Density Estimation Framework for Model Error Assessment

K. Sargsyan, H. Najm, Z. Liu, C. Safta,
B. van Bloemen Waanders, H. Michelsen, R. Bambha

Sandia National Laboratories
Livermore, CA
Albuquerque, NM

AGU Fall Meeting
San Francisco
December 15-19, 2014

Sandia National Laboratories is a multi-program laboratory operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000

Density Estimation Framework for Model Error Assessment

K. Sargsyan, H. Najm, Z. Liu, C. Safta,
B. van Bloemen Waanders, H. Michelsen, R. Bambha

Sandia National Laboratories
Livermore, CA
Albuquerque, NM

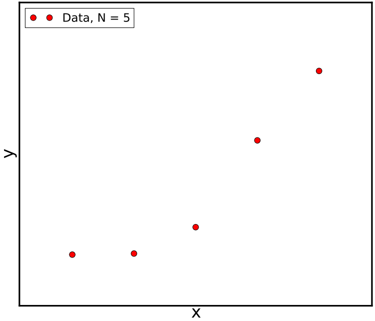
AGU Fall Meeting
San Francisco
December 15-19, 2014

- DOE Office of Advanced Scientific Computing Research (ASCR), Scientific Discovery through Advanced Computing (SciDAC)
- Sandia National Labs Laboratory Directed Research and Development (LDRD) Program
- DOE Office of Basic Energy Sciences (BES), Div. of Chem. Sci., Geosci., & Biosci.

Motivation

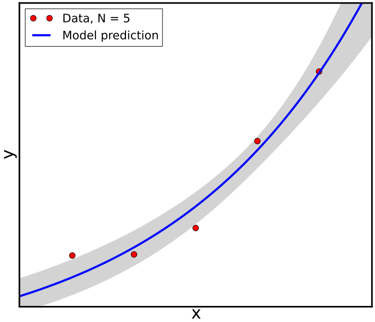
- All models are wrong, but some are useful *[George Box]*
- Models of physical systems rely on
  - Presumed theoretical framework
  - Mathematical formulation
  - Simplifying assumptions, parameterizations
  - Numerical discretization of governing equations
  - Computational software & hardware
- Model error is frequently non-negligible
- Estimating model error is useful for
  - Model validation
  - Model comparison
  - Scientific discovery and model improvement
  - Reliable computational predictions

