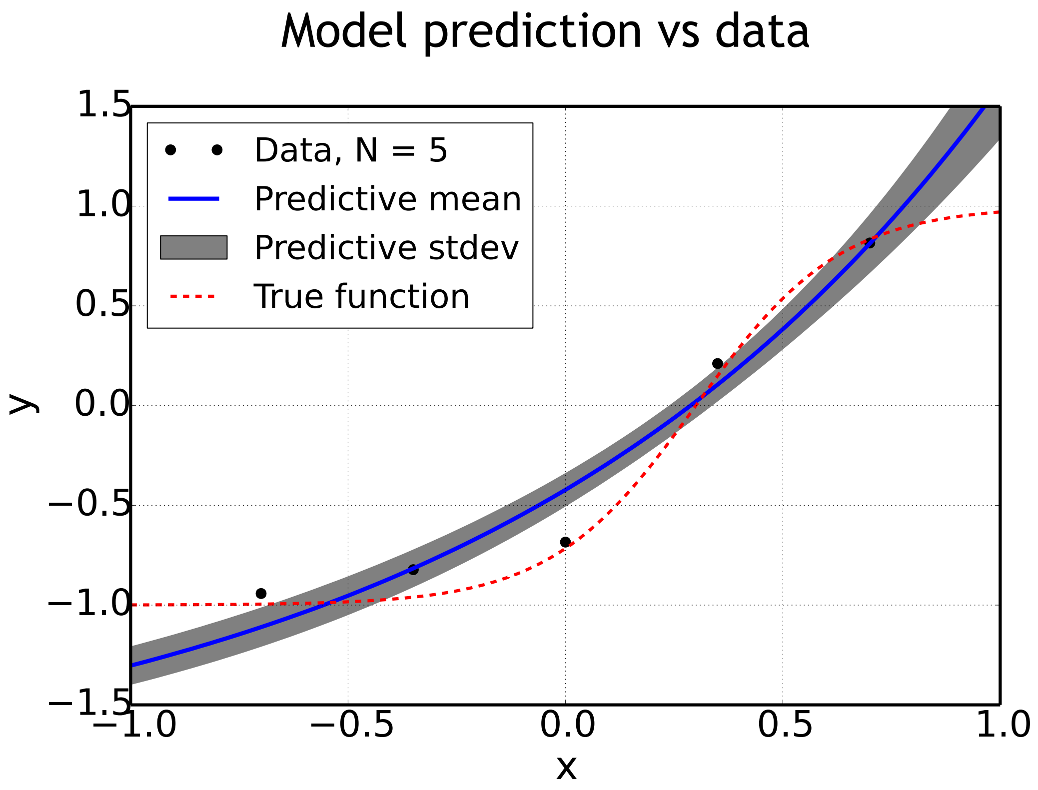


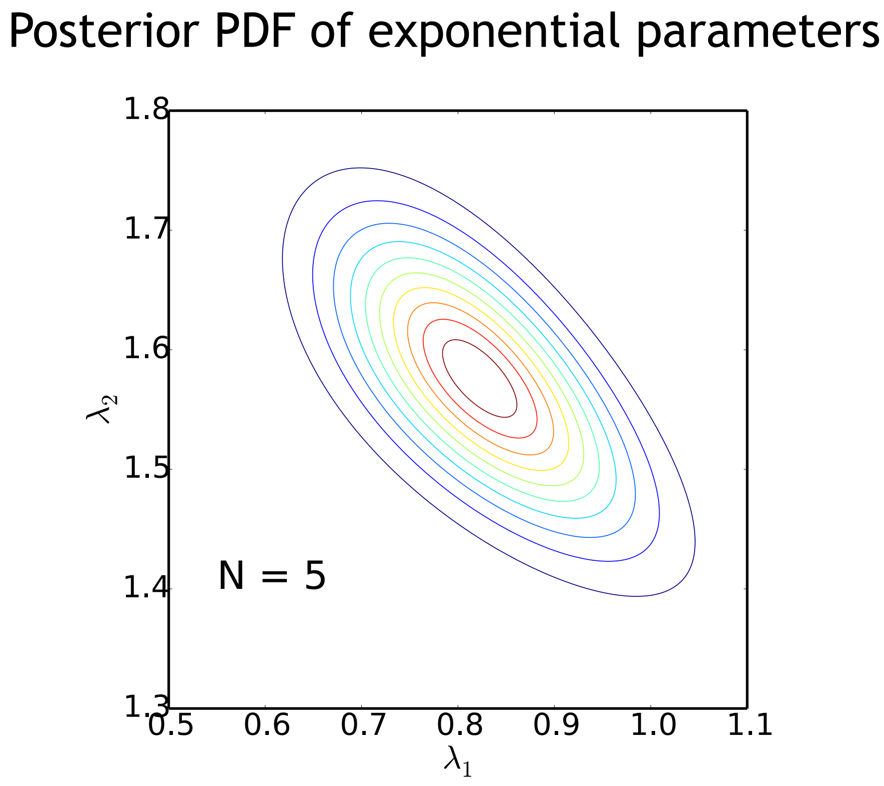
Posterior PDF of exponential parameters



Calibrate an exponential model

Wrong model leads to biased estimation

