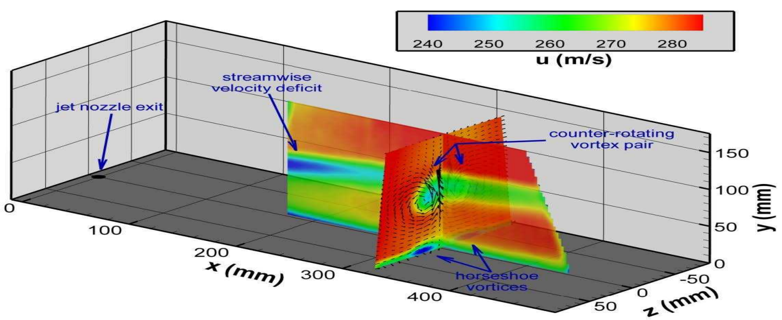
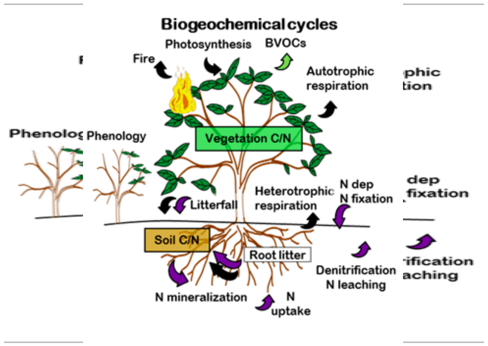
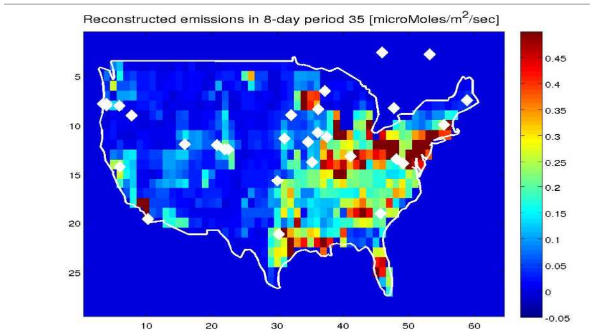
SAND2015-2931PE

Exceptional service in the national interest







Bayesian inference of parameters of computationally expensive models

Jaideep Ray

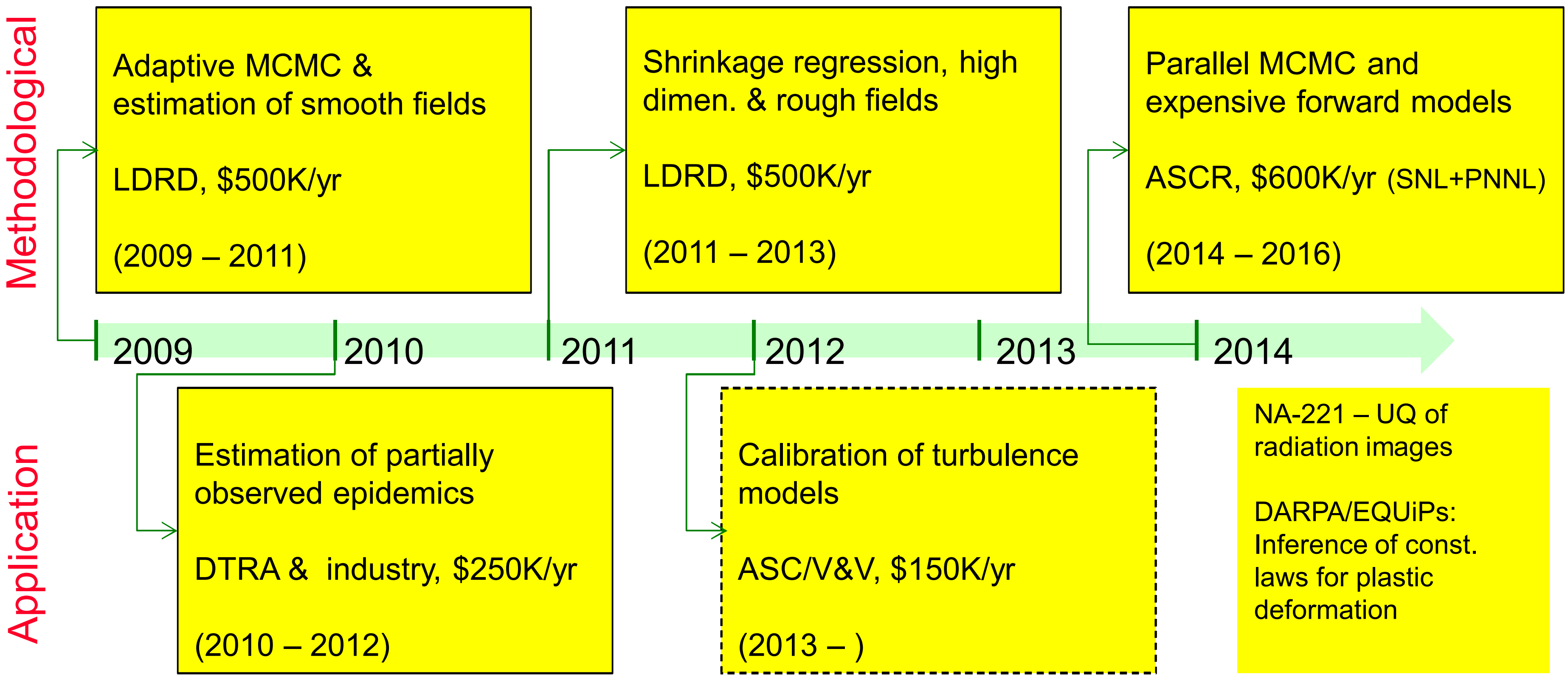
Introduction

- I develop methods to infer model parameters & inputs from experimental observations (inverse problems)
- Two issues: (1) inference with uncertainty (2) dimensionality reduction
- Funded by : SC/ASCR, LDRD (methods); ASC/V&V, DTRA (applications)
- Collaborators: 8900, 1400 (Comp & Info Sciences), Prof. Michalak (Carnegie Inst., Stanford), Prof. Marzouk (MIT) [methods]; 1500 (Engg. Sciences) & private industry [applications]
- Acknowledgments:
- SNL: L. Swiler, S. Lefantzi, S. Arunajatesan, J. Lee
- Outside: Z. Hou & M. Huang (PNNL); V. Yadav (JPL)

Take-home message

- Calibration of models critical to predictive simulation
  - Due to limitations of data, parameters are best inferred as PDFs
  - Can be done via Bayesian inference; the “solver” in this case are called Markov chain Monte Carlo (MCMC) methods (sequential)
  - Expensive; engineering & scientific models not calibrated this way
- We are developing the technology to do so. 2 core advances:
  - Development of surrogate models for engineering models
  - Development of multi-chain MCMC methods (parallelization)
- Illustrate with calibration of Community Land Model (CLM)
  - Programmatic use: calibration of turbulence models, disease models, etc.
- End with what next, papers, software etc.

Timeline of development



Similar work

- Methodological
  - Ghattas & Thanh (UT Austin) : focus on better proposal via local estimate of posterior distribution
  - Vrugt (UC Irvine) : multi-chain, with Differential Evolution; very similar
  - Haario (Finland) : multi-chain adaptive Metropolis; very similar
- Applications
  - Moser/Oliver (UT Austin): calibration of turbulence models
  - Edeling & Bijl (Delft): turbulence models; boundary layers
  - Solonen, Jarvinen, Haario: calibration, atmosphere model

Model calibration

- Model calibration: Estimation of model parameters or inputs from observations. Fundamental to predictive simulations
  - Can be scalars (e.g., Young's modulus) or entire fields (CO2 fluxes, permeability fields etc.)
  - Model fitting : $\mathbf{Y}^{(obs)} = \mathcal{M}(\mathbf{p}) + \varepsilon$

$$\min_{\mathbf{p}} \quad \left\| \mathbf{Y}^{(obs)} - \mathcal{M}(\mathbf{p}) \right\|_2^2$$

  - Inferred quantities are uncertain - Solution: estimate **p** as a PDF
- Bayesian Inference: A method to construct the PDF
  - Requires a *prior* PDF $\pi(\mathbf{p})$ and an error model, $\varepsilon \sim N(0, \sigma^2)$
  - Produces a *posterior* PDF $P(\mathbf{p}, \sigma^2 \mid Y^{(obs)})$; Bayes' formula

$$P(\mathbf{p}, \sigma^2 \mid \mathbf{Y}^{(\mathbf{obs})}) \propto \mathcal{L}(\mathbf{Y}^{(\mathbf{obs})} \mid \mathbf{p}, \sigma^2)\, \pi(\mathbf{p});$$