Motivation



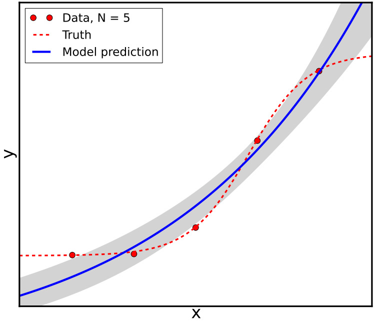
Model-data fit

Motivation

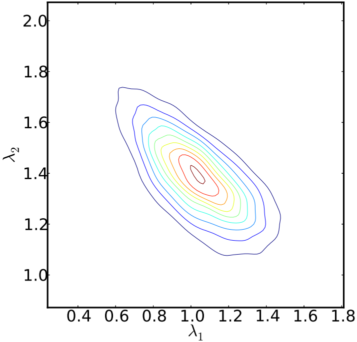


Model-data fit

Motivation

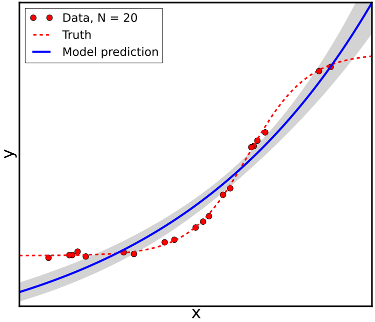


Model-data fit

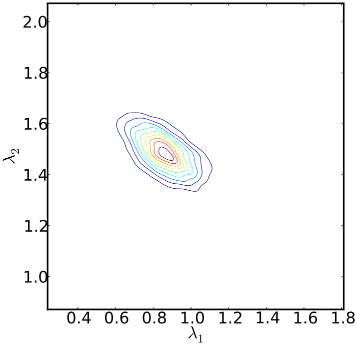


Calibrated parameters

Motivation

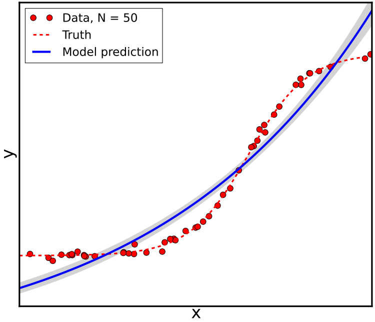


Model-data fit

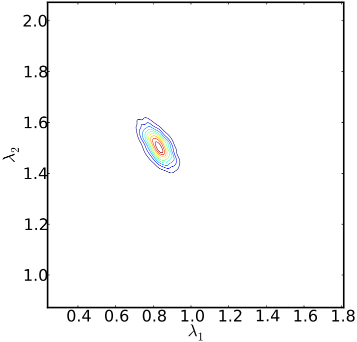


Calibrated parameters

Motivation



Model-data fit



Calibrated parameters

- If the model has structural errors, more data does not help!
- We target model-vs-truth discrepancy

State-of-the-art: Issues for physical models

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

- Explicit additive statistical model for model error $\delta(x)$ Kennedy-O'Hagan (2001).
- Calibrated predictive model $y_{mod}(x) = f(x; \lambda) + \delta(x)$

- Potential violation of physical constraints
  - *e.g.* incompressible flow: $\nabla \bullet v = 0$
- Disambiguation of model error $\delta(x_i)$ and data error ϵ_i^{d}
- Calibration of model error on measured observable does not impact the quality of model predictions on other QoIs

Model error embedding: key idea

- Ideally, modelers want predictive *errorbars*:
  inserting randomness on the outputs has issues, so...
- Cast input parameters λ as a random variable Λ
  Black-box

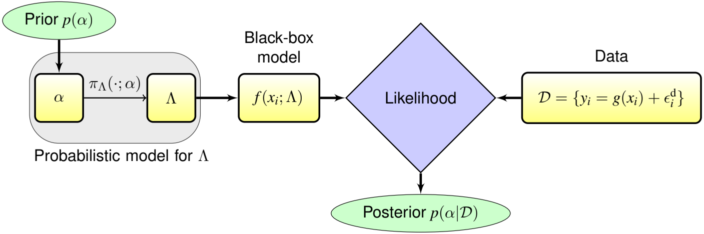
$$y_i = f(x_i; \Lambda) + \epsilon_i^{\mathrm{d}}$$

- More generally, explore additional parameterizations,
  Extra 'physics'

$$y_i = \tilde{f}(x_i; \lambda, \Theta) + \epsilon_i^{\mathrm{d}}$$

- Calibration turns into density estimation
  - Object of inference is PDF of Λ, not parameter λ
- Back to calibration: parameterize $\pi_\Lambda(\cdot; \alpha)$ and calibrate for α
  - E.g. Multivariate Normal, or Polynomial Chaos

Model error embedding: features



- Embed model error in specific submodel phenomenology
    - a modified transport or constitutive law
    - a modified formulation for a material property
- 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

Model error embedding: Bayesian formulation

- Consider the simplest setting with no data noise, *i.e.* $\epsilon_i^{\mathrm{d}} = 0$.
- In the simplest setting, cast λ as a random variable Λ
  Black-box $$y_i = f(x_i; \Lambda)$$
- Calibration turns into density estimation for the PDF of Λ
- Polynomial Chaos parameterization for $\Lambda = \sum_{k=0}^{K} \alpha_k \Psi_k(\xi)$
- Back to parameter estimation, now for $\alpha = (\alpha_0, \ldots, \alpha_K)$

- Bayesian setting $$\underbrace{p(\alpha|\mathcal{D})}_{\text{Posterior}} \propto \underbrace{L_{\mathcal{D}}(\alpha)}_{\text{Likelihood}} \underbrace{p(\alpha)}_{\text{Prior}}$$
  - Full Likelihood $L_{\mathcal{D}}(\alpha) = p(\mathcal{D}|\alpha) = p(y_1, \ldots, y_N|\alpha)$
  - Marginalized Likelihood $L_{\mathcal{D}}(\alpha) \approx \prod_{i=1}^{N} p(y_i|\alpha)$
  - ABC Likelihood, see next

