Embedded Model Error Representation and Propagation in Climate Models

Khachik Sargsyan[1], Xun Huan[1], Habib Najm[1], Cosmin Safta[1], Daniel Ricciuto[2], Peter Thornton[2]

[1]Sandia National Laboratories, Livermore, CA
[2]Oak Ridge National Laboratory, Oak Ridge, TN

AGU Fall Meeting
Dec 11-15, 2017

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.

Embedded Model Error Representation and Propagation in Climate Models

Khachik Sargsyan[1], Xun Huan[1], Habib Najm[1], Cosmin Safta[1], Daniel Ricciuto[2], Peter Thornton[2]

[1]Sandia National Laboratories, Livermore, CA
[2]Oak Ridge National Laboratory, Oak Ridge, TN

AGU Fall Meeting
Dec 11-15, 2017

DOE Office of Advanced Scientific Computing Research (ASCR)
DOE Office of Biological and Environmental Research (BER)

Main target: model *structural* error

deviation from 'truth' or from a higher-fidelity model

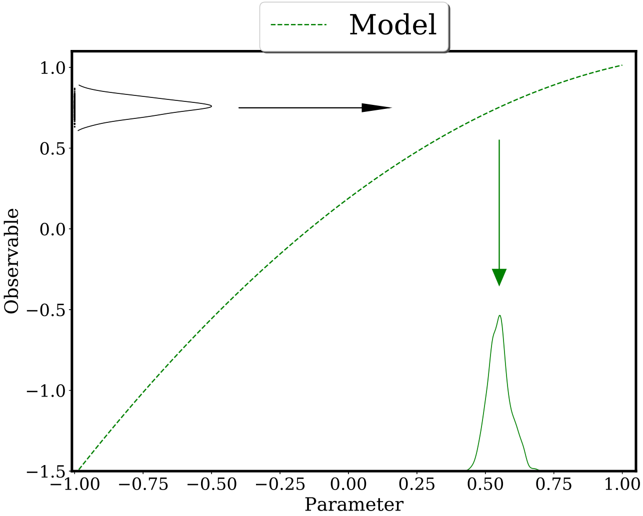
- Inverse modeling context
  - Given experimental or higher-fidelity model data, estimate the model error

---

- 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

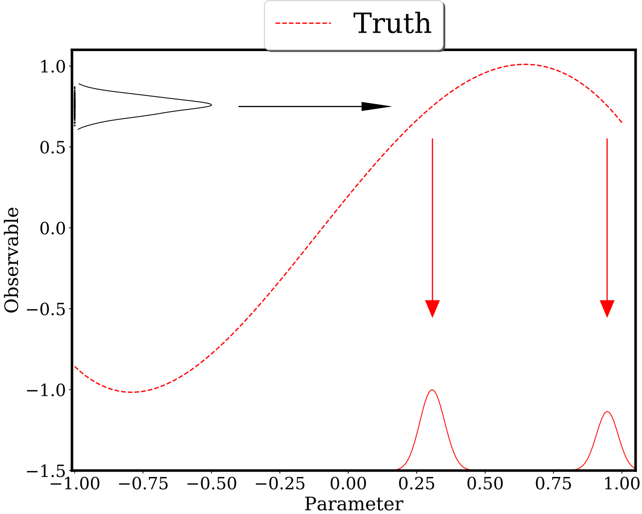
Data informs model parameters:

but what if the model is only an approximation?



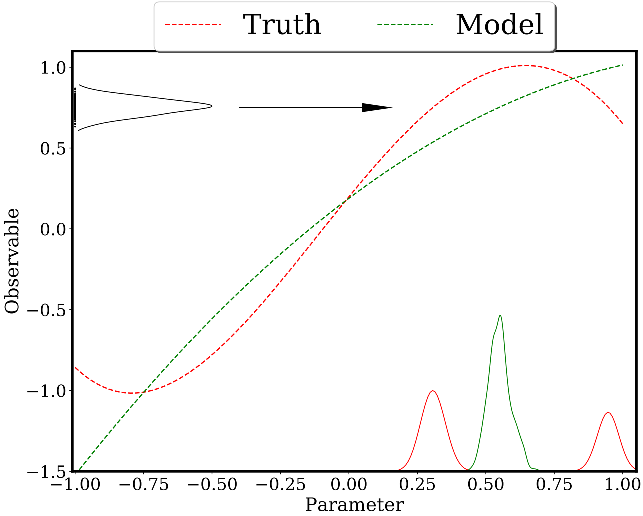
Data informs model parameters:

but what if the model is only an approximation?



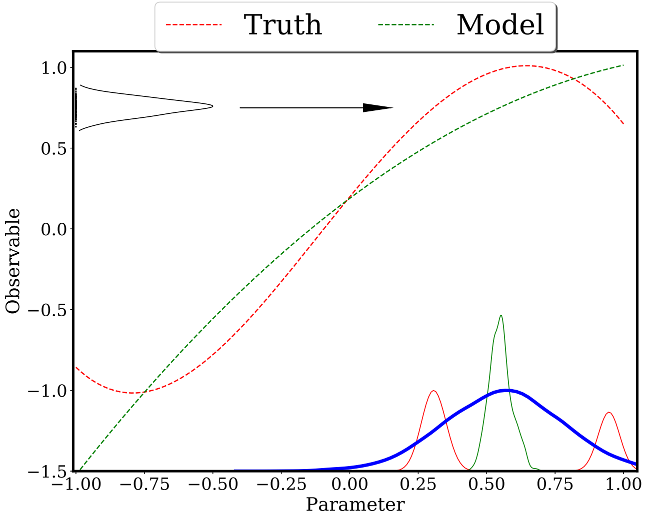
Data informs model parameters:

but what if the model is only an approximation?