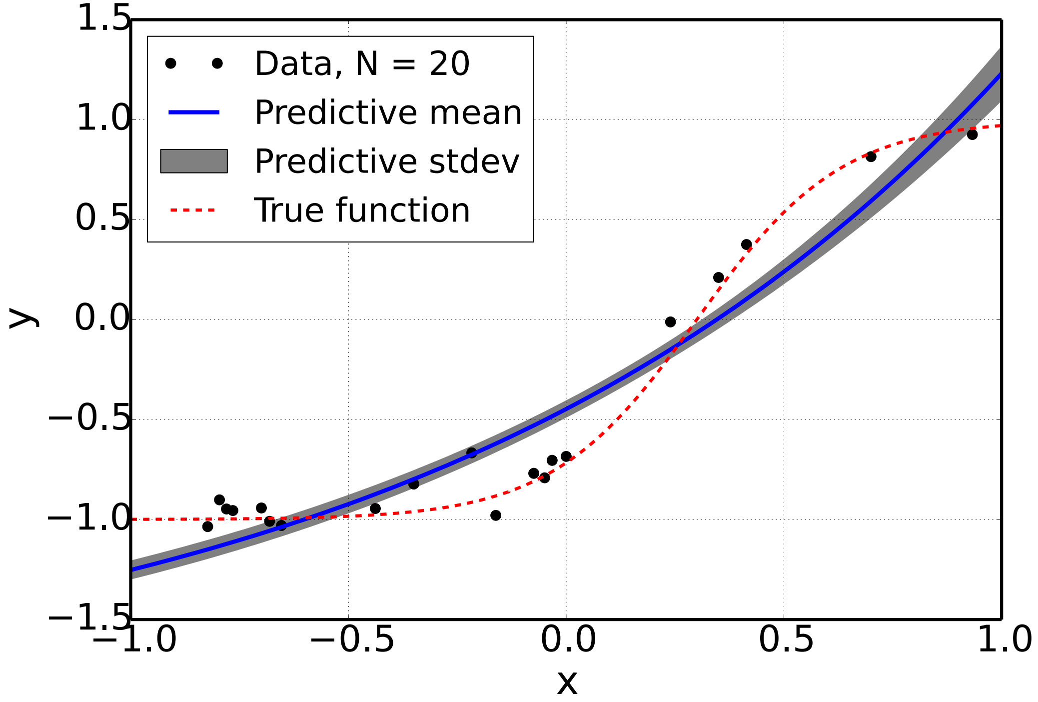




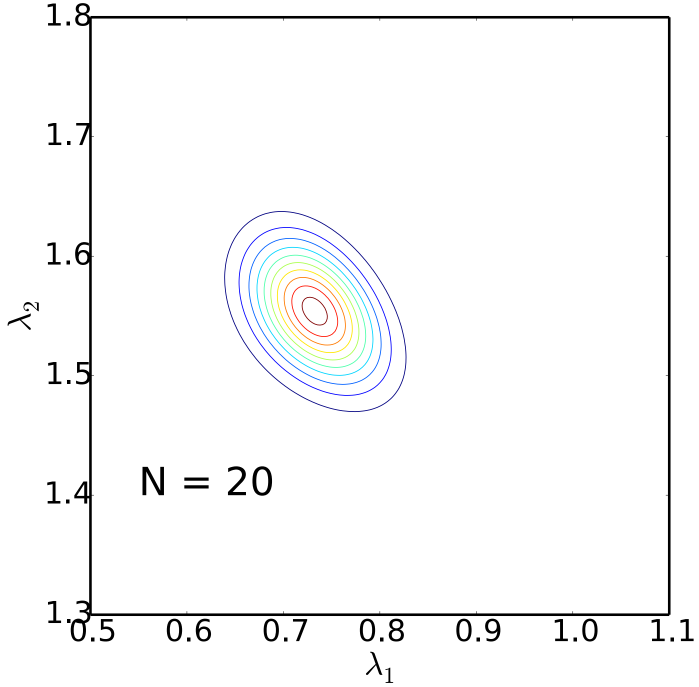
Calibrate an exponential model, but data comes from a different function (there is model error!)