ABC Likelihood

With $\rho(\mathcal{S})$ being a metric of the statistic $\mathcal{S}$, use the kernel function as an ABC likelihood:

$$L_{\mathrm{ABC}}(\alpha) = \frac{1}{\epsilon} K\left(\frac{\rho(\mathcal{S})}{\epsilon}\right)$$

where ϵ is a ‘tolerance’ parameter. Require

- Uncertain prediction $p(y|D)$ is centered on the data
- The width of the distribution $p(y|D)$ is consistent with the spread of the data around the nominal model prediction

Propose the Gaussian kernel density:

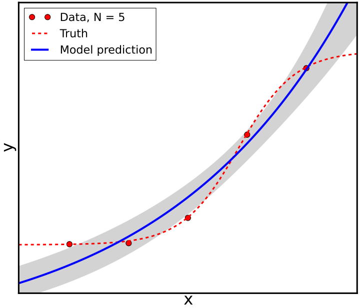
$$L_{\mathrm{ABC}}(\alpha) = \frac{1}{\epsilon\sqrt{2\pi}} \prod_{i=1}^{N} \exp\left(-\frac{(\mu_i(\alpha) - y_i)^2 + (\sigma_i(\alpha) - \gamma|\mu_i(\alpha) - y_i|)^2}{2\epsilon^2}\right)$$

Predictive moments: $\mu_i(\alpha) = \mathrm{E}_\Lambda[f(x_i;\Lambda)]$ $\sigma_i^2(\alpha) = \mathrm{V}_\Lambda[f(x_i,\Lambda)]$

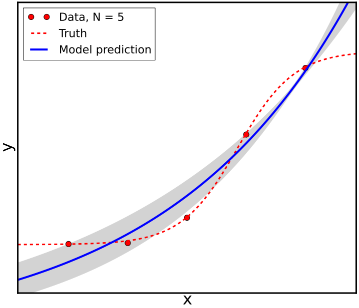
Predictions account for model error: example 1

Calibrating an exponential model $f(x;\lambda_1,\lambda_2) = \lambda_2 e^{\lambda_1 x} - 2$
with data from a hyperbolic tangent model $g(x) = \tanh(3(x-0.3))$

Additive Gaussian error



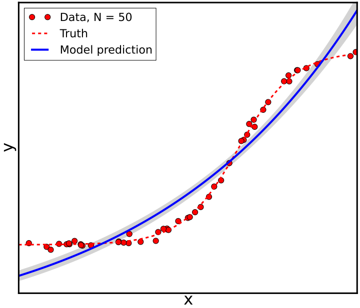
Embedded model error



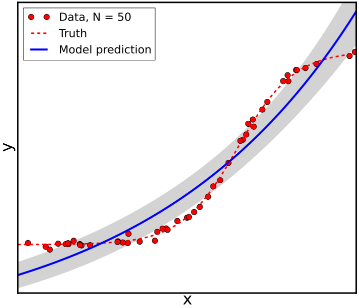
Predictions account for model error: example 1

Calibrating an exponential model $f(x;\lambda_1,\lambda_2) = \lambda_2 e^{\lambda_1 x} - 2$
with data from a hyperbolic tangent model $g(x) = \tanh(3(x-0.3))$

Additive Gaussian error



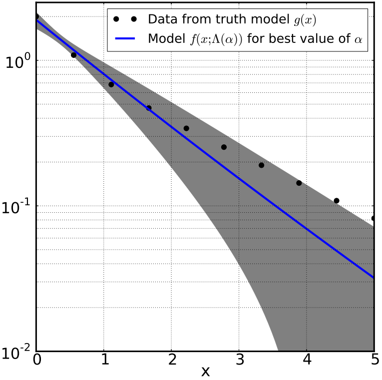
Embedded model error



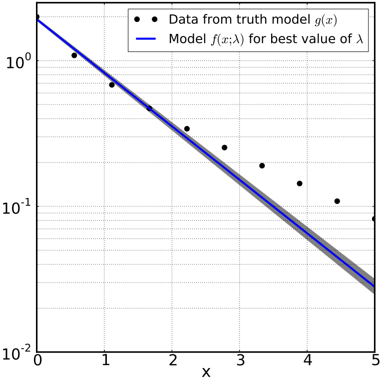
Predictions account for model error: example 2