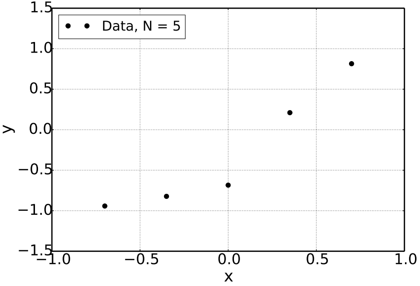


Data informs model parameters:

but what if the model is only an approximation?



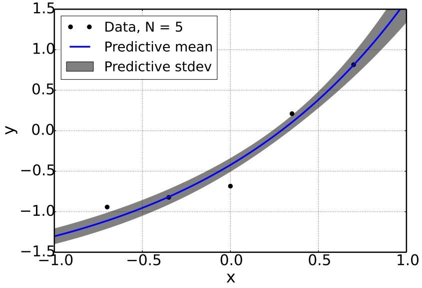
Ignoring model error leads to overconfident and biased predictions



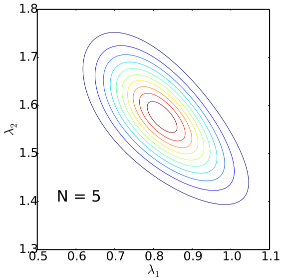
Model-data fit

- Given noisy data g, calibrate an exponential model f: $\quad g(x) \approx f(x;\lambda)$

Ignoring model error leads to overconfident and biased predictions



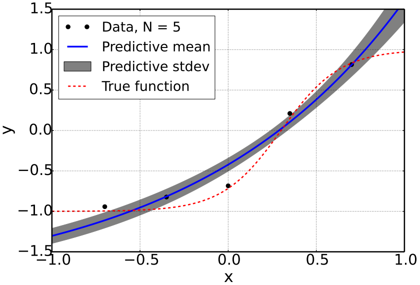
Model-data fit



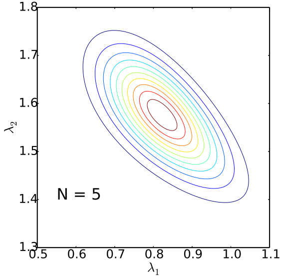
Posterior on parameters

- Given noisy data g, calibrate an exponential model f: $g(x) \approx f(x;\lambda)$
- Employ Bayesian inference to obtain posterior PDFs on λ

Ignoring model error leads to overconfident and biased predictions



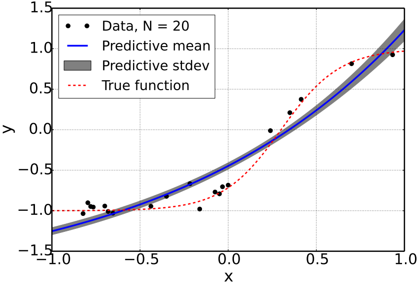
Model-data fit



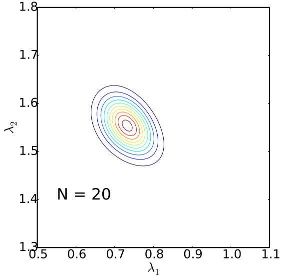
Posterior on parameters

- Given noisy data g, calibrate an exponential model f: $g(x) \approx f(x; \lambda)$
- Employ Bayesian inference to obtain posterior PDFs on λ
- True model – dashed-red – is *structurally* different from fit model $f(x, \lambda)$

Ignoring model error leads to overconfident and biased predictions



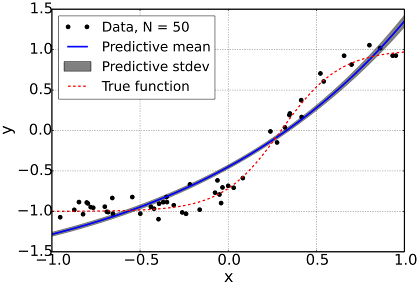
Model-data fit



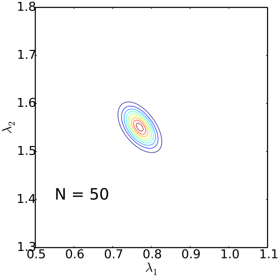
Posterior on parameters

- Given noisy data g, calibrate an exponential model f: $g(x) \approx f(x; \lambda)$
- Employ Bayesian inference to obtain posterior PDFs on λ
- True model – dashed-red – is *structurally* different from fit model $f(x, \lambda)$
- Higher data amount reduces posterior and predictive uncertainty
  - Increasingly sure about predictions based on the *wrong* model

Ignoring model error leads to overconfident and biased predictions



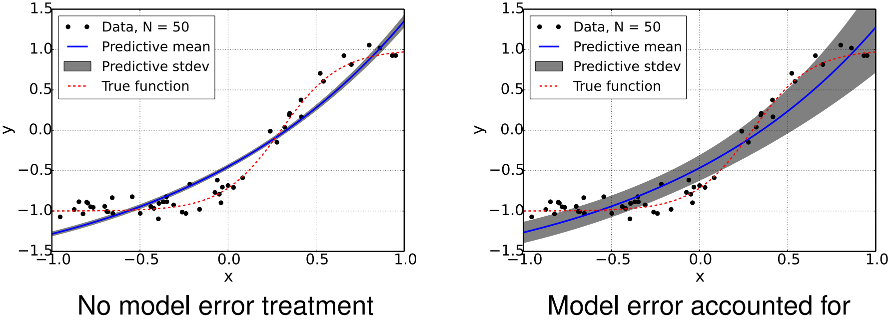
Model-data fit



Posterior on parameters

- Given noisy data g, calibrate an exponential model f: $\quad g(x) \approx f(x; \lambda)$
- Employ Bayesian inference to obtain posterior PDFs on λ
- True model – dashed-red – is *structurally* different from fit model $f(x, \lambda)$
- Higher data amount reduces posterior and predictive uncertainty
  - Increasingly sure about predictions based on the *wrong* model