Wrong model leads to biased estimation

Model prediction vs data

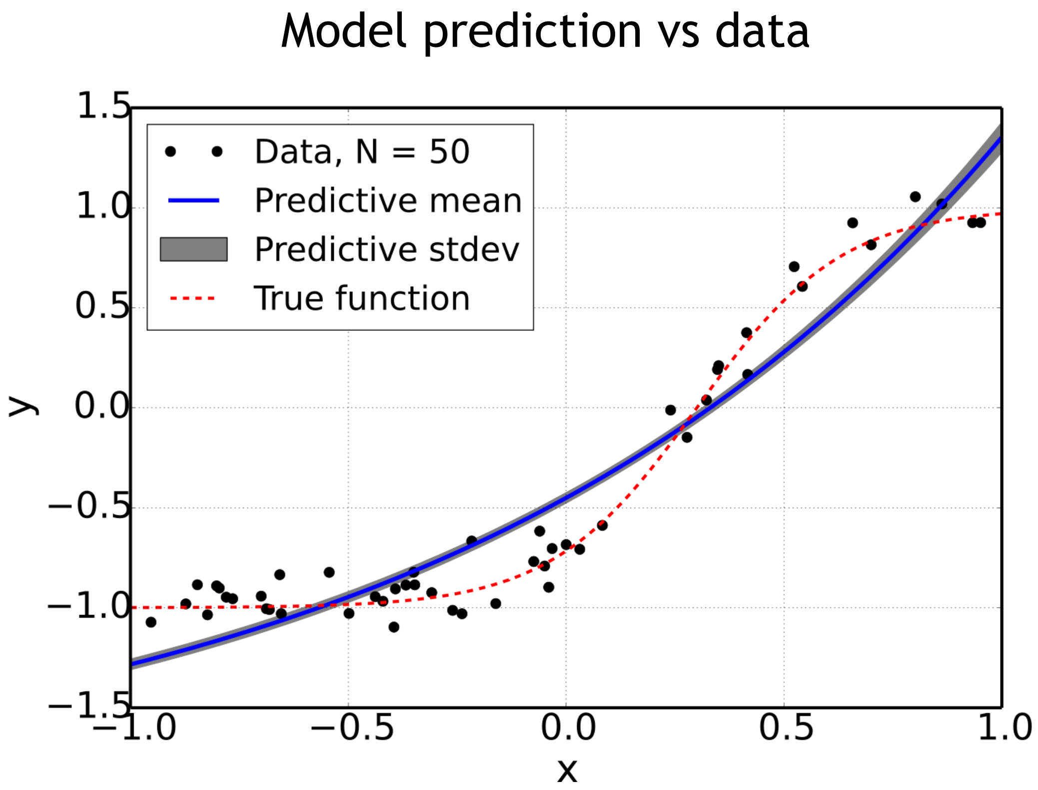


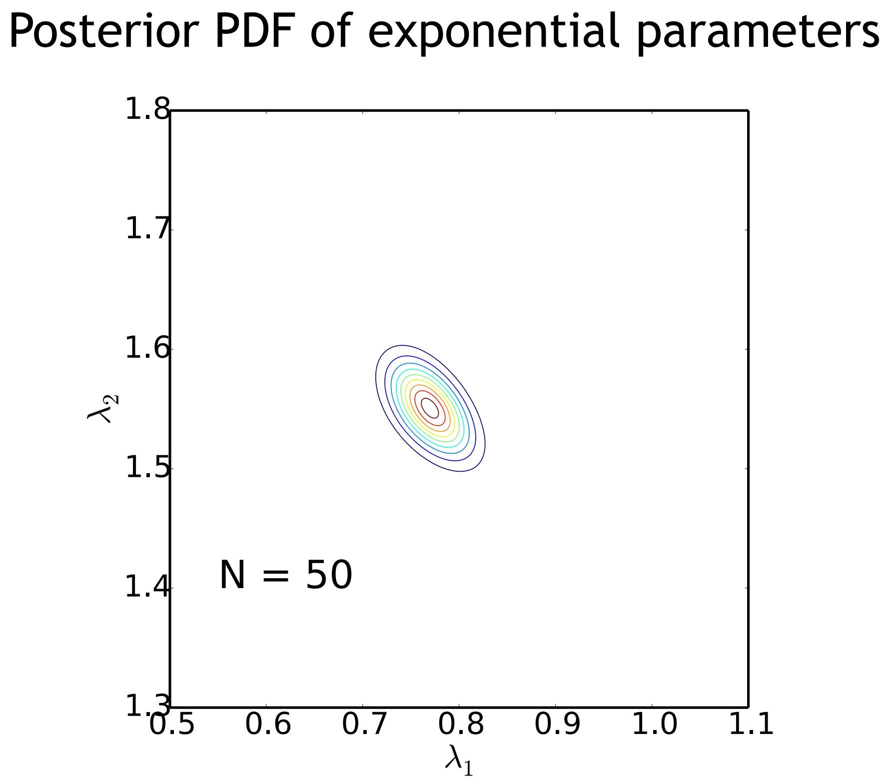
Posterior PDF of **model** parameters



Collecting more data: become increasingly sure about the wrong values of parameters