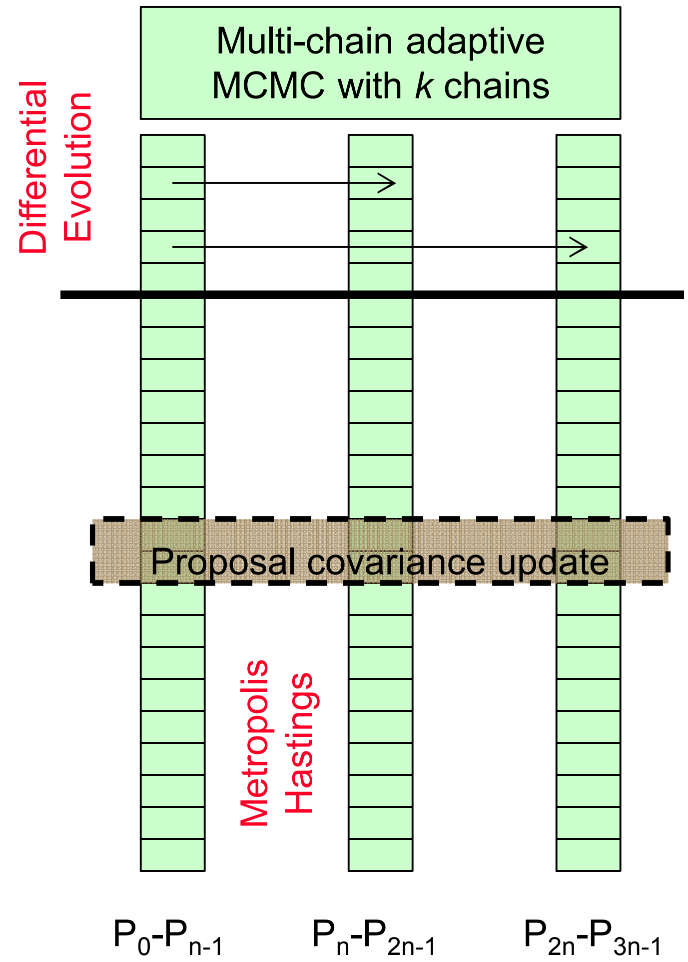
$$\mathcal{L}(:|:) \propto \exp\left(- \left\| Y^{(obs)} - \mathcal{M}(\mathbf{p}) \right\|_2^2 \Big/ 2\sigma^2 \right)$$

Markov chain Monte Carlo

- $P(\mathbf{p}, \sigma^2 \mid \mathbf{Y}^{(obs)})$ is "solved" by sampling from it
  - MCMC methods can sample from arbitrary PDFs efficiently
  - Each sample requires a run of $\mathcal{M}(\mathbf{p})$ – and samples are serial
    - Expensive!
- Ways out – our strategy
  - Replace $\mathcal{M}(\mathbf{p})$ with a statistical surrogate ("curve fit") $\mathcal{M}^{(s)}(\mathbf{p})$
    - $\mathcal{M}^{(s)}(\mathbf{p})$ is fast; using standard MCMC methods to solve inverse problem
    - Sophistication lies in making $\mathcal{M}^{(s)}(\mathbf{p})$; not always possible
    - Inferences include the effect of surrogate approximations
  - Multi-chain MCMC methods – MCMC carried out by communicating chains
    - Use original $\mathcal{M}(\mathbf{p})$, not surrogate – no surrogate approximation errors
    - Pool information collected by *N* chains to improve sampling efficiency
    - Distribute sampling burden over multiple chains

Core advances

- Surrogate models made by fitting to training data
  - Sample **p**-space; run $\mathbf{Y}=\mathcal{M}(\mathbf{p})$ for each $\mathbf{p}_i$
  - "Curve-fit" $\mathbf{Y}_i \cong \mathcal{M}^{(s)}(\mathbf{p}_i)$
  - $\mathcal{M}^{(s)}(\mathbf{p}_i)$: polynomials, GPs, neural nets, etc.
  - Overfitting: Cross-validation (CV), shrinkage, etc.
- Multi-chain adaptive MCMC w/ $\mathcal{M}(\mathbf{p})$
  - Initial exploration wasteful – fix using differential evolution (GA); asynchronous!
  - Adapt; propose $\mathbf{p}_{i+1}$ based on successful samples $\mathbf{p}_{i-k}$, k = 0 …. M
  - Have *N* chains pool samples; information sharing leads to better $\mathbf{p}_{i+1}$
  - Amortize sampling over *N* chains

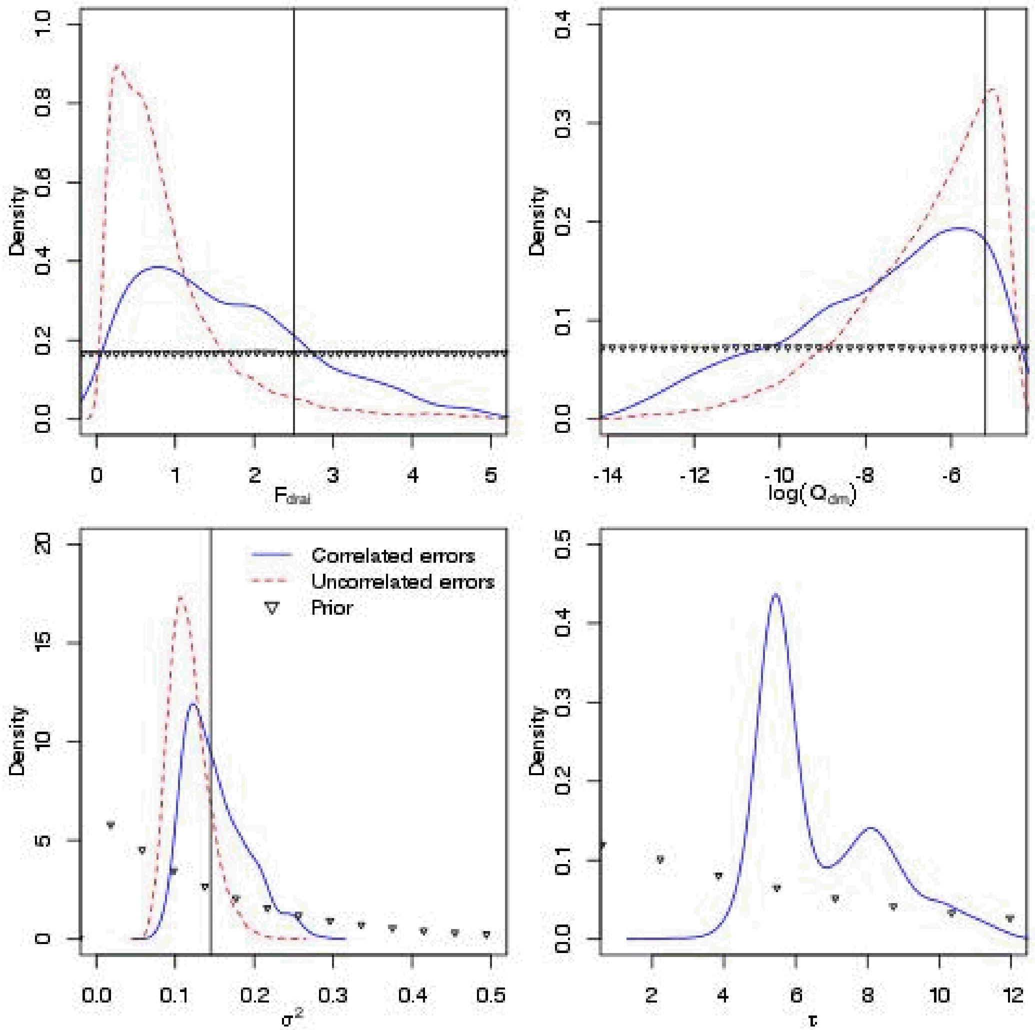


Bayesian calibration of the CLM

- Community Land Model (CLM) – for terrestrial, biogenic & hydrological processes
  - Used with atmosphere and ocean models in Earth System models
  - Can also be used with measured meteorology at a given site
- Default setting – reproduce observations @ global scale
  - Needs to be recalibrated when used in "site" mode; good illustration for Bayesian calibration of expensive models
- Calibration to monthly-averaged Latent Heat (LH) measurements from US-ARM
  - LH most affected by 3 hydrological parameters (F_{drai}, Q_{dm}, b)
- 2 calibrations:
  - Using 2003-2007 data, done with surrogates
  - Using synthetic data, done with CLM and 8-chain MCMC

