SAND2012-7811P

Solving inverse problems with quantified uncertainty

J. Ray[1]
jairay [at] sandia [dot] gov

In collaboration with B. van BloemenWaanders[2], S. A. McKenna[2], Vineet Yadav[3] and A. M. Michalak[3]

[1]Sandia National Laboratories, Livermore, CA,
[2]Sandia National Laboratories, Albuquerque, NM
[3]Carnegie Institution for Science, Stanford, CA

Funded by the LDRD program in Sandia National Labs and by DoE/Office of Science
SAND2012-ABCD

What is this talk about?

- Probabilistic/statistical methods to solve inverse/parameter estimation problems
  - Useful when you think the estimated parameters may be wrong or uncertain
    - When the model is not really a good fit to data
    - When the data is limited
    - When there are too many parameters to be estimated
- In such difficult parameter estimation scenarios, really 2 ways out
  - Estimate parameters as probability distributions i.e. as PDFs
  - Estimate only those parameters that can be constrained by the data
    - Called dimensionality reduction
    - But how do you find, *a priori*, the parameters to drop?

Some definitions

- Inverse problems – basically data fitting with a model
  - With the aim of estimating model parameters (uncorrelated variables, *a priori*)
  - Or model inputs e.g., a spatially variable material property field
    - Often discretized on a grid
- Components of an inverse problem
  - The data or observations, $\mathbf{y}^{(\mathbf{obs})}$
  - The model inputs or parameters, $\mathbf{p}$
  - The forward model, M($\mathbf{p}$)
    - $\mathbf{y}^{(\mathbf{obs})} = M(\mathbf{p}) + \varepsilon$, ε is noise or measurement error
  - A model for noise ε, if doing a probabilistic/statistical/Bayesian inverse problem
    - Often, nothing more than i.i.d. Gaussians, $N(0, \sigma^2)$

Outputs and issues

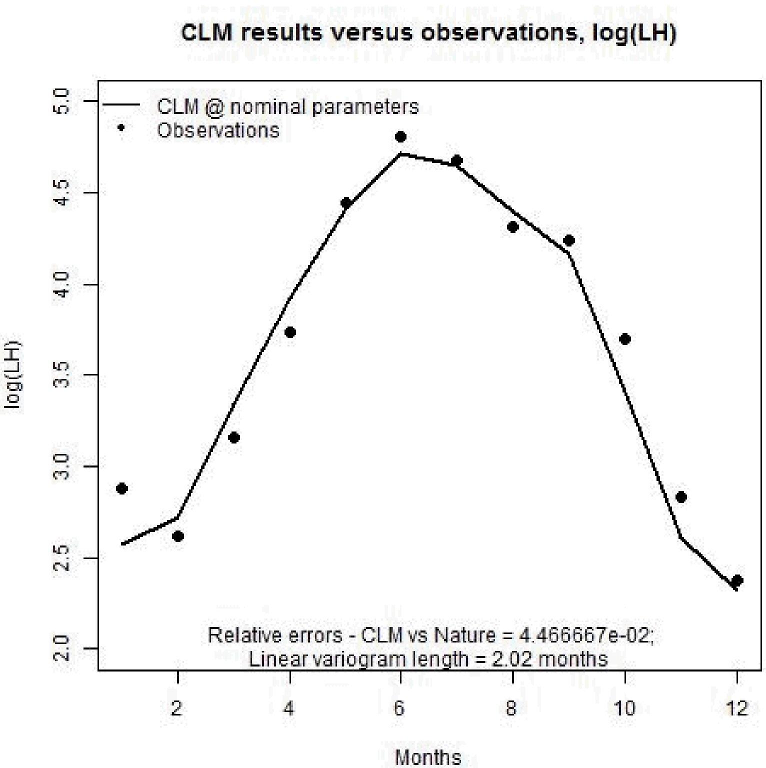
- Bayesian inverse problems estimate **p** as a joint PDF
  - All elements of **p** are included, even if the data contains no info on them
- When only “constrainable” elements of **p** are estimated
  - Sparsity-enforced optimization/reconstruction
  - Requires certain mathematical requirements before one attempts this
- Issues
  - Bayesian inverse problems require many evaluations of forward model – impossible if dealing with a computationally expensive PDE
    - Have to take recourse to surrogate models
  - Elegant simplifications for linear inverse problems i.e. if M(**p**) = **[M]p**