Wrong model leads to biased estimation

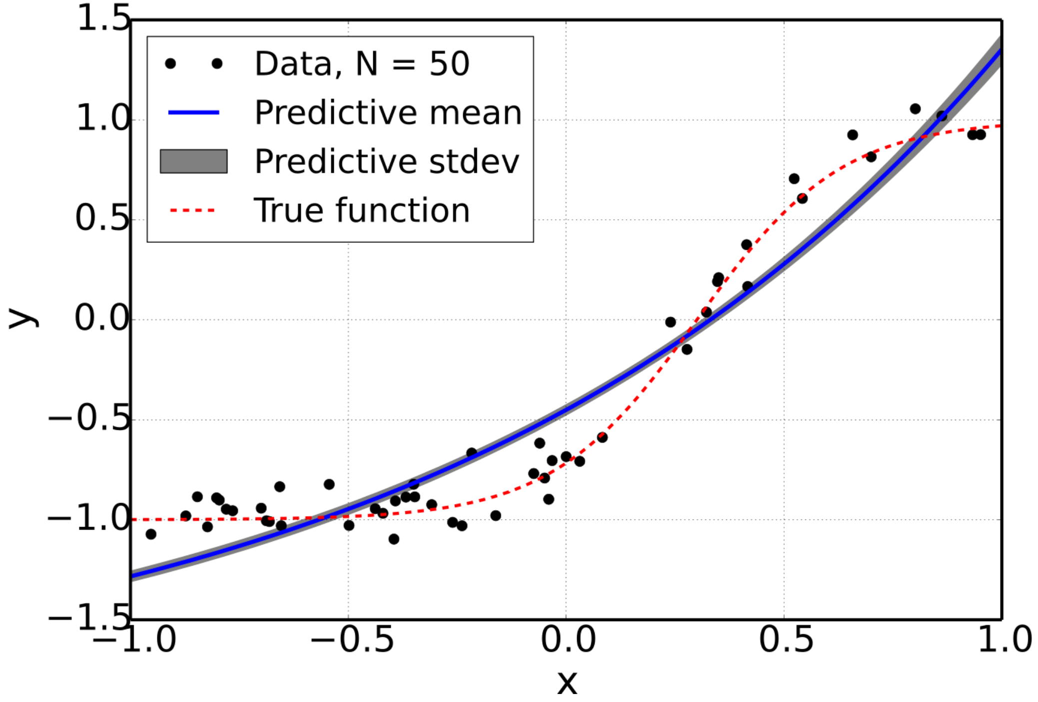




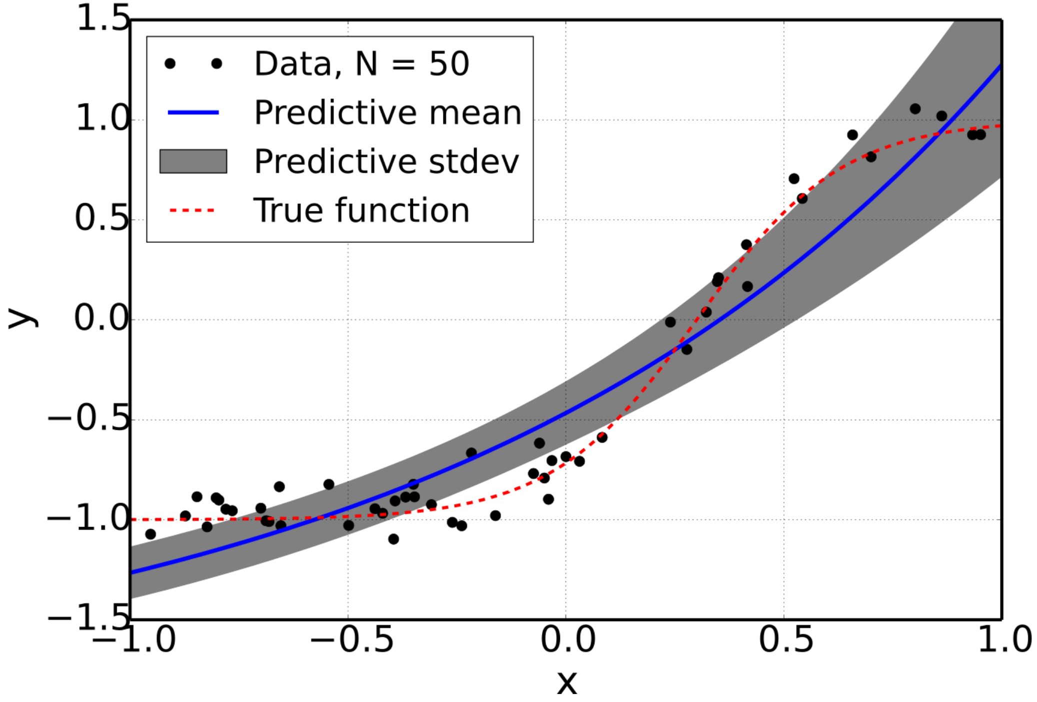
Collecting more data: become increasingly sure about the wrong values of parameters