Ignoring model error leads to overconfident and biased predictions



- Given noisy data g, calibrate an exponential model f: $\quad g(x) \approx f(x;\lambda)$
- Employ Bayesian inference to obtain posterior PDFs on λ
- True model – dashed-red – is *structurally* different from fit model $f(x,\lambda)$
- Accounting for model error allows extra uncertainty component to propagate through predictions

Explicit model discrepancy: issues 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 $\delta(x)$ [Kennedy-O'Hagan, 2001]
- Potential violation of physical constraints
- Disambiguation of model error $\delta(x_i)$ and data error ϵ_i
- 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 $\delta(x)$

Key Idea: Model Error Embedding

Ideally, modelers want predictive *errorbars*:
inserting randomness on the outputs has issues, so...

- Augment input parameters λ with a stochastic term δ_α

  x*-independent* $$y_i = f(x_i; \lambda + \delta_\alpha) + \epsilon_i$$

- Generalize parameter forms,

  Random field $$y_i = f(x_i; \lambda + \delta_\alpha(x_i)) + \epsilon_i$$

- More generally, explore additional parameterizations,

  Intrusive $$y_i = \tilde{f}(x_i; \lambda, \delta_\alpha(x_i)) + \epsilon_i$$

Non-Intrusive Probabilistic Embedding

Additive corrections δ_α for input parameters λ

$$y_i = f(x_i; \lambda + \delta_\alpha) + \epsilon_i$$

- Embed model error in specific submodel phenomenology
  - a modified transport or constitutive law
  - a modified formulation for a material property
  - turbulent model constants
- Allows placement of model error term in locations where key modeling assumptions and approximations are made
  - as a correction or high-order term
  - as a possible alternate phenomenology
- Naturally preserves model structure and physical constraints
- Disambiguates model/data errors

Bayesian Framework for Model Error Estimation

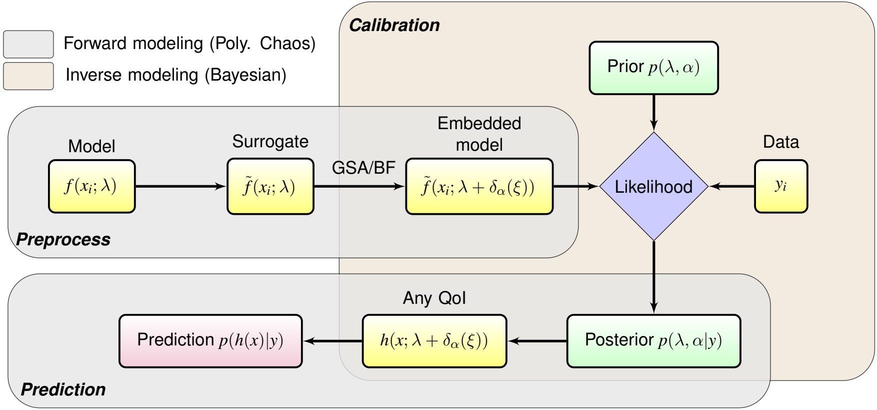
$$y_i = f(x_i; \lambda + \delta_\alpha) + \epsilon_i$$

- Given data y_i, perform *simultaneous* estimation of $\tilde{\alpha} = (\lambda, \alpha)$, i.e. model parameters λ and model-error parameters α.
- Bayes' theorem

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

- In order to estimate the likelihood $L_y(\tilde{\alpha}) = p(y|\tilde{\alpha}) = p(y|\lambda, \alpha)$, one needs uncertainty propagation through $f(x_i; \underbrace{\lambda + \delta_\alpha}_{\text{stochastic}})$,
- ... hence, we employ Polynomial Chaos (PC) representation for δ_α.

Model error embedding – workflow



- Predictive uncertainty decomposition: Total Variance =

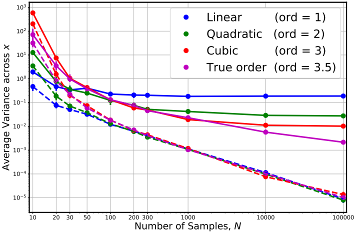
Parametric uncertainty + Data noise + Model error + Surrogate error

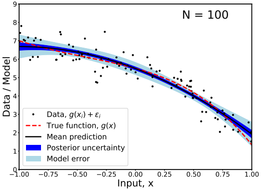
More data leads to ‘leftover’ model error

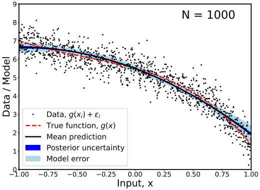
Calibrating a quadratic $f(x) = \lambda_0 + \lambda_1 x + \lambda_2 x^2$
w.r.t. ‘truth’ $g(x) = 6 + x^2 - 0.5(x+1)^{3.5}$ measured with noise $\sigma = 0.1$.