Outline of the talk

- Problem I – demonstrate Bayesian inversion in action
  - Estimate 3 parameters of a computationally expensive climate model
  - Issues in making a surrogate
  - Derivation of the inverse problem; numerical scheme
  - Calibration, results & implications
- Problem II – demonstrate sparsity-enforced reconstruction of a field
  - Estimate anthropogenic CO_2 emissions, on a grid
  - Concept of compressive sensing and sparse reconstruction
  - Estimate emissions in only those grid-cells which are constrained by observations, but ….
    - We don't know a priori which grid-cells are "constrainable"

Problem I – Calibration of the CLM

- CLM - Community Land Model
  - Used in Earth System models; vegetation, hydrology, evaporation, heat balance etc.
- Expensive; 1 hr/invocation for 1 site
- Desired: estimate **p** = {F_{drai}, Q_{dm}, S_y}, 3 hydrological parameters
  - Data: monthly Latent Heat (LH) measurements @ US-Moz site
  - “Nominal” values of p known
- What is the distribution of **p**?
  - Compare with nominal value



Formulation (1/2)

- Let y(obs) be the observations – $\mathbf{y}^{(\mathbf{obs})} = \{y^{(obs)}{}_i\}$, i = 1 … 12 months
- Model predictions $\mathbf{y}^{(\mathbf{pred})} = M(\mathbf{p})$, $\mathbf{p} = \{F_{drai}, \log(Q_{dm}), S_y\}$
- Error $\varepsilon = \mathbf{y}^{(\mathbf{obs})} - M(\mathbf{p})$, $\varepsilon_i \sim N(0, \sigma^2)$

$$P(\varepsilon_i \mid \mathbf{p}) \propto \exp\left(-\frac{\varepsilon_i^2}{\sigma^2}\right); \quad P(\boldsymbol{\varepsilon} \mid \mathbf{p}) \propto \prod_{i=1}^{12} \exp\left(-\frac{\varepsilon_i^2}{\sigma^2}\right) = \exp\left(-\frac{\|\varepsilon\|_2}{\sigma^2}\right)$$

$$P(\mathbf{y}^{(\mathbf{obs})} \mid \mathbf{p}) \propto \exp\left(-\frac{\left\|\mathbf{y}^{\mathbf{obs}} - M(\mathbf{p})\right\|_2}{\sigma^2}\right)$$

- $P(\varepsilon|\mathbf{p}) = P(\mathbf{y}^{(\mathbf{obs})}|\mathbf{p})$ is called the Likelihood

Sandia National Laboratories

Formulation (2/2)

- We desire P(**p** | **y**$^{(obs)}$) (aka the posterior distribution) and σ^2
- Bayes rule

$$\underbrace{P(\mathbf{p} \mid \mathbf{y}^{\mathbf{(obs)}})}_{Posterior} P(\mathbf{y}^{\mathbf{(obs)}}) = \underbrace{P(\mathbf{y}^{\mathbf{(obs)}} \mid \mathbf{p})}_{Likelihood} \underbrace{P(\mathbf{p})}_{\mathrm{Pr}ior}$$

- Prior distribution for **p**
  - Each of the components $\{F_{drai}, \log(Q_{dm}), S_y\}$, are independent
  - They have uniform distributions between a specified UB and LB
- We need to evaluate P(**p** | **y$^{(obs)}$**)
  - How? Using a sampler – Markov Chain Monte Carlo (MCMC)

What is MCMC?