Wrong model leads to biased estimation

Model prediction vs data



… what we actually want

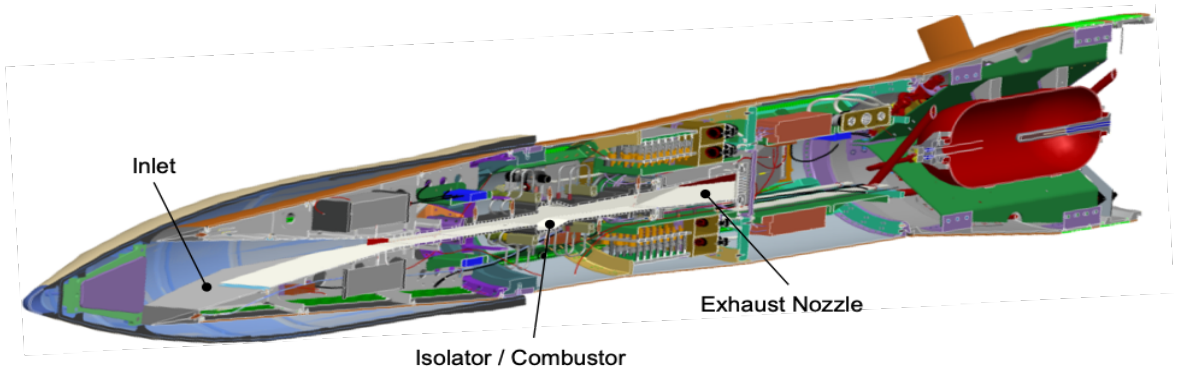


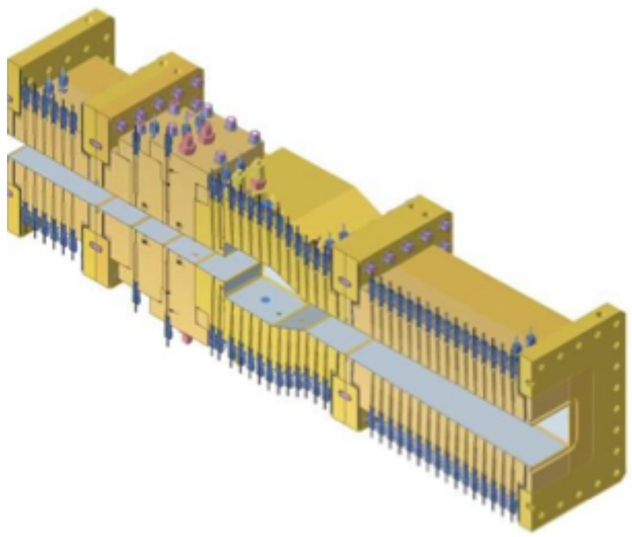
Predictive uncertainties **do not** capture model-data discrepancy

Predictive uncertainties capture model-data discrepancy

Application: Turbulent Flow

funded by **DARPA**





Large Eddy Simulation (LES) of a laboratory scale Scramjet combustor NASA Langley Hypersonic International Flight Research and Experimentation (HIFiRE) configuration