Summary of features:

- Well-defined model-to-model calibration
- Model-driven discrepancy correlations
- Respects physical constraints
- Disambiguates model and data errors
- Calibrated predictions of multiple QoIs

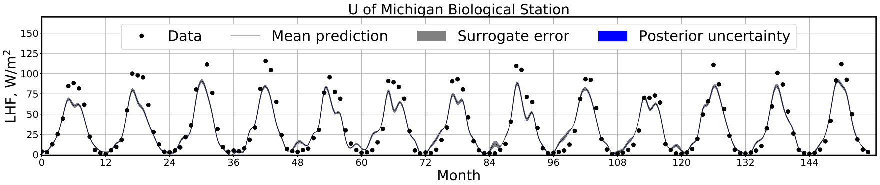






E3SM Land Model (ELM)

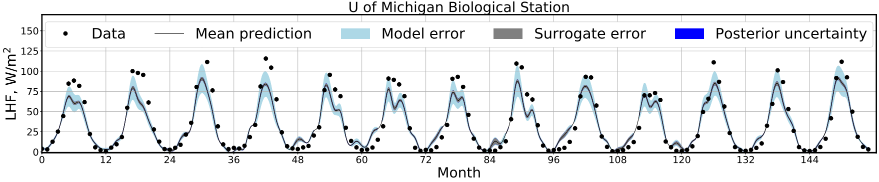
- US Department of Energy (DOE) sponsored Earth system model
- Land, atmosphere, ocean, ice, human system components
- High-resolution, employ DOE leadership-class computing facilities



- Predictive variance decomposition with model-error component

E3SM Land Model (ELM)

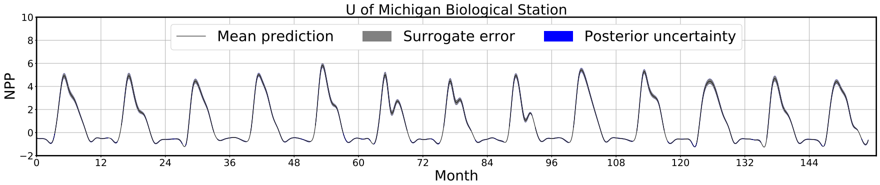
- US Department of Energy (DOE) sponsored Earth system model
- Land, atmosphere, ocean, ice, human system components
- High-resolution, employ DOE leadership-class computing facilities



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

E3SM Land Model (ELM)

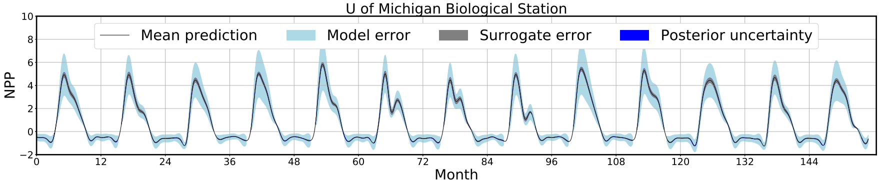
- US Department of Energy (DOE) sponsored Earth system model
- Land, atmosphere, ocean, ice, human system components
- High-resolution, employ DOE leadership-class computing facilities



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

E3SM Land Model (ELM)

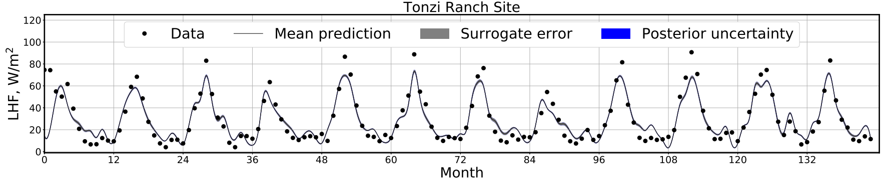
- US Department of Energy (DOE) sponsored Earth system model
- Land, atmosphere, ocean, ice, human system components
- High-resolution, employ DOE leadership-class computing facilities



- 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)

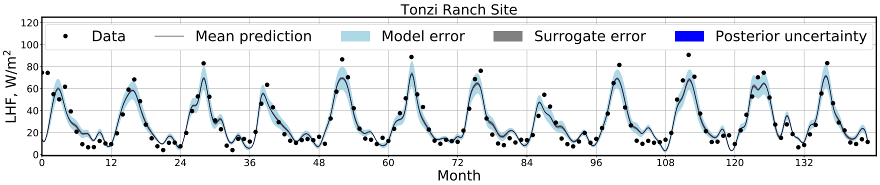
- US Department of Energy (DOE) sponsored Earth system model
- Land, atmosphere, ocean, ice, human system components
- High-resolution, employ DOE leadership-class computing facilities



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

E3SM Land Model (ELM)

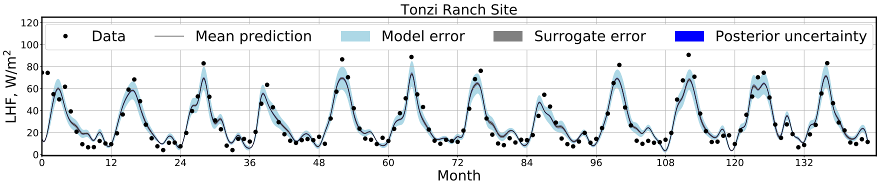
- US Department of Energy (DOE) sponsored Earth system model
- Land, atmosphere, ocean, ice, human system components
- High-resolution, employ DOE leadership-class computing facilities



- 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



- Predictive variance decomposition with model-error component
- For surrogate construction (forward UQ) under the hood, see poster by C. Safta [NG33A-0190: Machine Learning Techniques for Global Sensitivity Analysis in Climate Models] Wednesday afternoon.