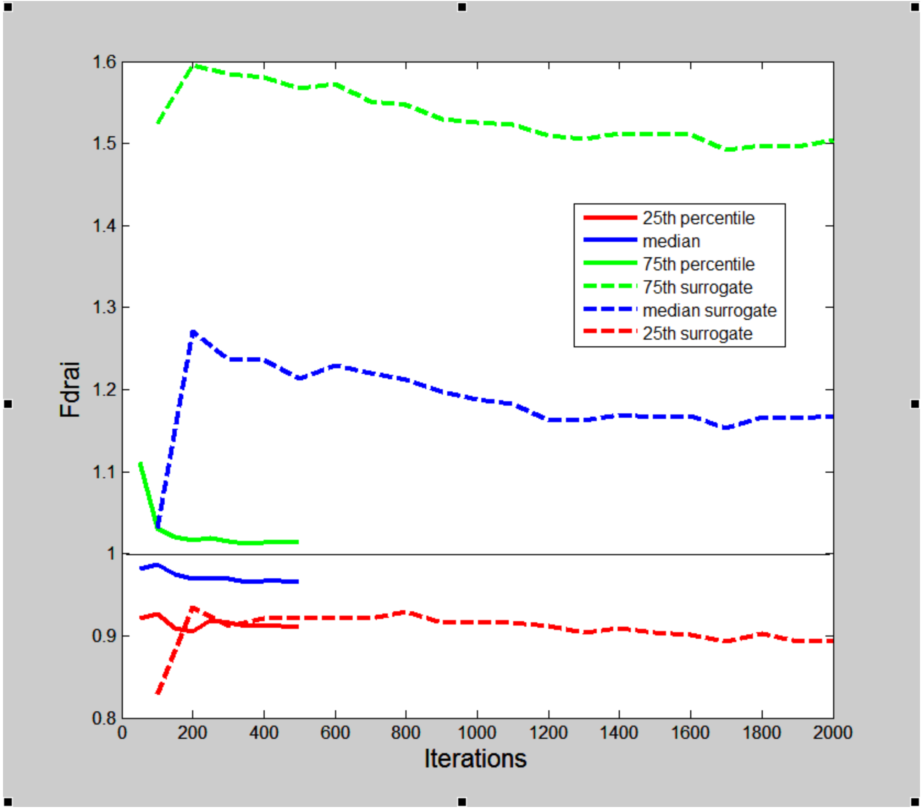
Calibration with surrogates

- Calibration performed using 2 different error models
  - To probe the causes of model – observation mismatch
  - Allows model selection
  - Showed that seasonal processes responsible for mismatch (τ ~ 5 months)
- Remember – done with surrogates
  - How much is the inference affected by surrogate model errors?



Calibration with multi-chain MCMC & CLM

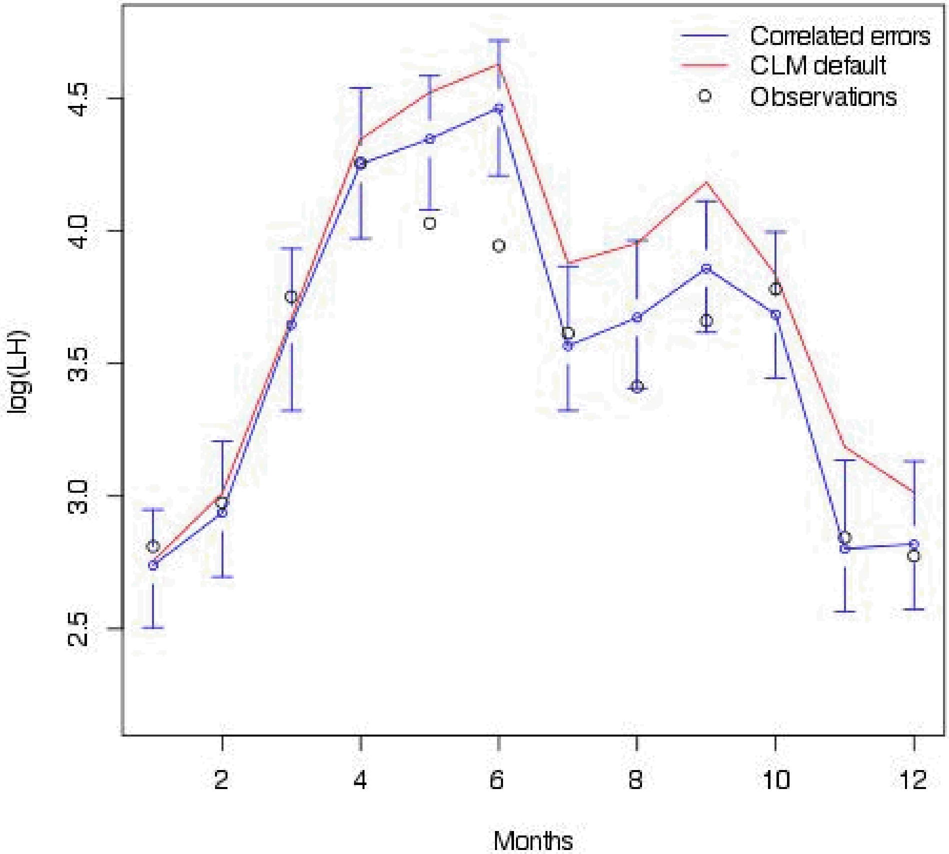
- Calibration with CLM provided as a check
  - Surrogate approximations affect the PDF of parameters
- Synthetic data problem
  - US-ARM meteorology for 2003
  - $\mathbf{Y}^{(obs)}$ generated using $F_{drai} = 1$
- Bayesian inference
  - Estimate F_{drai}; track its various quantiles for convergence
    - Should converge to 1
  - Redo with surrogates
- Chains: 500 (CLM); 2000 (surrogate)
- Inversion with surrogates leads to an over-estimate of uncertainty



Surrogates allow rigorous statistical checks, but surrogate errors can lead to excessive uncertainty

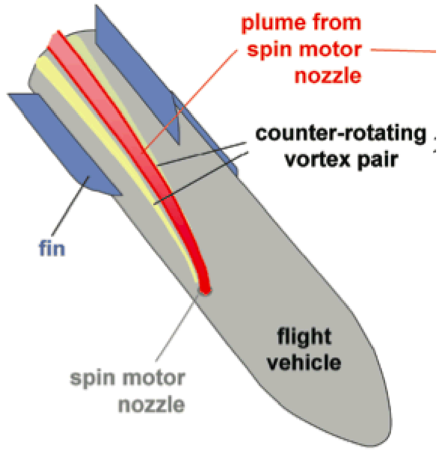
Calibration of CLM - Impact

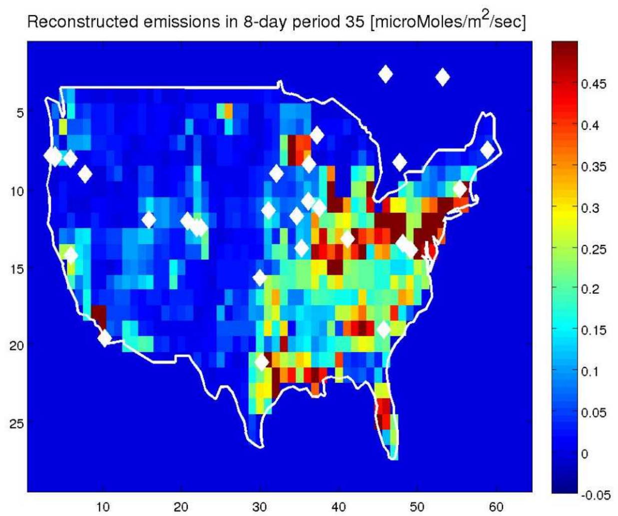
- First time a Bayesian CLM calibration had been done
  - Provably converged results for PDFs
  - PDFs for 2 sites show that parameter are site dependent
  - Proposed models for structural error, showed how to select them, and extracted info on causes of structural error
  - Showed the role of ensemble predictions
- Using real CLM versus surrogate, computed the effect of surrogate models on the PDFs



Applications & other work

- Estimation of turbulence model parameters [3 parameters]
  - Used for predictive RANS simulations of flow over slender aerodynamic bodies
  - Calibration to jet-in-crossflow experiments
  - Funds: ASC/V&V
- Estimation of rough fields using wavelet random field models [10^3 parameters]
  - Used in estimation of fossil-fuel CO_2 emissions
  - Specialized shrinkage methods for inverse problems with $O(10^3)$ dimensions
  - Funds: LDRD
- Bayesian characterization of outbreaks from limited data
  - Funds: DTRA & private industry





Summary & what next?