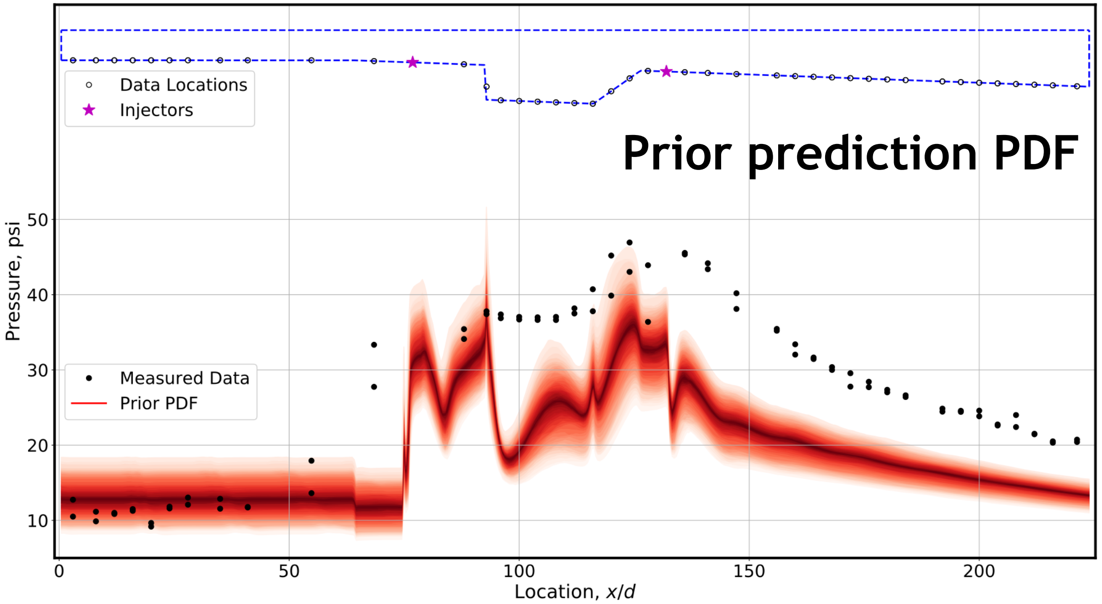
Application: Turbulent Flow

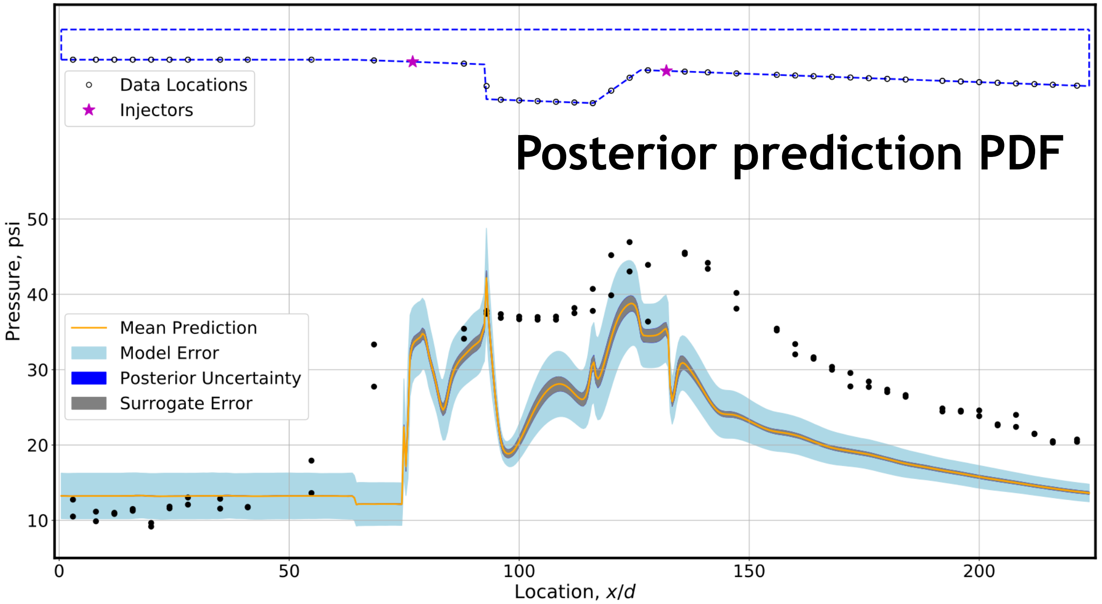
funded by **DARPA**

- Major UQ challenges for turbulent flow (LES)
  - Nonlinear dynamics
  - Large number of uncertain parameters
  - **LES model structural error**
  - Optimize design under uncertainty

Experimental data obtained from NASA

Model error is the main contributor of the predictive variance





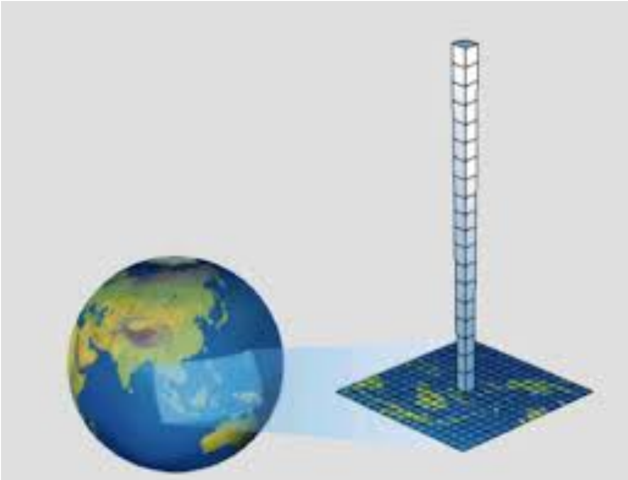
X. Huan, C. Safta, *K. Sargsyan*, G. Geraci, Michael S. Eldred, Zachary P. Vane, G. Lacaze, Joseph C. Oefelein, Habib N. Najm, "Global Sensitivity Analysis and Estimation of Model Error, toward Uncertainty Quantification in Scramjet Computations", *AIAA Journal*, Vol. 56, No. 3, p.1170–1184, 2018