Inference library in UQTk v3.0 (www.sandia.gov/uqtoolkit)

- Workflow for model error representation, quantification and propagation
- Custom components: forward model, likelihood and prior
- A range of common forward models, including polynomial surrogates
- Various likelihood options, including classical, Kennedy-O'Hagan, model-error-embedding and its approximations
- Several prior options for embedded parameters α, including Wishart, Jeffreys, range-constrained
- All pieces – forward model, likelihood and prior – can be made custom

Summary

Thank You

- **Represent, quantify and propagate model structural errors**
- Bayesian machinery for simultaneous estimation of physical parameters and model error
- A principled guide for model exploration (embedded representation, but can be performed *non-intrusively*!)
- Differentiates from data noise; allows model-to-model calibration
- Connections with Bayesian model averaging, model 'nudging', and stochastic physics
- Besides climate models, applied successfully in LES, transport models, chemistry, fusion

---

- K. Sargsyan, H. Najm, and R. Ghanem. "On the Statistical Calibration of Physical Models". *International Journal for Chemical Kinetics*, 47(4): 246-276, 2015.
- K. Sargsyan, X. Huan, and H. Najm. "Embedded Model Error Representation for Model Calibration", to be submitted, *Journal of Computational Physics*, 2017.

Summary Thank You

- **Represent, quantify and propagate model structural errors**
- Bayesian machinery for simultaneous estimation of physical parameters and model error
- A principled guide for model exploration (embedded representation, but can be performed *non-intrusively*!)
- Differentiates from data noise; allows model-to-model calibration
- Connections with Bayesian model averaging, model 'nudging', and stochastic physics
- Besides climate models, applied successfully in LES, transport models, chemistry, fusion

---

We are hiring!

- **Postdoctoral Position** UQ-in-Climate at Sandia National Labs
- Go to Sandia careers' website and look for job ID 659182
- Experience with UQ, climate modeling, coding.
- Salary $85700+/year, in Livermore, CA

Additional Material

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

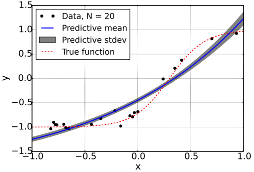
- **Conventional:** Correct for model errors: $g(x) = f(x;\lambda) + \delta(x) + \epsilon$
  - Physical parameters are ok
  - Wrong model predictions (data-specific corrections)
  - Model and data errors mixed up

- **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

Data-Model-Truth

- Measurements

$$\overset{\text{data}}{y_i} \;=\; \overset{\text{truth}}{g(x_i)} \;+\; \overset{\text{data error}}{\epsilon_i^{\text{d}}}$$



- Model

$$\overset{\text{truth}}{g(x_i)} = \overset{\text{model}}{f(x_i;\lambda)} \;+\; \overset{\text{model error}}{\delta(x_i)}$$

- Total error budget

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

Explicit statistical modeling of model discrepancy/error $\delta(x)$

Model Error: $\delta(x) \sim \text{GP}(\mu(x), C(x,x'))$

Data Error: $\epsilon_i^{\text{d}} \sim \text{N}(0,\sigma^2)$

Estimate model parameters λ along with those of $\delta(x)$, ϵ_i^{d}

Polynomial Chaos Representation of Augmented Input

$$y_i = f(x_i; \lambda + \delta_\alpha) + \epsilon_i$$

- Zero-mean PC form $\delta_\alpha = \sum_{k=1}^{K} \alpha_k \Psi_k(\xi)$
- Functional representation of a large class of random variables
- The PC *germ* ξ is a standard random variable
  - e.g. Uniform$(-1, 1)$ or Normal$(0, 1)$
- The PC bases (e.g. Legendre or Hermite polynomials) are orthogonal w.r.t. PDF of ξ

$$\int \Psi_m(\xi)\Psi_k(\xi)\pi_\xi(\xi)d\xi = 0 \quad \text{for } m \neq k.$$

- PC representation allows efficient
  - Sampling
  - Moment estimation
  - Variance-based decomposition
  - Uncertainty propagation (via NISP, see next slide)

Non-intrusive Spectral Projection (NISP) for Uncertainty Propagation

- Input random variable represented as PC

$$\Lambda(\xi) = \sum_k \alpha_k \Psi_k(\xi)$$

- Black-box forward model $Z = f(\Lambda)$
- Seeking PC representation of output random variable

$$Z(\xi) = \sum_k z_k \Psi_k(\xi)$$

- Use orthogonality property and quadrature integration to find PC coefficients

$$z_k = \frac{1}{||\Psi_k||^2} \int f(\Lambda(\xi))\Psi_k(\xi)\pi_\xi(\xi)d\xi \approx \frac{1}{||\Psi_k||^2} \sum_q f(\Lambda(\xi^{(q)}))\Psi_k(\xi^{(q)})w^{(q)}$$

Likelihood construction: data model

- Data $y_i = g(x_i) + \epsilon_i$
- Model $f(x_i; \Lambda)$
- Model input as a PC $\Lambda = \lambda + \delta_\alpha = \sum_k \alpha_k \Psi_k(\xi_1, \ldots, \xi_d)$

- Data generation model

$$\begin{aligned} y_i &= f(x_i, \lambda + \delta_\alpha) + \epsilon_i = \\ &= f\left(x_i, \sum_k \alpha_k \Psi_k(\xi_1, \ldots, \xi_d)\right) + \sigma \xi_{d+i} = \\ &\overset{NISP}{\approx} \sum_k f_{ik}(\tilde{\alpha}) \Psi_k(\xi_1, \ldots, \xi_d) + \sigma \xi_{d+i} \end{aligned}$$

- Likelihood $L_y(\tilde{\alpha}) = p(y|\tilde{\alpha})$ for $\tilde{\alpha} = (\lambda, \alpha)$ and its construction directly follows, via sampling or moment extraction.