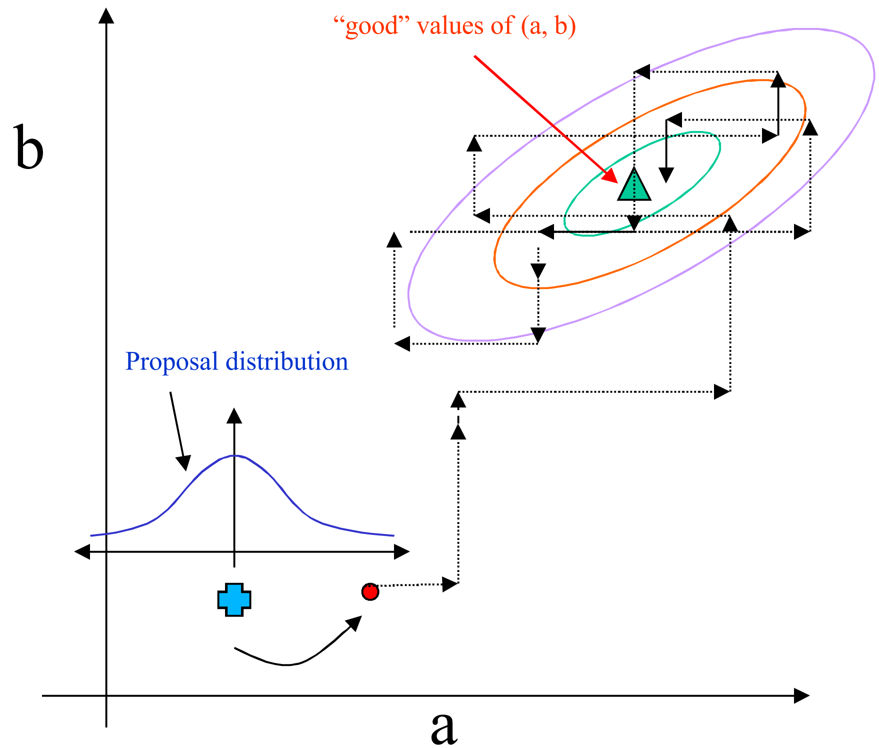
- A way of sampling from an arbitrary distribution
  - The samples, if histogrammed, recover the distribution $P(\mathbf{p} \mid \mathbf{y}^{(obs)})$
- Efficient and adaptive
  - Given a starting point (1 sample), the MCMC chain will sequentially find the peaks and valleys in the distribution and sample proportionally
- Ergodic
  - Guaranteed that samples will be taken from the entire range of the distribution
- Drawback
  - Generating each sample requires one to evaluate the expression for the density P
  - i.e., a model evaluation – very expensive!

An example, using MCMC

- Given: (Y^{obs}, X), a bunch of n observations
- Model: $y_i^{obs} = ax_i + b_i + \varepsilon_i$, $\varepsilon \sim \mathcal{N}(0, \sigma)$
- Priors : uniform distributions for a, b, σ
- For a given value of (a, b, σ), compute "error" $\varepsilon_i = y_i^{obs} - (ax_i + b_i)$
  - Likelihood of the set $(a, b, \sigma) = \Pi \exp(- \varepsilon_i^2/\sigma^2)$
- Solution: $\pi (a, b, \sigma \mid Y^{obs}, X) = \Pi \exp(- \varepsilon_i^2/\sigma^2)$ * (bunch of uniform priors)
- Solution method:
  - Sample from $P(a, b, \sigma \mid Y^{obs}, X)$ using MCMC; save them
  - Generate a "3D histogram" from the samples to determine which region in the (a, b, σ) space gives best fit
  - Histogram values of a, b and σ, to get individual PDFs for them
  - Estimation of model parameters, with confidence intervals!

MCMC, pictorially

- Choose a starting point, O^n = (a_{curr}, b_{curr})
- Propose a new a, $a_{prop} \sim \mathcal{N}(a_{curr}, \sigma_a)$
- Evaluate P(a_{prop}, b_{curr} | ...) / P(a_{curr}, b_{curr} | …) = m
- Accept a_{prop} (i.e. a_{curr} <- a_{prop}) with probability min(1, m)
- Repeat with b
- Loop over till you have enough samples



Sandia National Laboratories

Surrogate model

- Usually MCMC needs $10^3 - 10^7$ steps to converge to a distribution
  - Can't use CLM as-is; need to make a surrogate ("curve-fit" model)
- Procedure
  - Sample 128 points in **p**-space
    - Used a method called quasi-Monte Carlo to spread out the samples evenly in the 3D parameter space
  - Run CLM; obtain log(LH) = $\mathbf{y}^{(\mathrm{clm})}_j$, j = 1…128
  - Propose a polynomial model, but where to stop?