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

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



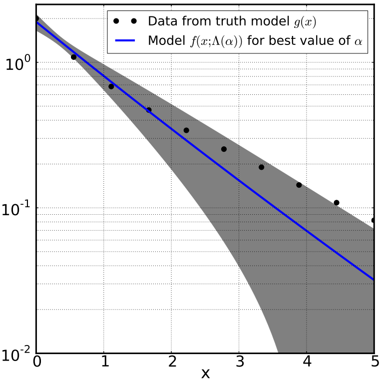
Additive Gaussian error



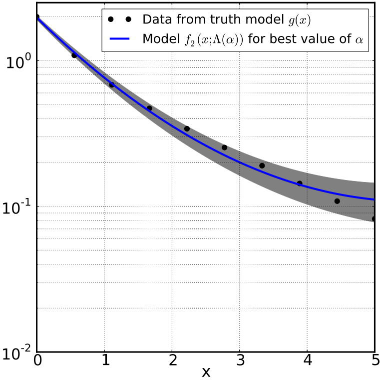
Predictions account for model error: example 2

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



TransCom3 Experiment of CO_2 Flux Inversion

[Gurney *et al.*, Tellus B, 2003]

- Observations $\mathbf{d}$ at $N = 77$ sites around the world
- Inverse problem: find fluxes $\mathbf{s}$ at $M = 22$ locations
- Linearized 'response' model $\mathbf{R}$, such that $\mathbf{d} \approx \mathbf{R}\mathbf{s}$

$$\mathbf{d} = \mathbf{R}\mathbf{s} + \epsilon_{\mathbf{d}}$$

- Model $\mathbf{R}$ is never perfect thus contaminating the inversion
- The inferred values of $\mathbf{s}$ compensate for model deficiencies
- $\epsilon_{\mathbf{d}}$ is meant to capture data errors, but is 'entangled' with model errors

TransCom3 Experiment of CO_2 Flux Inversion

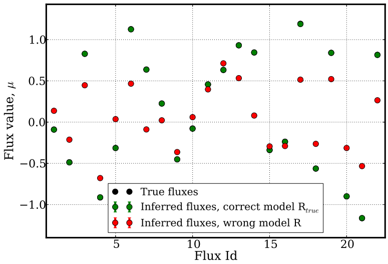
[Gurney *et al.*, Tellus B, 2003]

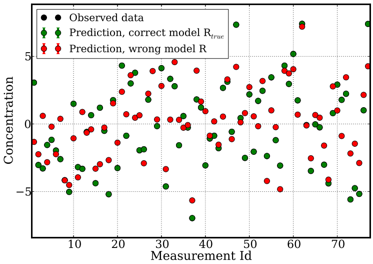
- Synthetic study: assume no measurement error
- Generate data from a true model $\mathbf{R}_{\text{true}}$ with exact flux values:

$$\mathbf{d} = \mathbf{R}_{\text{true}}\mathbf{s}_{\text{exact}}$$

- Infer $\mathbf{s}$
  - with the true model $\mathbf{d} \approx \mathbf{R}_{\text{true}}\mathbf{s}$, or
  - with a wrong model $\mathbf{d} \approx \mathbf{R}\mathbf{s}$

Additive Gaussian error





TransCom3 Experiment of CO_2 Flux Inversion

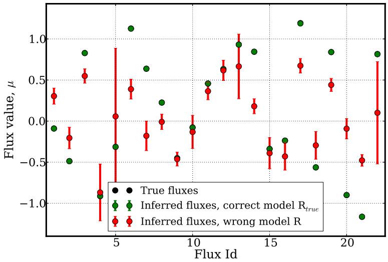
[Gurney *et al.*, Tellus B, 2003]

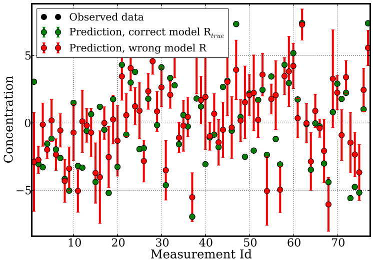
- Synthetic study: assume no measurement error
- Generate data from a true model $\mathbf{R}_{\text{true}}$ with exact flux values:

$$\mathbf{d} = \mathbf{R}_{\text{true}}\mathbf{s}_{\text{exact}}$$

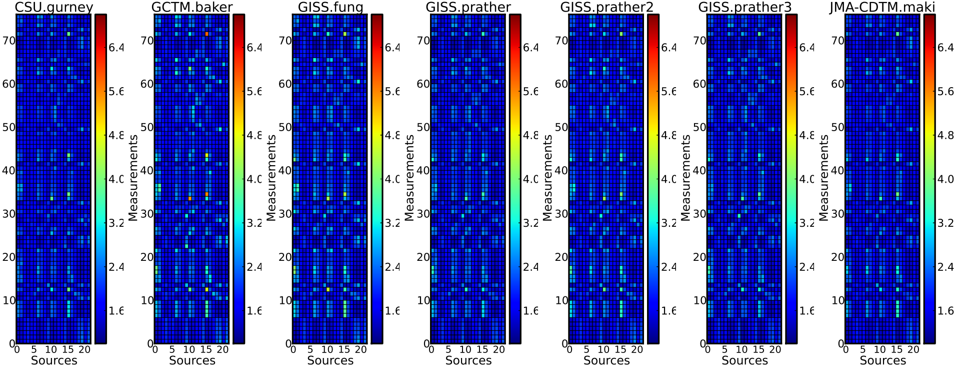
- Infer $\mathbf{s}$
  - with the true model $\mathbf{d} \approx \mathbf{R}_{\text{true}}\mathbf{s}$, or
  - with a wrong model $\mathbf{d} \approx \mathbf{R}\mathbf{s}$

Embedded model error





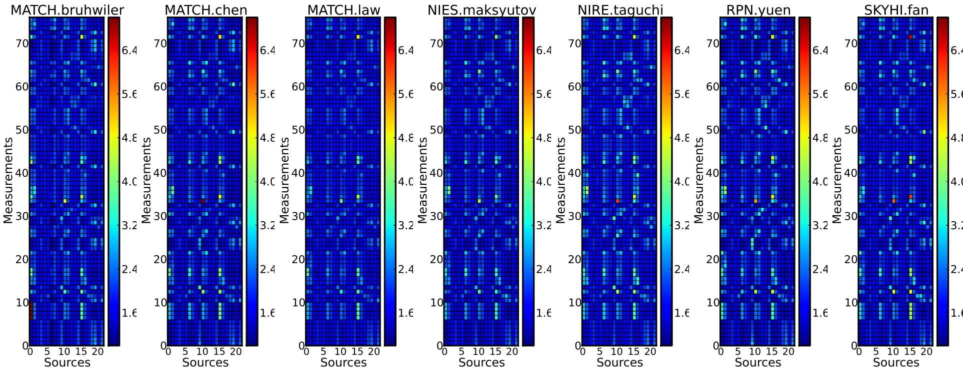
Consider 14 different response models $\mathbf{R}$



Infer fluxes $\mathbf{s}$, given measurements $\mathbf{d}$ to satisfy $\mathbf{d} \approx \mathbf{Rs}$

- Conventional additive Gaussian error (least-squares): $\mathbf{d} = \mathbf{Rs} + \xi$
- Embed probabilistic model for fluxes s: $\mathbf{d} = \mathbf{R}(\mu_\mathbf{s} + \mathbf{C_s}\xi)$

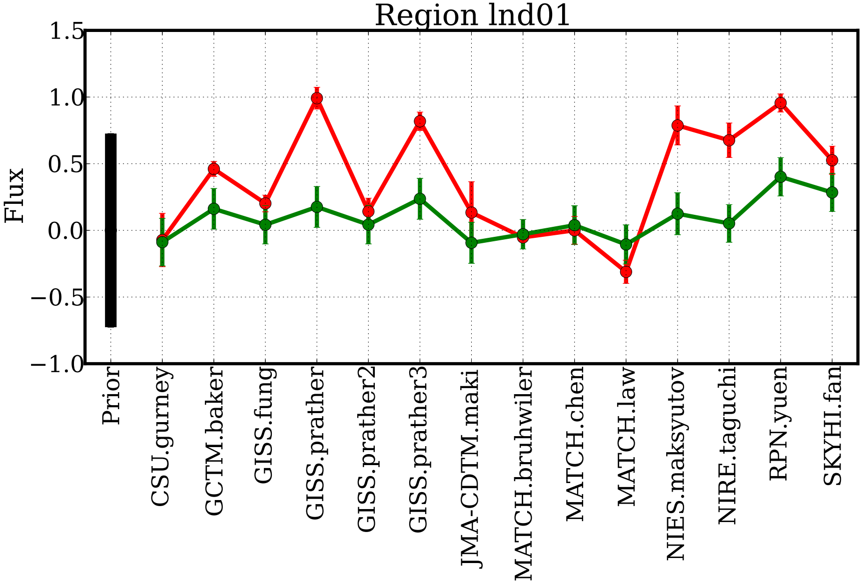
Consider 14 different response models $\mathbf{R}$