Model Error – Likelihood options

$$y_i = \sum_k f_{ik}(\tilde{\alpha})\Psi_k(\xi_1,\ldots,\xi_d) + \sigma\xi_{d+i}$$

- True Likelihood:

$$L_y(\tilde{\alpha}) = p(y|\tilde{\alpha}) = p(y_1,\ldots,y_N|\tilde{\alpha}) = \pi(y)$$

  - Degenerate if no data noise
  - Requires multivariate kernel density estimation (KDE) or high-d integration
  - Gaussian approximation:

$$L_y(\tilde{\alpha}) \propto \exp\left(-\frac{1}{2}(y-\mu(\tilde{\alpha}))^T\Sigma^{-1}(\tilde{\alpha})(y-\mu(\tilde{\alpha}))\right)$$

  - NISP PC relieves the expense and provides easy access to mean $\mu(\tilde{\alpha})$ and covariance $\Sigma(\tilde{\alpha})$

Model Error – Likelihood options

$$y_i = \sum_k f_{ik}(\tilde{\alpha})\Psi_k(\xi_1, \ldots, \xi_d) + \sigma\xi_{d+i}$$

- Marginalized Likelihood:

$$L_y(\tilde{\alpha}) = p(y|\tilde{\alpha}) \approx \prod_{i=1}^{N} p(y_i|\tilde{\alpha}) = \prod_{i=1}^{N} \pi(y_i)$$

  - Requires univariate KDE
  - Neglects built-in correlations - looks for a pointwise match
  - Gaussian approximation:

$$L_y(\tilde{\alpha}) \propto \exp\left(-\frac{1}{2}\sum_{i=1}^{N} \Sigma_{ii}^{-1}(\tilde{\alpha})(y_i - \mu_i(\tilde{\alpha}))^2\right)$$

  - NISP PC relieves the expense and provides easy access to marginal means $\mu_i(\tilde{\alpha})$ and variances $\Sigma_{ii}(\tilde{\alpha})$

Model Error – Likelihood options

$$y_i = \sum_k f_{ik}(\tilde{\alpha})\Psi_k(\xi_1,\ldots,\xi_d) + \sigma\xi_{d+i}$$

- Approximate Bayesian Computation (ABC):

$$L_y(\tilde{\alpha}) = \frac{1}{\epsilon}K\left(\frac{\rho(\mathcal{S}_{\mathcal{M}},\mathcal{S}_{\mathcal{D}})}{\epsilon}\right)$$

  - Mean of $f(x_i;\Lambda)$ is “centered” on the data
  - The width of the distribution of $f(x_i;\Lambda)$ is consistent with the spread of the data around the nominal model prediction

$$L_y(\tilde{\alpha}) \propto \exp\left(-\frac{1}{2\epsilon^2}\sum_{i=1}^{N}\left[(\mu_i(\tilde{\alpha})-y_i)^2 + \left(\sqrt{\Sigma_{ii}(\tilde{\alpha})}-\gamma|\mu_i(\tilde{\alpha})-y_i|\right)^2\right]\right)$$

  - NISP PC relieves the expense and provides easy access to marginal means $\mu_i(\tilde{\alpha})$ and variances $\Sigma_{ii}(\tilde{\alpha})$

Optimal Embedding via Bayes Factors

- **Question:** which parameters should be augmented with stochastic structure to capture model error?
- Initially, we base the decision on GSA (heuristic)
- Implementing formal model comparison via Bayes Factor

Bayes' formula for a given model M_k

$$\overbrace{p(\tilde{\alpha}|y, M_k)}^{\text{Posterior}} = \frac{\overbrace{p(y|\tilde{\alpha}, M_k)}^{\text{Likelihood}} \overbrace{p(\tilde{\alpha}|M_k)}^{\text{Prior}}}{\underbrace{p(y|M_k)}_{\text{Evidence}}}$$

Bayes factor between two models is the ratio of two evidence terms:

$$\text{BF}(M_1, M_2) = \frac{p(y|M_1)}{p(y|M_2)}$$

Computing log-evidence $\log p(y|M_k)$ is key for model selection.

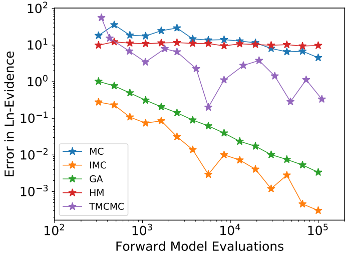
Model Selection: Model Evidence Computation

- Model evidence is a high-dimensional integral, requiring many model evaluations – challenging to compute
- We investigated five methods
  - GA (Gaussian approximation to posterior)
  - HM (Harmonic Mean estimator)
  - MC (Plain Monte-Carlo)
  - IMC (Importance sampling Monte-Carlo)
  - TMCMC (Transitional Markov chain Monte-Carlo)

Model Selection: Model Evidence Computation

- Model evidence is a high-dimensional integral, requiring many model evaluations – challenging to compute
- We investigated five methods
  - GA (Gaussian approximation to posterior)
  - HM (Harmonic Mean estimator)
  - MC (Plain Monte-Carlo)
  - IMC (Importance sampling Monte-Carlo)
  - TMCMC (Transitional Markov chain Monte-Carlo)

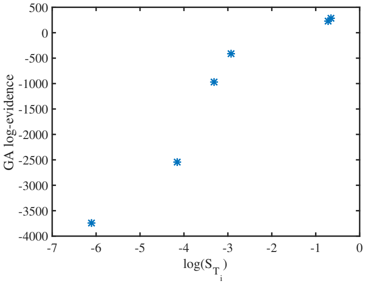
Param	GSA $\bar{S}_{T_i}$	GA
C_R	5.24×10^{-1}	2.82×10^{2}
Pr_t^{-1}	1.58×10^{-2}	-2.55×10^{3}
Sc_t^{-1}	4.90×10^{-1}	2.30×10^{2}
$I_i = u_i'/U_i$	3.63×10^{-2}	-9.68×10^{2}
$I_r = v'/u'$	2.24×10^{-3}	-3.74×10^{3}
L_i	5.32×10^{-2}	-4.15×10^{2}
C_R, Sc_t^{-1}		2.79×10^{2}