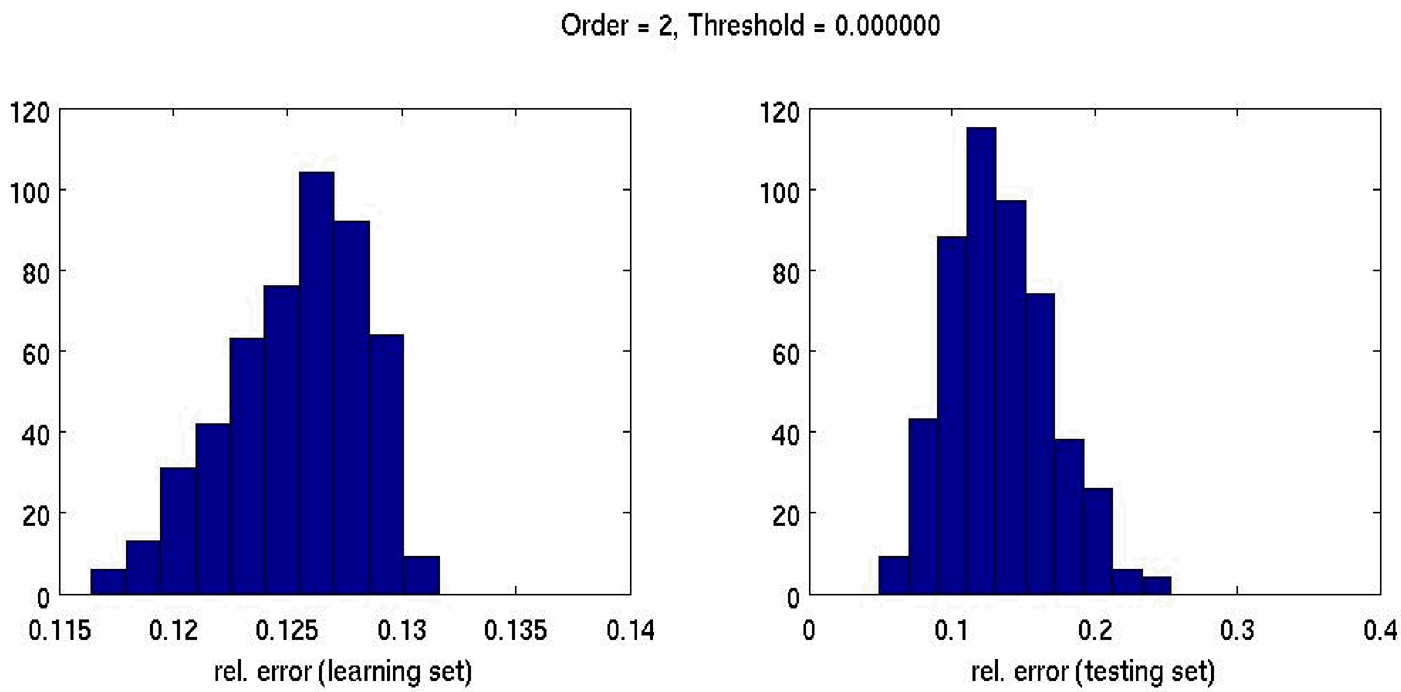
$$\log(LH) = \mathbf{y}^{(\mathbf{clm})} = \sum_{i=1}^{3} \alpha_i p_i + \sum_{i=1}^{3} \sum_{j=i}^{3} \beta_{ij} p_i p_j + \sum_{i=1}^{3} \sum_{j=i}^{3} \sum_{k=(i+j)}^{3} \gamma_{ijk} p_i p_j p_k + \ldots..$$

Making the surrogate model

- Pick a month – say April
- Make 5 competing models – 1st to 5th order
- Partition the 128 runs into a 118-run Learning Set (LS) and 10-run Testing Set (TS)
- Resample the 128 runs again, make 500 {LS + TS} pairs
- For a given model, say quadratic
  - Use LS to estimate α_i, β_{ij} etc using simple regression; compute polynomial model versus CLM prediction errors (relative)
  - Predict log(LH)at the TS parameters; compute relative error vis-à-vis CLM predictions
  - Over 500 {LS+TS} pairs, one gets a distribution of LS and TS relative errors
    - What do these look like?