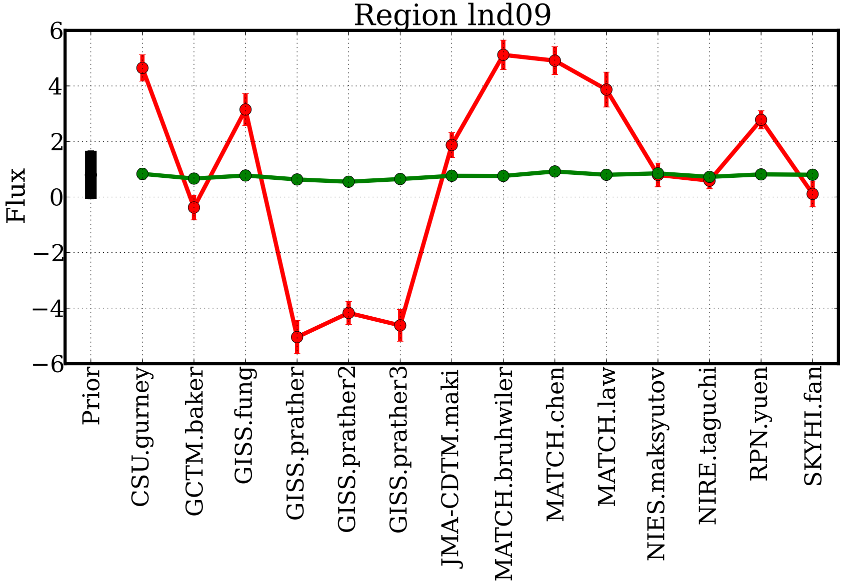
Infer fluxes $\mathbf{s}$, given measurements $\mathbf{d}$ to satisfy $\mathbf{d} \approx \mathbf{R}\mathbf{s}$

- Conventional additive Gaussian error (least-squares): $\mathbf{d} = \mathbf{R}\mathbf{s} + \xi$
- Embed probabilistic model for fluxes s: $\mathbf{d} = \mathbf{R}(\mu_\mathbf{s} + \mathbf{C}_\mathbf{s}\xi)$

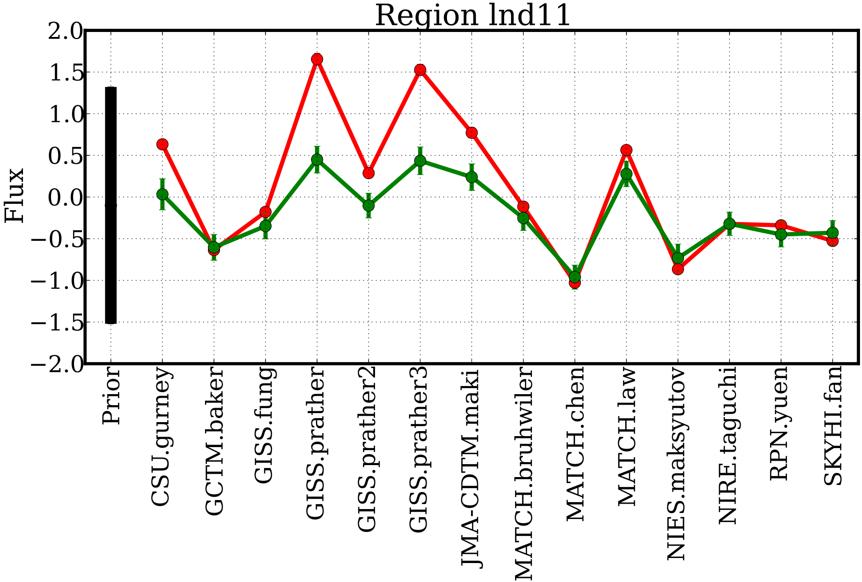
Inferred fluxes show less variability across models