- We are developing scalable statistical inversion methods to calibrate/infer model inputs from data
  - Models: Engineering & scientific; PDE-based; expensive!
  - The methods also help uncover model-form errors
  - Funding: a mix of foundational & application-oriented sources
- What next?
  - Scale up multi-chain MCMC
  - Expand the use Bayesian inference to uncover model deficiencies
  - Expand the fields where I apply Bayesian inference
- SNL Team
  - Methods: L. Swiler, B. van Bloemen Waanders & S. Lefantzi
  - Applications: S. Arunajatesan & L. Dechant,

Papers and software

1. J. Ray et al, "Bayesian calibration of the Community Land Model using surrogates" accepted, *SIAM Journal on Uncertainty Quantification*, 2015.
2. J. Ray et al, "A sparse reconstruction method for the estimation of multiresolution emission fields via atmospheric inversion", under review, *Geoscientific Model Development*, 2015.
3. V. Yadav et al, "A statistical approach for isolation fossil-fuel emissions in atmospheric inverse problems", under review, *Journal of Geophysical Research – Atmosphere*
4. J. Ray et al, "A multiresolution spatial parameterization for the estimation of fossil-fuel carbon dioxide emissions via atmospheric inversions", *Geoscientific Model Development*, 7, 1901-1918, 2014.
5. J. Ray et al., "Bayesian reconstruction of binary media with unresolved fine-scale spatial structures" in *Advances in Water Resources* , 44:1--19, 2012.
6. S. A. McKenna et al, "Truncated multiGaussian fields and effective conductance of binary media", in *Advances in Water Resources* , 34:617-626, 2011.
7. J. Ray, Y. M. Marzouk, and H. N. Najm, "A Bayesian approach for estimating bioterror attacks from patient data", in *Statistics in Medicine*, 30(2):101-126, 2011
8. J. Ray & J. Lee, **sparse-msrf**: A package for sparse modeling and estimation of fossil-fuel CO2 emission fields, distributed under BSD license

Estimation of k-ε parameters

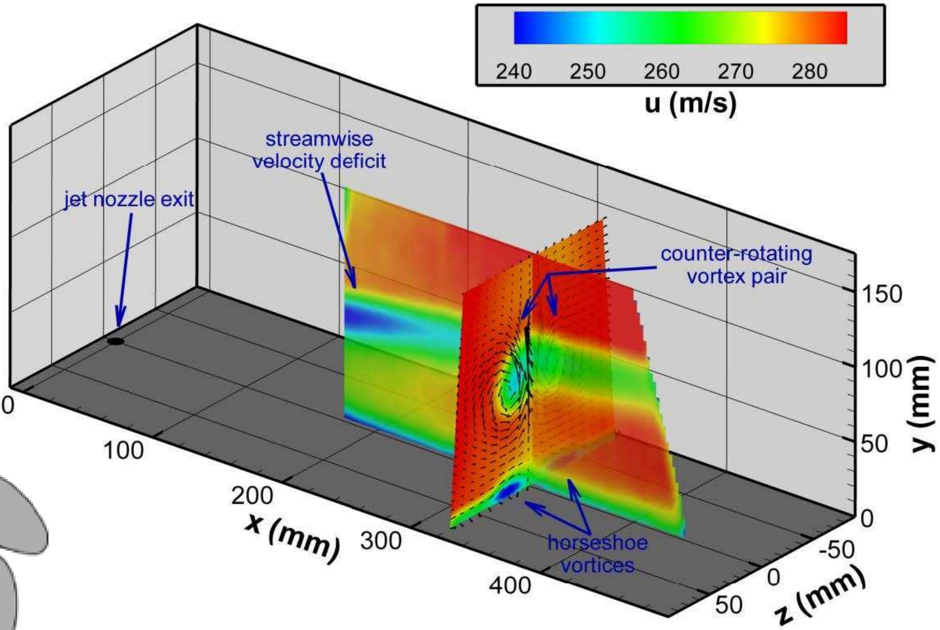
BACK-UP SLIDES

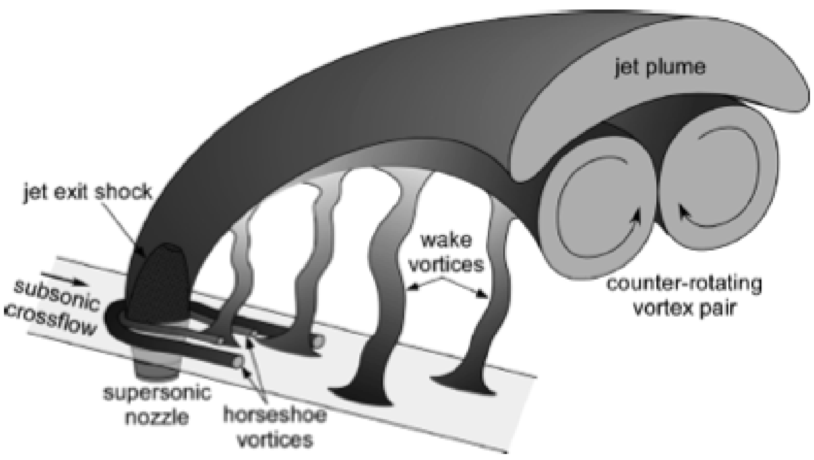
Introduction

- Aim: Develop a principled way of enriching a turbulence model to reduce model-form error
  - Needed for a predictive RANS simulator for transonic jet-in-crossflow
- Drawback: RANS simulations are simply not predictive
  - They have "model-form" error i.e., missing physics
  - They use parameters derived from canonical flows quite unlike jet-in-crossflow interactions.
- Hypothesis
  - Once a RANS model has been calibrated to a jet-in-crossflow experiment, any lack of predictive skill is due to model-form uncertainty i.e., shortcomings of the linear eddy viscosity model (LEVM)
  - If the LEVM can be enriched with higher-order terms and re-calibrated, we could reduce the error further

Target problem - jet-in-crossflow

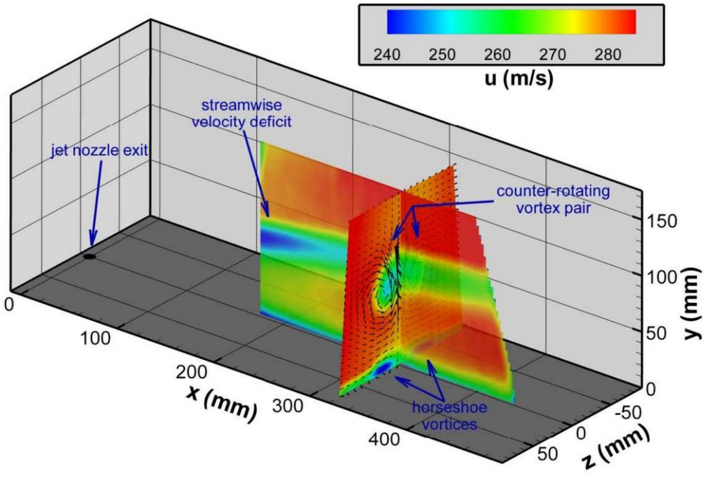
- A canonical problem for spin-rocket maneuvering, fuel-air mixing etc.
- We have experimental data (PIV measurements) and corresponding RANS simulations