Sandia National Laboratories

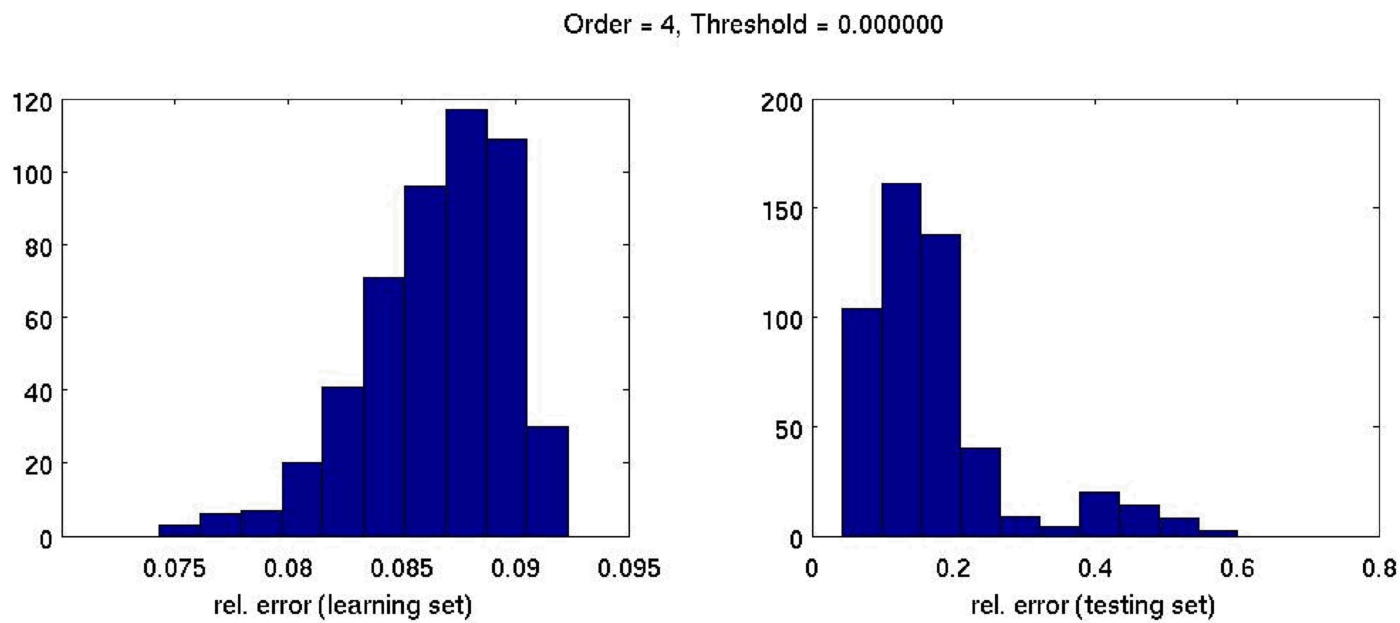
Quadratic model predictions



- LS error about the same magnitude as TS errors (~0.13)
  - Model has about the same predictive skill in LS as TS