Application: Earth System Land Model

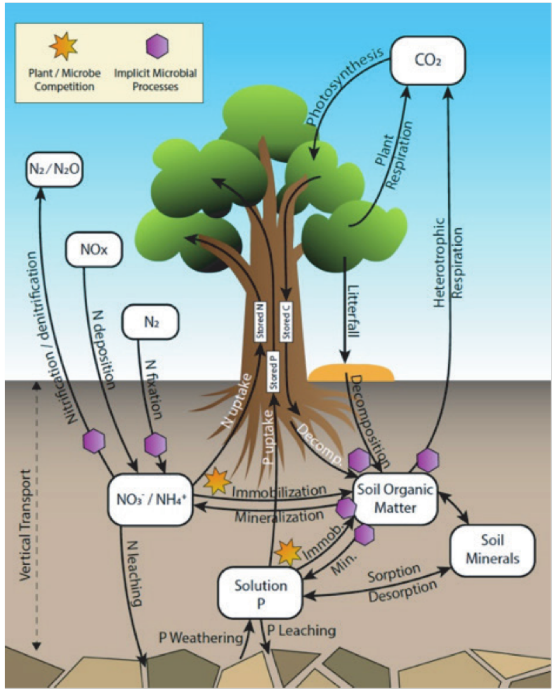
funded by **BER+ASCR**

E3SM
Energy Exascale Earth System Model

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



NERSC
National Energy Research Scientific Computing Center

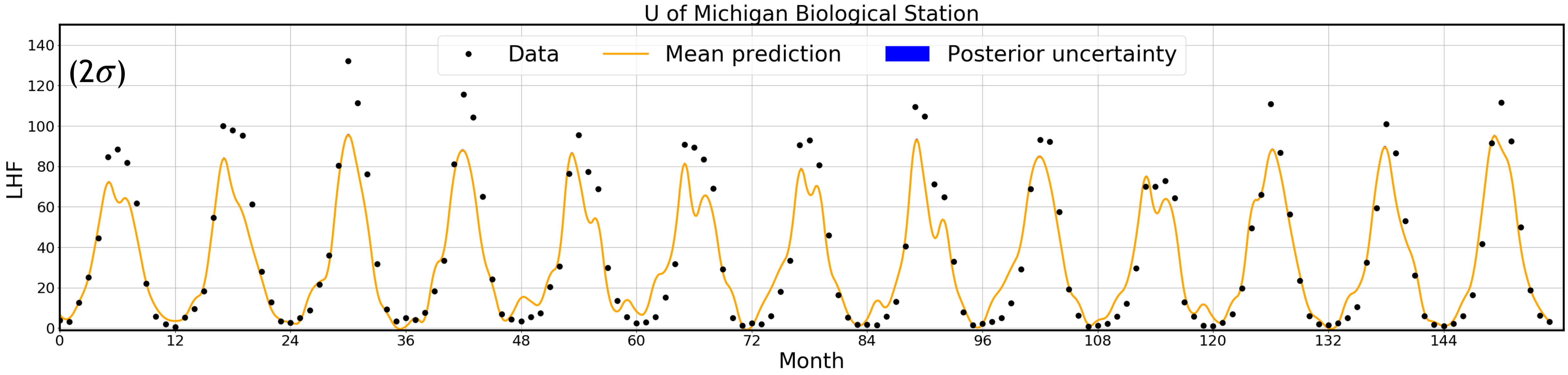


Serving as a UQ Lead for Land Model:
Direct UQ impact on several land components,
collaboration with multiple NLs

Oak Ridge National Laboratory
Berkeley Lab Lawrence Berkeley National Laboratory
Argonne National Laboratory
Pacific Northwest National Laboratory