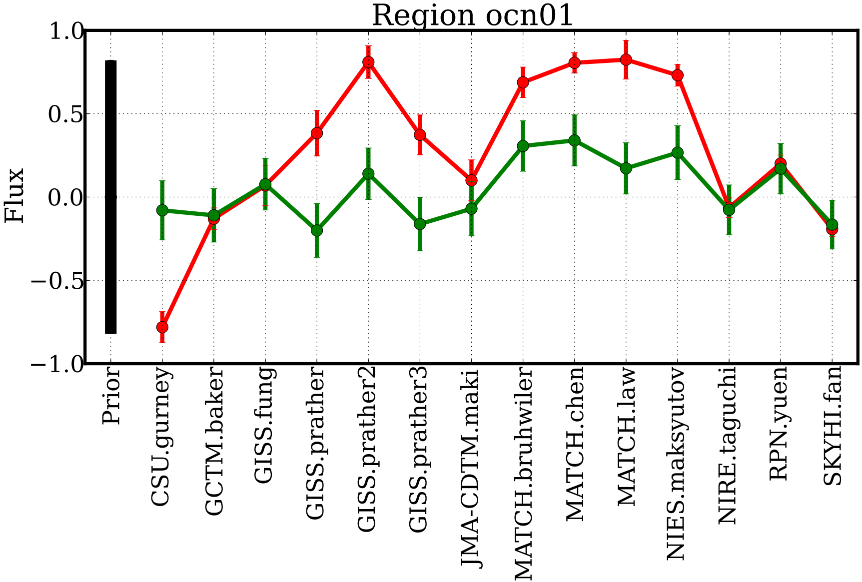
Inferred fluxes show less variability across models



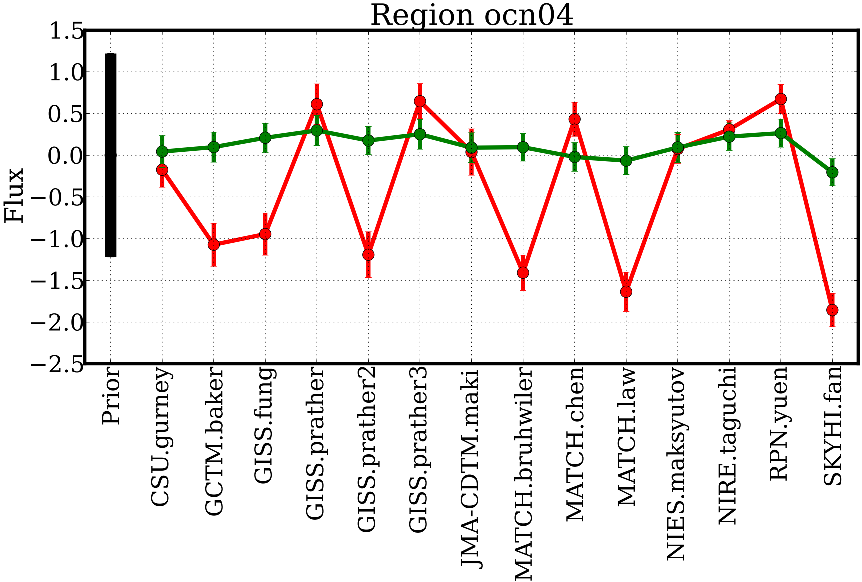
Inferred fluxes show less variability across models



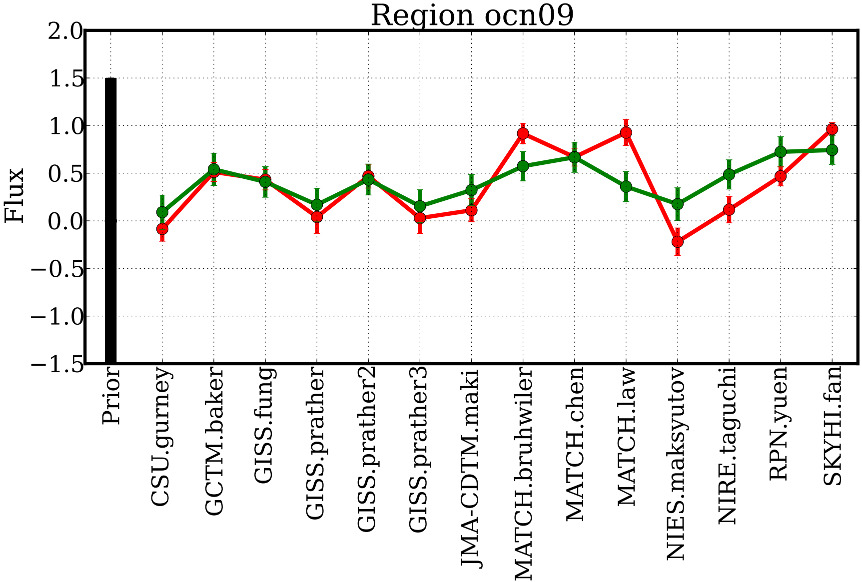
Inferred fluxes show less variability across models