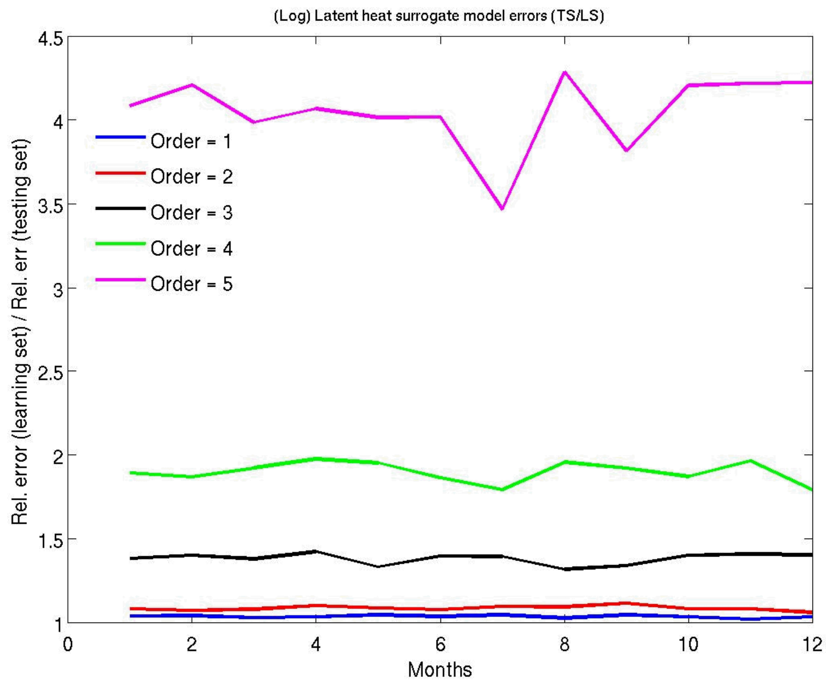
Sandia National Laboratories

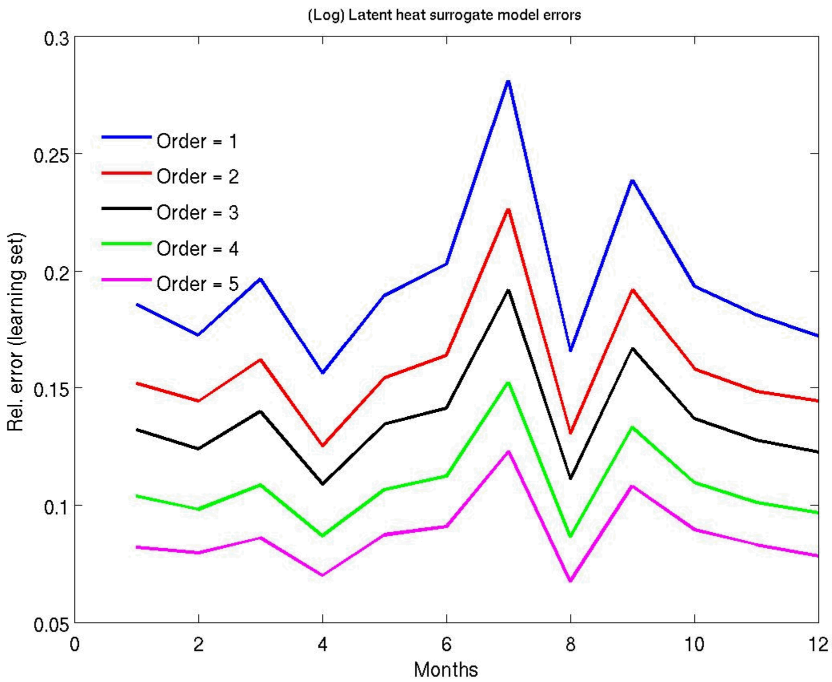
4th-order model predictions



- Learning Set error very low, Testing set error 3x bigger
- Clear case of overfitting the LS
- So which model to retain – linear to 5th order?

Plotting errors across months



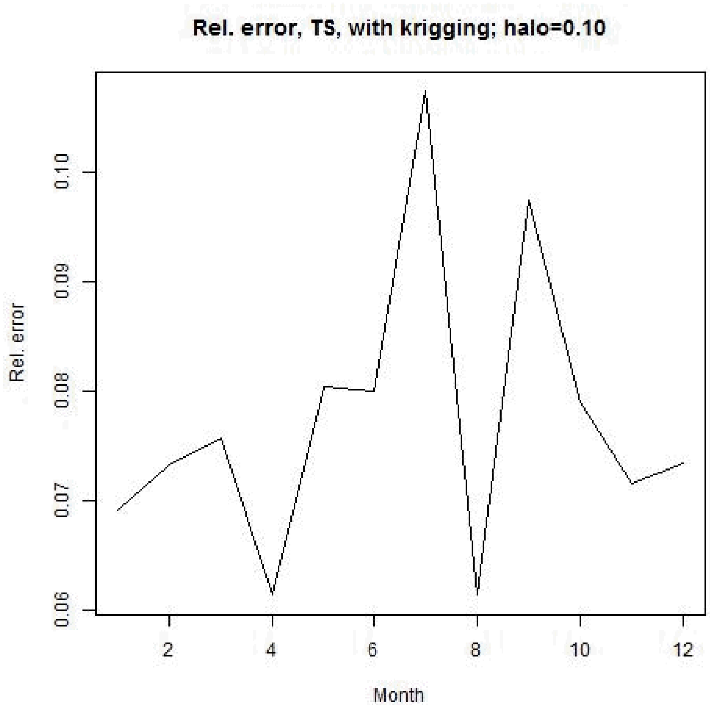


- Linear and quadratic models have similar errors for LS and TS
  - No overfitting here
- But quadratic model has lower errors overall, so choose it.