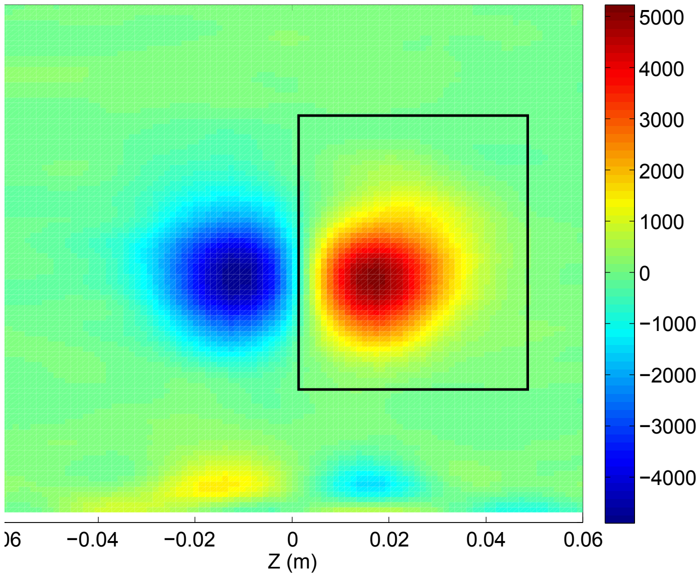




Bayesian calibration

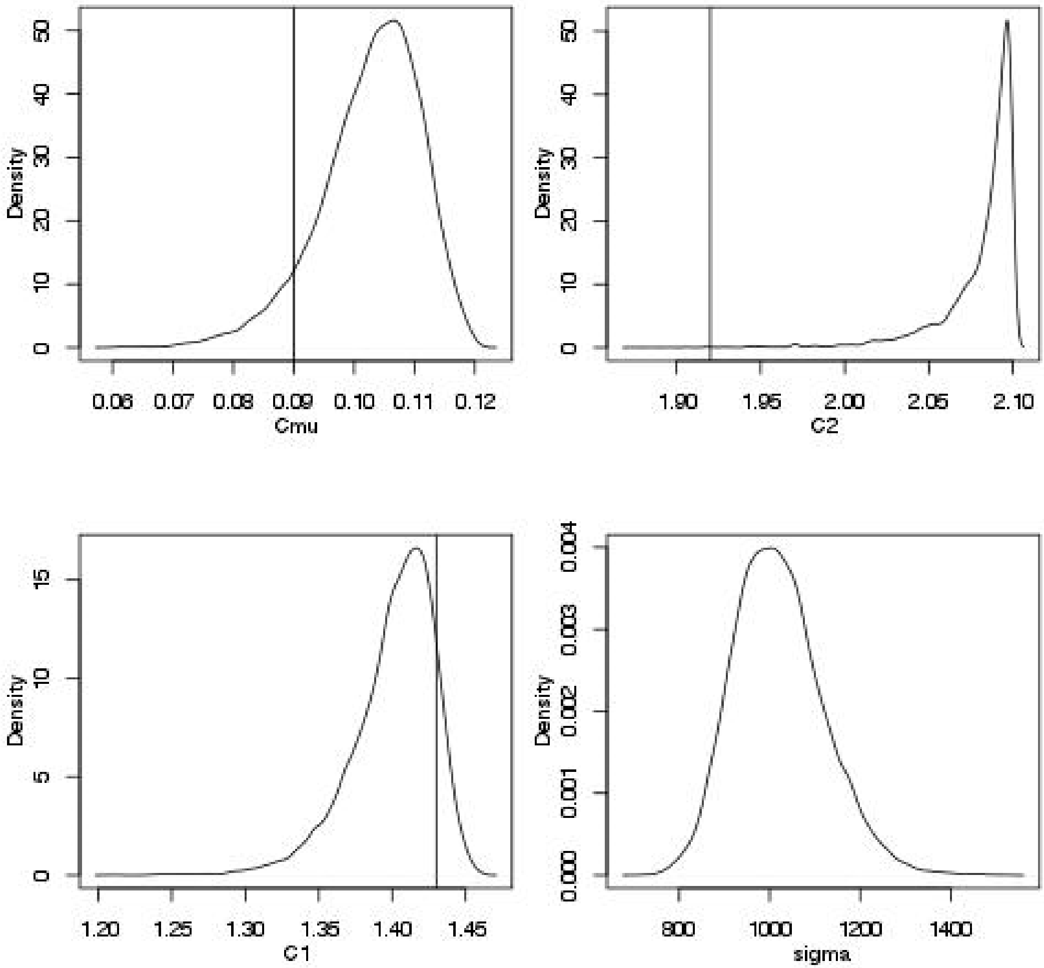
- We have velocity measurements on the crossplane
  - We computed a vorticity field
  - And used that (in a window) as the calibration variable
- We create a training set of 2744 3D RANS simulations by sampling in the (C_μ, $C_{\varepsilon 2}$, $C_{\varepsilon 1}$) space
- We create statistical models for ω_i = $\omega_i(C_\mu, C_{\varepsilon 2}, C_{\varepsilon 1})$ using polynomials
  - ω_i is the streamwise vorticity in grid-cell I
- The statistical models were used in Bayesian inversion, in lieu of the RANS simulator



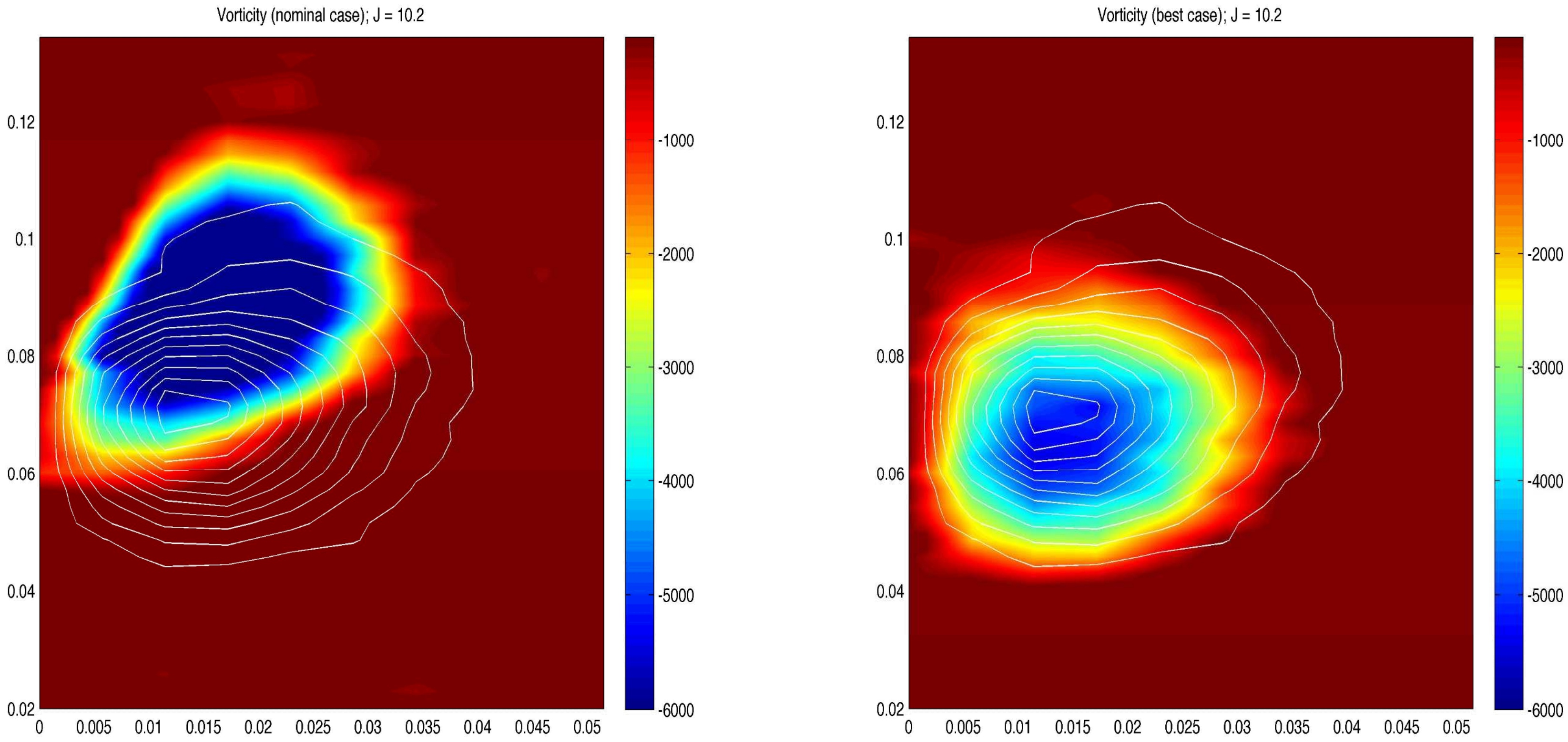


PDF of (C_μ, $C_{\varepsilon 2}$, $C_{\varepsilon 1}$)

- Marginalized versions of the 3D PDF shown here
  - Vertical lines are the "nominal" values of the parameters
- We sampled 100 (C_μ, $C_{\varepsilon 2}$, $C_{\varepsilon 1}$) realizations from the PDF
  - Generated 100 realizations of the crossplane vorticity field using the RANS simulator
  - Also found the best (C_μ, $C_{\varepsilon 2}$, $C_{\varepsilon 1}$) combination my matching the experimental vorticity field