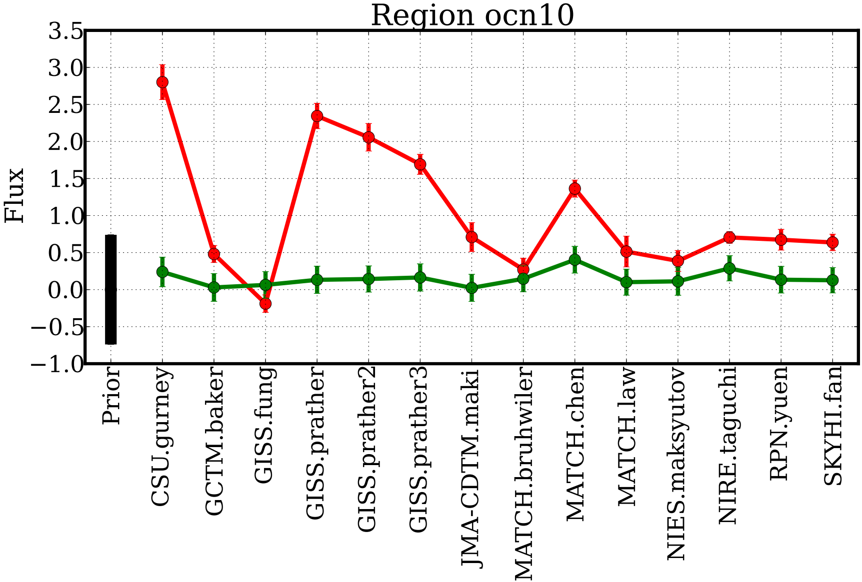
Inferred fluxes show less variability across models



Inferred fluxes show less variability across models



Inferred fluxes show less variability across models



Summary

- A method for dealing with model discrepancy error that is targeted at physical models
- Reformulate the calibration as a density estimation problem
- Bayesian machinery to find parameters of the PDFs
- Approximate Bayesian Computation (ABC) targets constraints of interest to the modeler
- Model-to-model calibration
- (in progress) Extension to include data noise

- K. Sargsyan, H. Najm, and R. Ghanem. "On the Statistical Calibration of Physical Models". *International Journal for Chemical Kinetics*, in review.

Thank You

Model error embedding: Bayesian formulation

- Consider the simplest setting with no data noise, *i.e.* $\epsilon_i^{\mathrm{d}} = 0$.
- In the simplest setting, cast λ as a random variable Λ
  Black-box
  $$y_i = f(x_i; \Lambda)$$
- Calibration turns into density estimation for the PDF of Λ
- Polynomial Chaos parameterization for $\Lambda = \sum_{k=0}^{K} \alpha_k \Psi_k(\xi)$
- Back to parameter estimation, now for $\alpha = (\alpha_0, \ldots, \alpha_K)$
- Bayesian setting
  $$\underbrace{p(\alpha|\mathcal{D})}_{\text{Posterior}} \propto \underbrace{L_{\mathcal{D}}(\alpha)}_{\text{Likelihood}} \underbrace{p(\alpha)}_{\text{Prior}}$$
  - Full Likelihood $L_{\mathcal{D}}(\alpha) = p(\mathcal{D}|\alpha) = p(y_1, \ldots, y_N|\alpha)$
  - Marginalized Likelihood $L_{\mathcal{D}}(\alpha) \approx \prod_{i=1}^{N} p(y_i|\alpha)$
  - ABC Likelihood, see next

Data-Model-Truth

- Measurements

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

- Model

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

- Total error budget

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

Statistical modeling of errors in calibrating $f(x;\lambda)$

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

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

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