Land model calibration given FLUXNET observations

Conventional calibration **without** model error

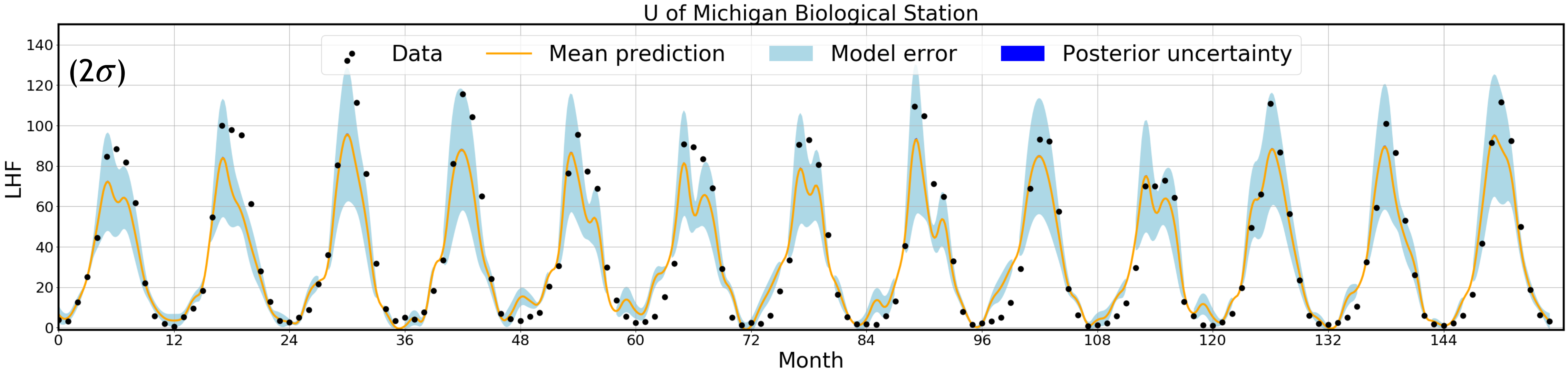


- Summer month peaks are not captured
- Posterior uncertainty negligible

LHF = Latent Heat Flux

Land model calibration given FLUXNET observations

Calibration **with** embedded model error

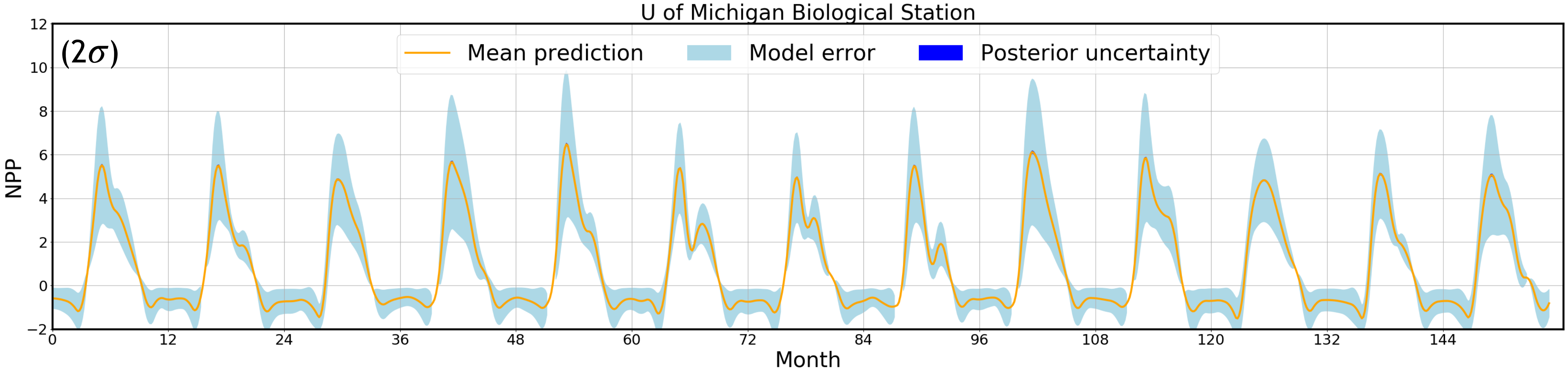


- Model error component dominates
- Captures model deficiency in summer months
- Indicates model improvement opportunities
- For further improvement: more intrusive embedding

LHF = Latent Heat Flux

Land model calibration given FLUXNET observations

Calibration **with** embedded model error

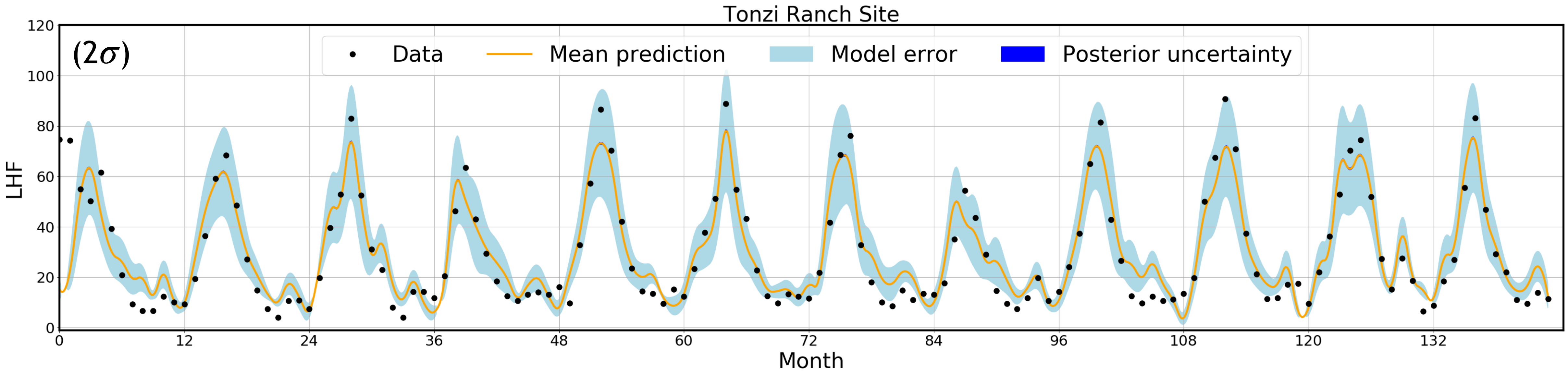


- Allows more accurate prediction of unobservable QoIs
- Can be piped to human component or atmosphere model as a boundary condition

NPP = Net Primary Productivity

Land model calibration given FLUXNET observations

Calibration **with** embedded model error



- Allows prediction at other FLUXNET sites
- Assumption: model goes wrong in a similar way

LHF = Latent Heat Flux