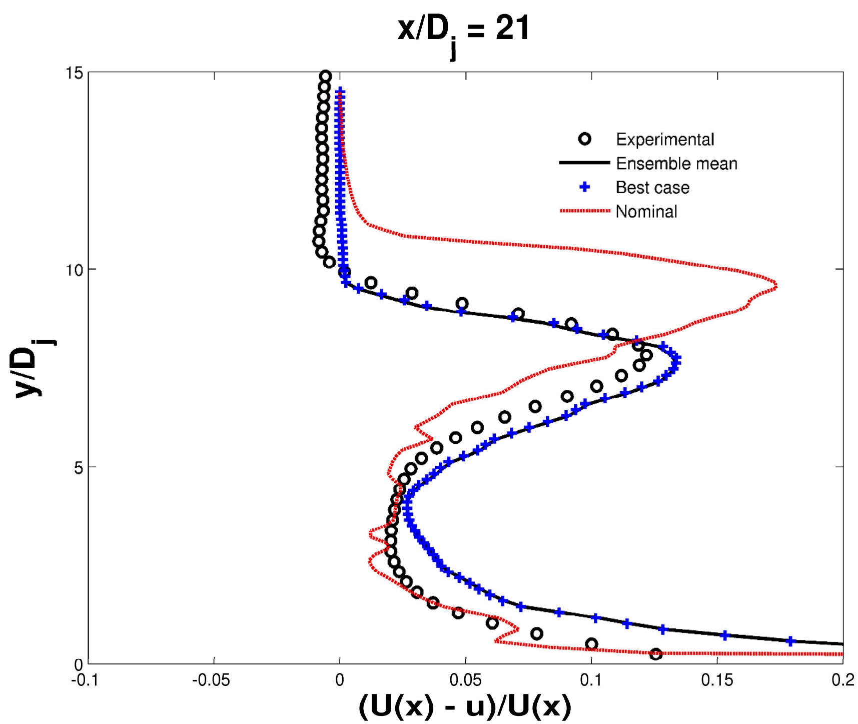


Crossplane predictions

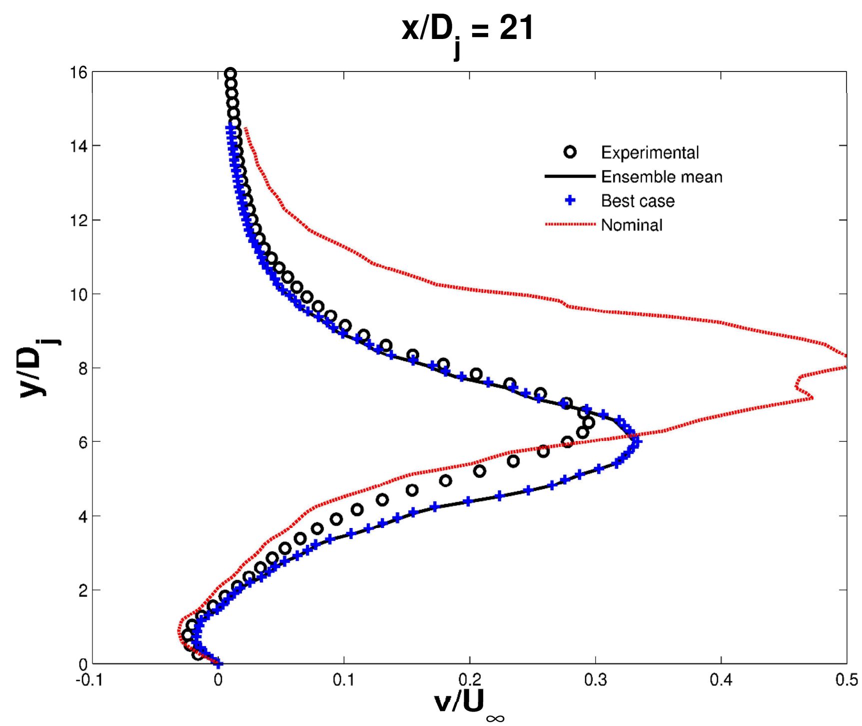


- Experimental vorticity in contours
- Stunning improvement in vorticity predictions

Mid-plane predictions



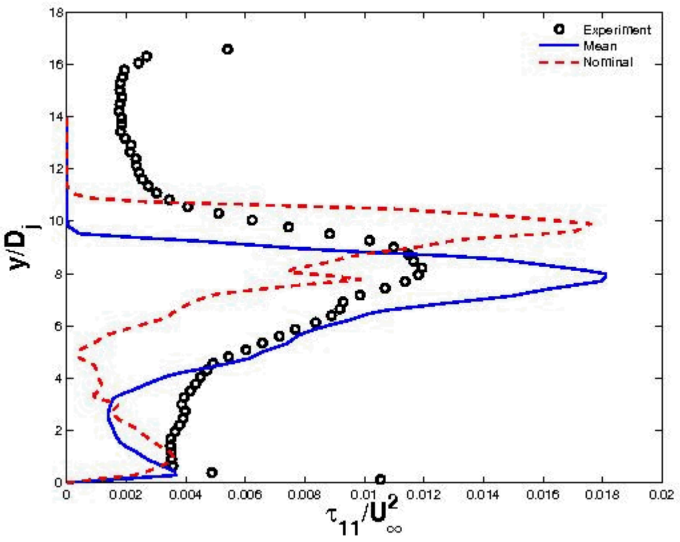
Velocity deficit



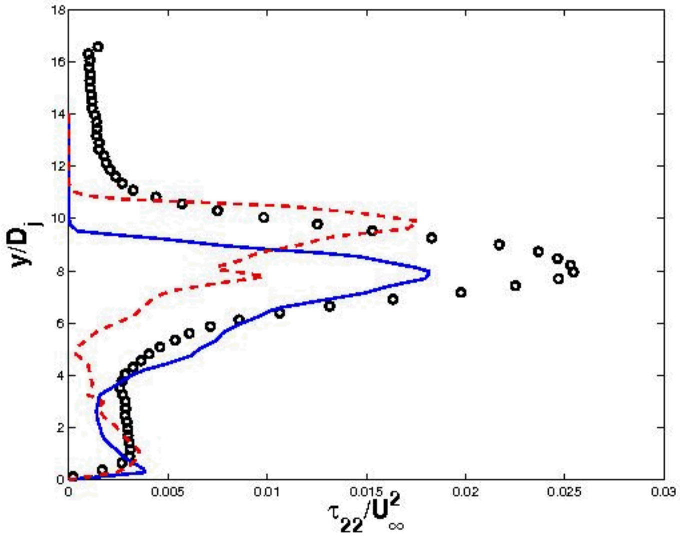
V-velocity

- Stunning improvement in vertical velocity predictions

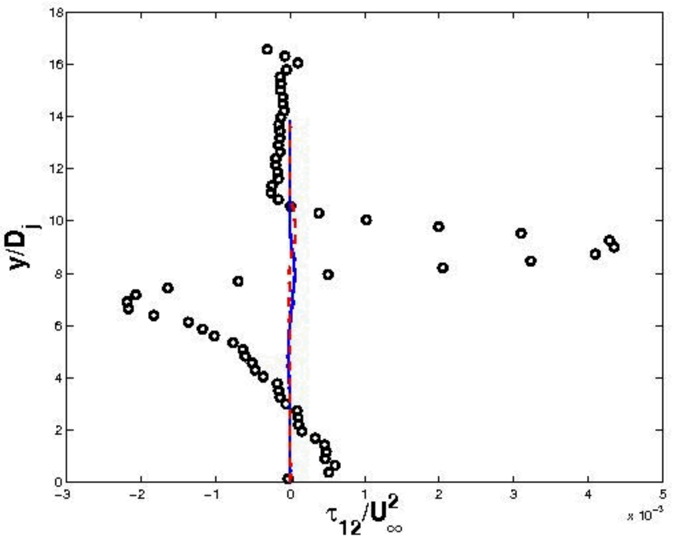
Prediction of turbulent stresses



τ_{11}



τ_{22}



τ_{12}

- M=0.8, J=10.2
- Not very good agreement; LEVM is deficient
- Improve it

Estimation of ffCO2 emissions (shrinkage & wavelet-based random field models)