Model Selection: Model Evidence Computation

- Model evidence is a high-dimensional integral, requiring many model evaluations – challenging to compute
- We investigated five methods
  - GA (Gaussian approximation to posterior)
  - HM (Harmonic Mean estimator)
  - MC (Plain Monte-Carlo)
  - IMC (Importance sampling Monte-Carlo)
  - TMCMC (Transitional Markov chain Monte-Carlo)

Param	GSA $\bar{S}_{T_i}$	GA
C_R	5.24×10^{-1}	2.82×10^{2}
Pr_t^{-1}	1.58×10^{-2}	-2.55×10^{3}
Sc_t^{-1}	4.90×10^{-1}	2.30×10^{2}
$I_i = u_i'/U_i$	3.63×10^{-2}	-9.68×10^{2}
$I_r = v'/u'$	2.24×10^{-3}	-3.74×10^{3}
L_i	5.32×10^{-2}	-4.15×10^{2}
C_R, Sc_t^{-1}		2.79×10^{2}



Embedded Model: Predictions

$$f(x;\Lambda) = f\left(x; \textstyle\sum_k \alpha_k \Psi_k(\xi_{1:d})\right) = \textstyle\sum_k f_k(x;\tilde{\alpha})\Psi_k(\xi_{1:d})$$

- Non-intrusive spectral projection (NISP) will allow
  - Posterior/pushed-forward predictions
  - Easy access to first two moments:

$$\mu(x;\tilde{\alpha}) = f_0(x;\tilde{\alpha}), \qquad \sigma^2(x;\tilde{\alpha}) = \sum_{k>0} f_k^2(x;\tilde{\alpha})||\Psi_k||^2$$

- Predictive mean

$$\mathbb{E}[y(x)] = \mathbb{E}_{\tilde{\alpha}}[\mu(x;\tilde{\alpha})]$$

- Decomposition of predictive variance

$$\mathbb{V}[y(x)] = \underbrace{\mathbb{E}_{\tilde{\alpha}}[\sigma^2(x;\tilde{\alpha})]}_{\text{Model error}} + \underbrace{\mathbb{V}_{\tilde{\alpha}}[\mu(x;\tilde{\alpha})]}_{\text{Posterior error}}$$

Embedded Model: Predictions at Data Locations

$$f(x_i;\Lambda) = f\left(x_i;\sum_k \alpha_k \Psi_k(\xi_{1:d})\right) + \sigma\xi_{i+d} = \sum_k f_k(x_i;\tilde{\alpha})\Psi_k(\xi_{1:d}) + \sigma\xi_{i+d}$$

- Non-intrusive spectral projection (NISP) will allow
  - Likelihood computation
  - Easy access to first two moments:

$$\mu(x_i;\tilde{\alpha}) = f_0(x_i;\tilde{\alpha}), \qquad \sigma^2(x_i;\tilde{\alpha}) = \sum_{k>0} f_k^2(x_i;\tilde{\alpha})||\Psi_k||^2$$

- Predictive mean

$$\mathbb{E}[y(x_i)] = \mathbb{E}_{\tilde{\alpha}}[\mu(x_i;\tilde{\alpha})]$$

- Decomposition of predictive variance

$$\mathbb{V}[y(x_i)] = \underbrace{\mathbb{E}_{\tilde{\alpha}}[\sigma^2(x_i;\tilde{\alpha})]}_{\text{Model error}} + \underbrace{\mathbb{V}_{\tilde{\alpha}}[\mu(x_i;\tilde{\alpha})] + \sigma^2}_{\text{Posterior/Data error}}$$

Two common embedding forms

$$y_i = f(x_i; \Lambda = \lambda + \delta_\alpha) + \epsilon_i$$

- Unconstrained inputs:
  - First-order Gauss-Hermite PC (Multivariate Normal):

$$\begin{cases} \Lambda_1 = \lambda_1 + \alpha_{11}\xi_1 \\ \Lambda_2 = \lambda_2 + \alpha_{21}\xi_1 + \alpha_{22}\xi_2 \\ \quad \vdots \\ \Lambda_d = \lambda_d + \alpha_{d1}\xi_1 + \alpha_{d2}\xi_2 + \cdots + \alpha_{dd}\xi_d \end{cases}$$

- Constrained inputs:
  - First-order Legendre-Uniform PC (Independent Uniform):

$$\begin{cases} \Lambda_1 = \lambda_1 + \alpha_1\xi_1 \\ \Lambda_2 = \lambda_2 + \alpha_2\xi_2 \\ \quad \vdots \\ \Lambda_d = \lambda_d + \alpha_d\xi_d \end{cases}$$

Surrogate construction is necessary

Remember output PC construction

$$z_k = \frac{1}{||\Psi_k||^2} \int f(\Lambda(\xi))\Psi_k(\xi)\pi_\xi(\xi)d\xi \approx \frac{1}{||\Psi_k||^2} \sum_q f(\Lambda(\xi^{(q)}))\Psi_k(\xi^{(q)})w^{(q)}$$

requires multiple model evaluations, hence...