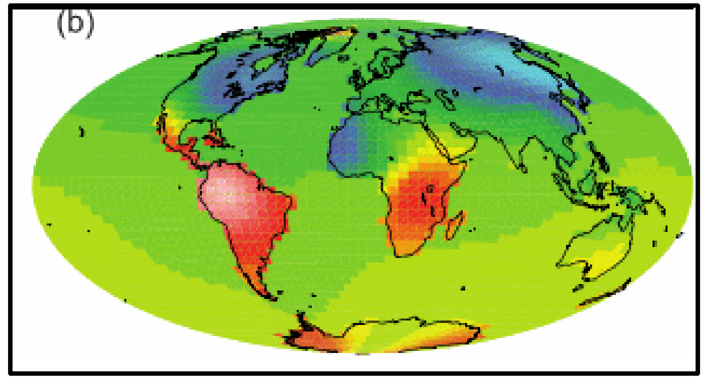
BACK-UP SLIDES

The ffCO2 estimation problem

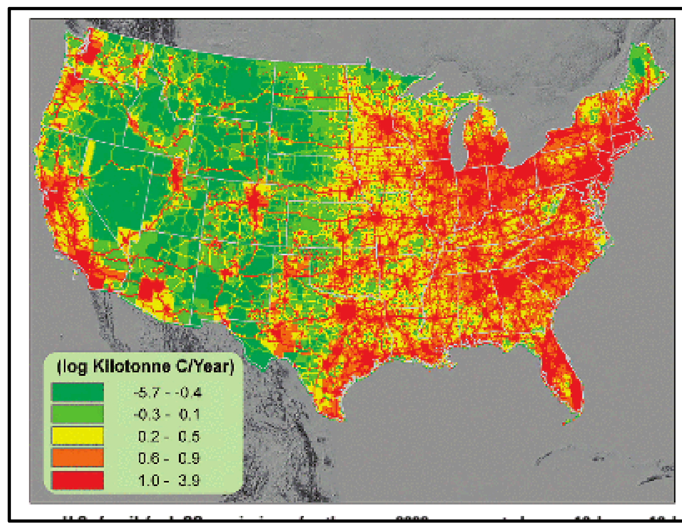
- Aim: Develop a technique to estimate anthropogenic CO_2 emissions from sparse observations
- Motivations:
  - An alternative to estimating CO_2 emission using bottom-up (economic model) techniques
  - Can provide independent verification in case of CO_2 abatement treaties
- How is it done?
  - Measure CO_2 concentrations in flasks at measurement sites; also column-averaged satellite measurements
  - Use an atmospheric transport model to invert for source locations

CO_2 flux inversions

- Biogenic CO_2 fluxes:
  - Smoothly variable in space
  - Modeled using multivariate Gaussian
  - Separate correlation lengths over land and oceans
- Anthropogenic (fossil fuel) emissions
  - Currently, only bottom-up estimates exist
  - A few databases – Vulcan (US-only, 2002); EDGAR (world)
  - Gaussian process will probably not work
    - What non-stationary covariance model to use?



Biogenic emissions: Mueller et al, *JGR*, 2008



Anthropogenic emissions: Gurney et al, *EST*, 2009

Spatial modeling

- An emission field on $2^N \times 2^N$ pixels grid
  - Can be decomposed on a wavelet basis, *N* deep
  - Each level *s* has $2^s \times 2^s - (2^{s-1} \times 2^{s-1})$ weights
- Spatial model for emissions

$$e(x) = \sum_{s=1}^{N} \sum_{i=1}^{2^s} \sum_{j=1}^{2^s} w_{s,i,j} \phi_{s,i,j}(x) = \mathbf{\Phi w}$$

  - ϕ are orthogonal bases (wavelet basis) of different resolution (scale)
  - A priori, the model is not low-dimensional (**w** is large)
- Conjecture
  - $w_{s,i,j}$ are mostly zero (i.e., is sparse)
  - Most can be removed by comparing to a wavelet transform of nightlights
  - Of the remaining, a fraction (near cities) may be estimated from observations; rest are small and can be set to zero

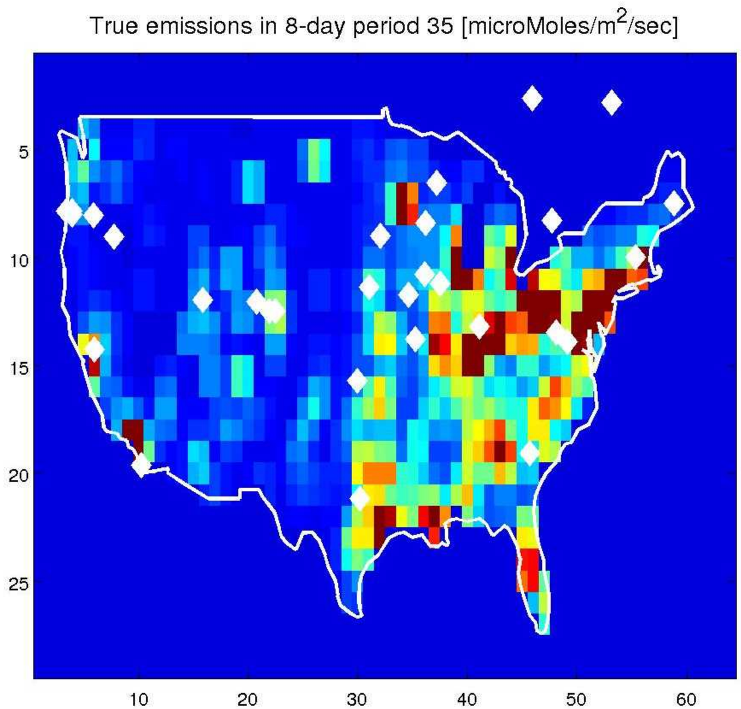
Dimensionality reduction

- Nightlights are a good proxy for FF emissions
  - Except emissions from electricity generation and cement production
  - Nightlights easily observed – DoD's DMSP-OLS
- Use thresholded radiance-calibrated nightlights from 1997-98 to mask out unpopulated regions

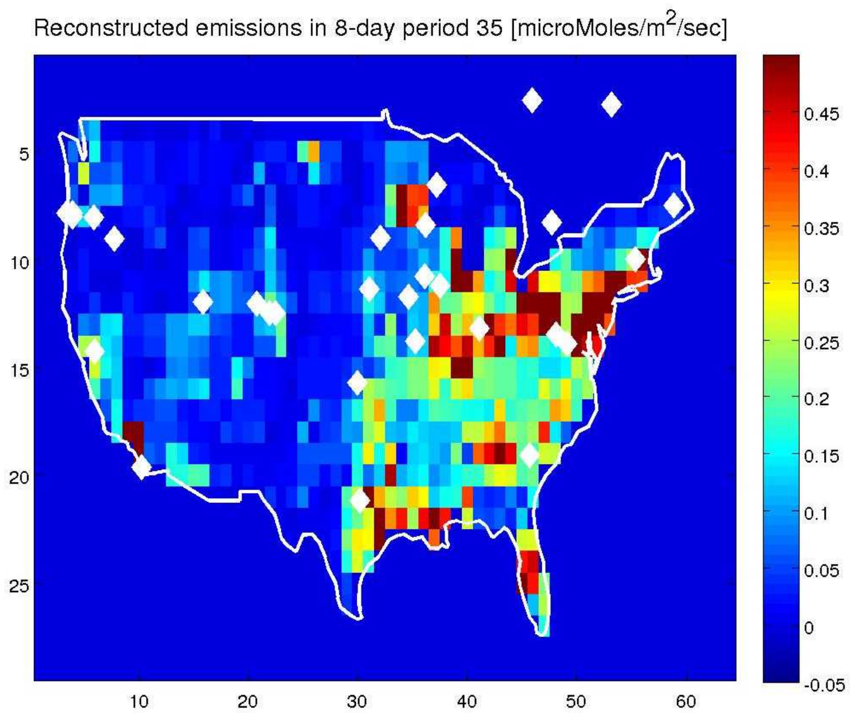
Reconstruction – via penalized optimization

- Typically, when fitting, we would solve
  - minimize $| \mathbf{y}^{(obs)} - [\mathbf{A}][\boldsymbol{\Phi}]\mathbf{w} |_2$ wrt **w**
- Sparsity-enforced (we want a sparse w)
  - minimize $| \mathbf{y}^{(obs)} - [\mathbf{A}][\boldsymbol{\Phi}]\mathbf{w} |_2 + |\mathbf{w}|_1$
  - The last penalty cuts down on the # of elements in **w**
- Many algorithms to solve this – usually formulated as
  - Minimize $|\mathbf{w}|_1$ under the constraint $| \mathbf{y}^{(obs)} - [\mathbf{A}][\boldsymbol{\Phi}]\mathbf{w} |_2 < \varepsilon_s$
  - We use StOMP
- The ffCO2 problem
  - **[Φ]** are the basis set – in our case, Haar wavelets; **w** are the wavelet coefficients; **[A]** is the transport matrix **[H]**
  - $\mathbf{y}^{(obs)}$ are tower measurements of CO_2 concentrations
  - minimize $| \mathbf{y}^{(obs)} - [\mathbf{H}][\boldsymbol{\Phi}]\mathbf{w} |_2 + |\mathbf{w}|_1$

How good is the reconstruction?



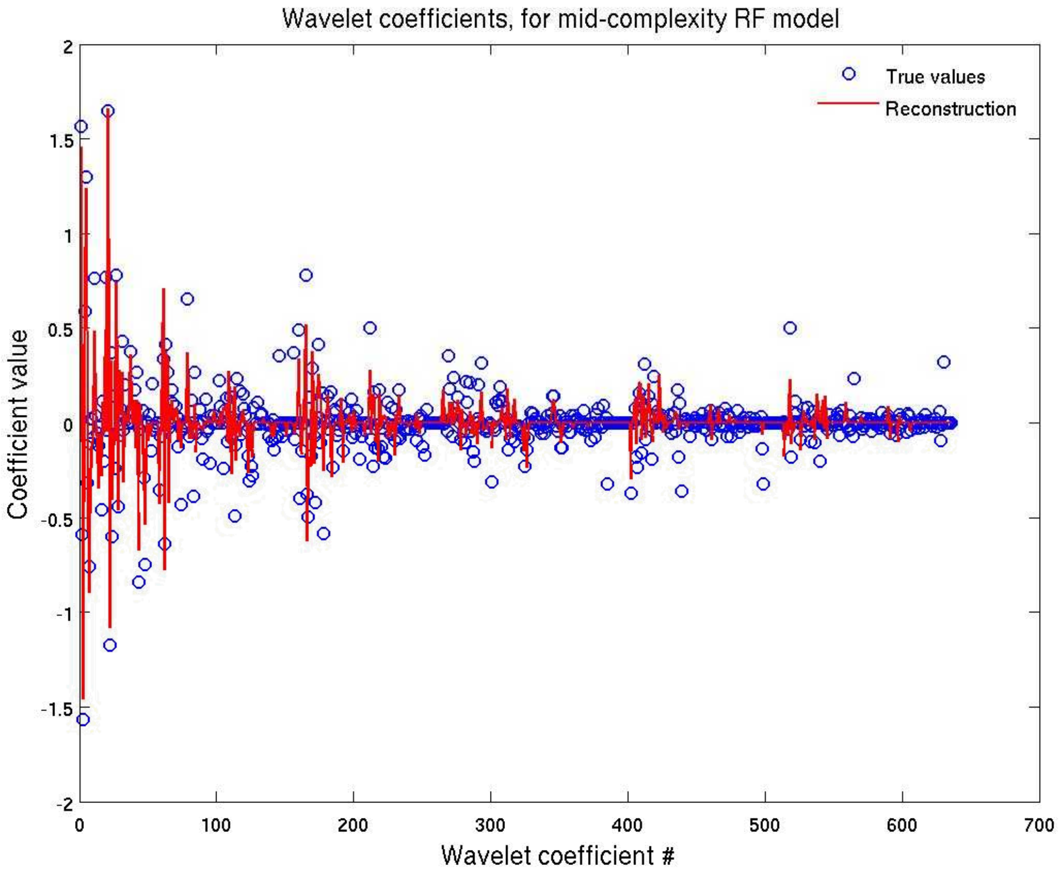
True emissions



Reconstructed emissions

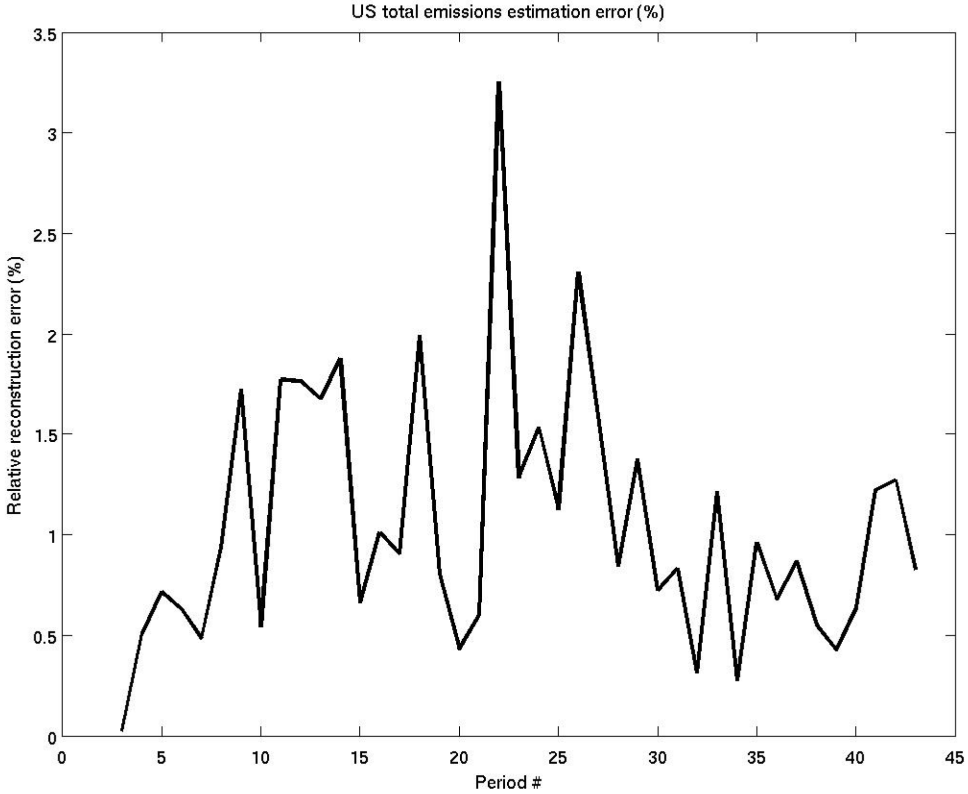
- A week in September 2008

Did sparsification work?



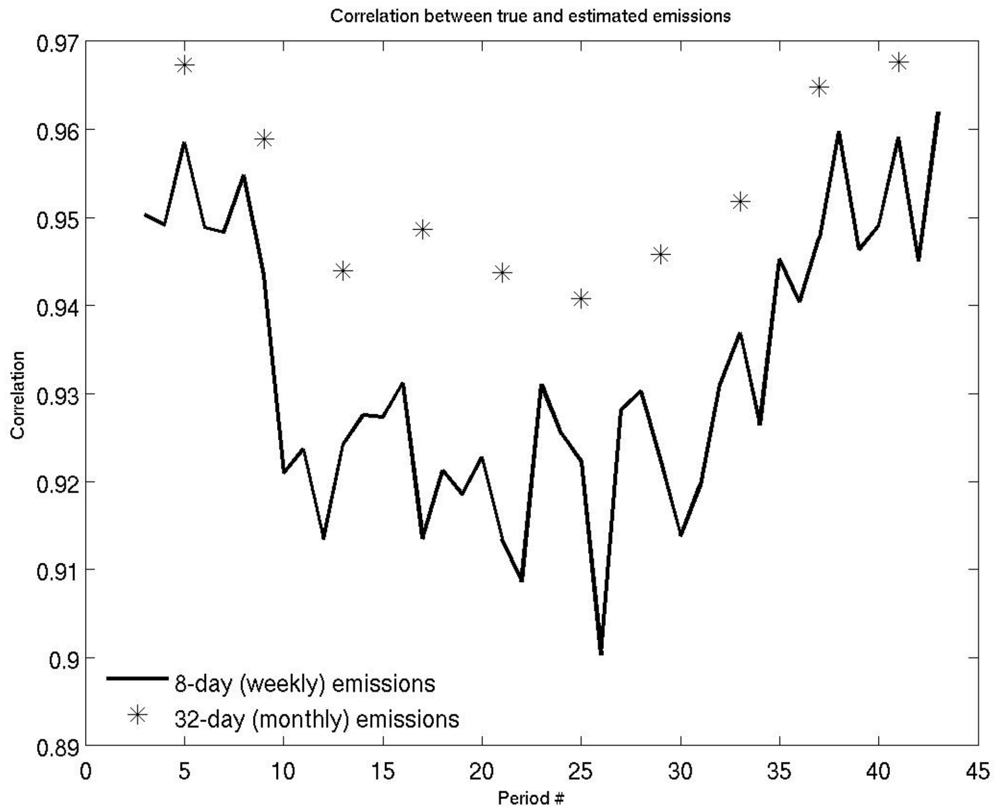
- Only about half the wavelets could be estimated
- We are probably not over-fitting the problem
  - Data-driven sparsification works

Reconstruction error in total US emission



- We get about 3.5% error, worst case

Is the spatial distribution correct?



- The spatial distribution of emissions is very close to truth
- Especially, if considering monthly fluxes