- We pre-construct a surrogate or a response surface to $f(\Lambda)$ via standard polynomial regression
- Subsequent NISP can be made exact if the bases of surrogate and PC match
- Access to leave-one-out (LOO) surrogate error as yet another component of the predictive uncertainty

Attribution of error components

$$y_i = \underbrace{\sum_k f_{ik}(\alpha)\Psi_k(\xi_1,\ldots,\xi_d) + \sigma_{\mathcal{D}}\xi_{d+i}}_{h_i(\hat{\xi};\hat{\alpha})}$$

Stochastic dimensions:

- Model error $\xi_1,\ldots,\xi_d$
- Measurement error $\xi_{d+1},\ldots,\xi_{d+N}$
- Posterior uncertainty (α): can be represented via its own PC expansion (using MCMC samples and Rosenblatt transformation)

Full PC expansion: $y_i = \sum f_j \Psi_j(\hat{\hat{\xi}})$

Full stochastic *germ*:

$$\hat{\hat{\xi}} = (\underbrace{\xi_1,\ldots,\xi_d}_{\text{Model error}}, \underbrace{\xi_{d+1},\ldots,\xi_{d+N}}_{\text{Measurement error}}, \underbrace{\xi_{d+N+1},\ldots,\xi_{d+N+N_\alpha}}_{\text{Posterior uncertainty}})$$

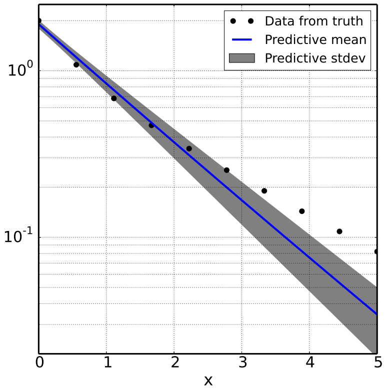
Posterior predictive variance:

$$\sigma^2_{\text{PP}}(x_i) = \mathbb{E}_\alpha[\sigma^2(x_i,\alpha)] + \mathbb{E}_{\sigma_{\mathcal{D}}}[\sigma^2_{\mathcal{D}}] + \mathbb{V}_\alpha[\mu(x_i,\alpha)]$$

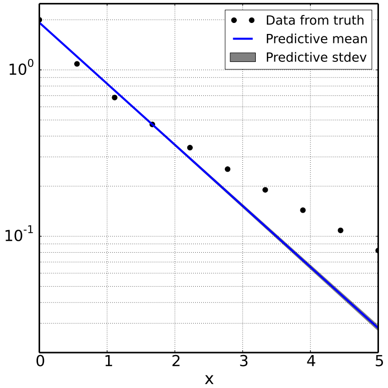
Predictions account for model error

Calibrating single-exponential models
with data from a double exponential model $g(x) = e^{-0.5x} + e^{-2x}$

Linear-exponential $f(x, \lambda) = e^{\lambda_1 + \lambda_2 x}$



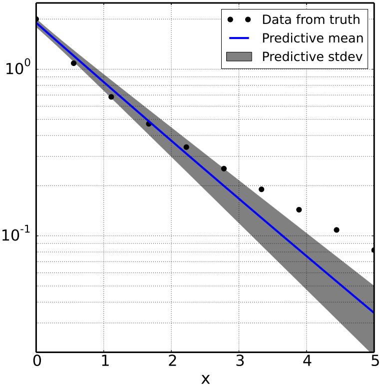
Additive Gaussian error



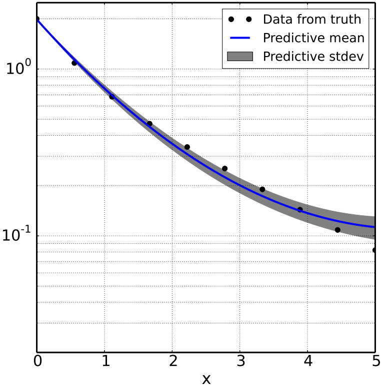
Predictions account for model error

Calibrating single-exponential models
with data from a double exponential model $g(x) = e^{-0.5x} + e^{-2x}$

Linear-exponential $f(x, \lambda) = e^{\lambda_1 + \lambda_2 x}$



Quadratic-exponential $f_2(x, \lambda) = e^{\lambda_1 + \lambda_2 x + \lambda_3 x^2}$



Key Steps

- **Formulation:** Identify a pair of models with different degree of fidelity
  - e.g., low-vs-high grid resolution, simplified-vs-detailed geometry, or data-vs-model.
- **Representation:** Embed model error a few parameters at a time
  - Build surrogate, perform GSA for initial screening
- **Quantification:** Calibrate for embedded PC coefficients
  - Challenging Bayesian formulation: adaptive MCMC sampling.
- **Prediction:** Embedded model error propagation via PC NISP
  - Posterior predictive checks
- **Attribution:** Attribute model errors to specific components
  - Variance-based decomposition into contributions from model error, surrogate error, data noise, posterior uncertainty.

Treatment of Discrete or Categorical Parameters

- We have developed an approach to incorporate discrete parameters in the embedded model error framework.
- Augment discrete parameters with a probability mass function (PMF) and infer the mass weights (just like the continuous case of inferring PDF).
- Allows MCMC on continuous parameters.
- Connections to Bayesian model averaging and model selection.

The overall mean for a given (α, a, x) is

$$\mu(\alpha, a; x) = \mathbb{E}_{\Lambda, L}\left[f(\Lambda(\alpha), L(a); x)\right] = \sum_{r=1}^{R} a_r \mu_r(\alpha; x),$$

and the variance is

$$\begin{aligned} \sigma^2(\alpha, a; x) &= \mathbb{V}_{\Lambda, L}\left[f(\Lambda(\alpha), L(a); x)\right] \\ &= \underbrace{\sum_{r=1}^{R} a_r \sigma_r^2(\alpha; x)}_{\text{due to cont. param.}} + \underbrace{\sum_{r=1}^{R} a_r \mu_r^2(\alpha; x) - \mu(\alpha, a; x)^2}_{\text{due to categorical param.}}. \end{aligned}$$