Sandia National Laboratories

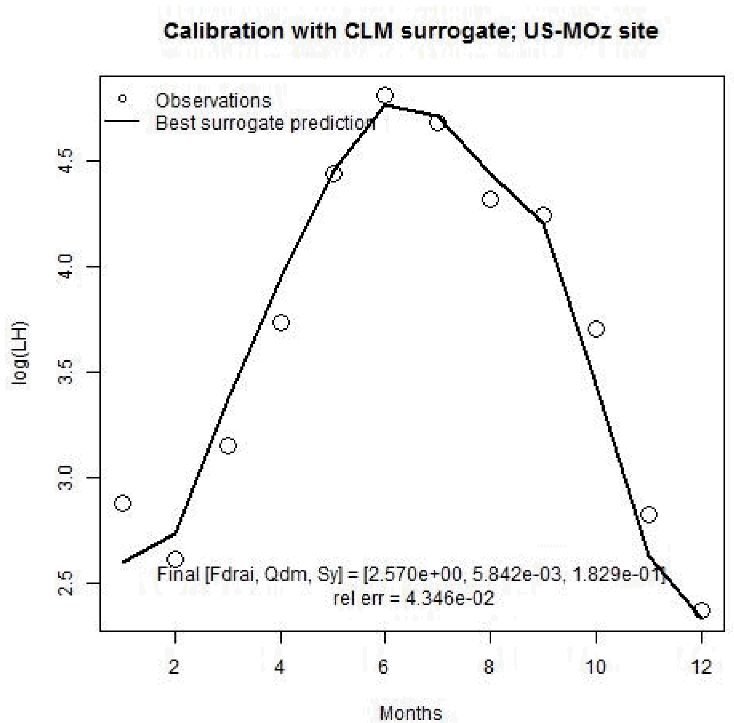
Augmenting the quadratic model

- Quadratic model has pretty large error (~17%)
  - Because it captures no more than the trend of log(LH) in **p**-space
- $\mathbf{y}^{(\mathbf{surr})}(\mathbf{p}) = \mathbf{y}^{(\mathbf{quad})}(\mathbf{p}) + \mathbf{c}(\mathbf{p})$, **c** is a correction
  - It is smooth (correlated) function of **p**
  - Model c(**p**) as a multivariate Gaussian
- With c(**p**) model, we can evaluate $\mathbf{y}^{(\mathbf{surr})}(\mathbf{p})$ at arbitrary **p**
  - Includes a quadratic prediction
  - And a correction interpolated from the 128 runs

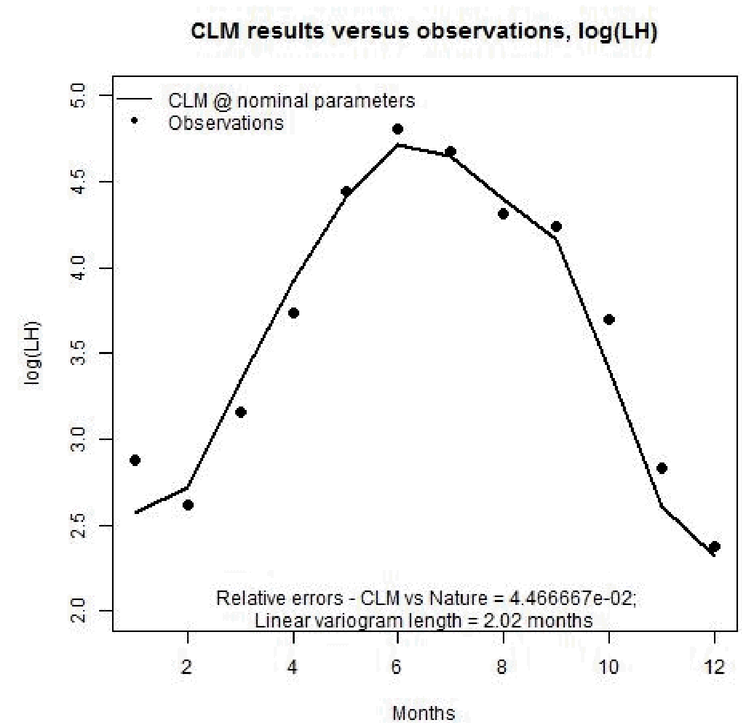


Augmented model give max 10% error

Deterministic fit to US-Moz data



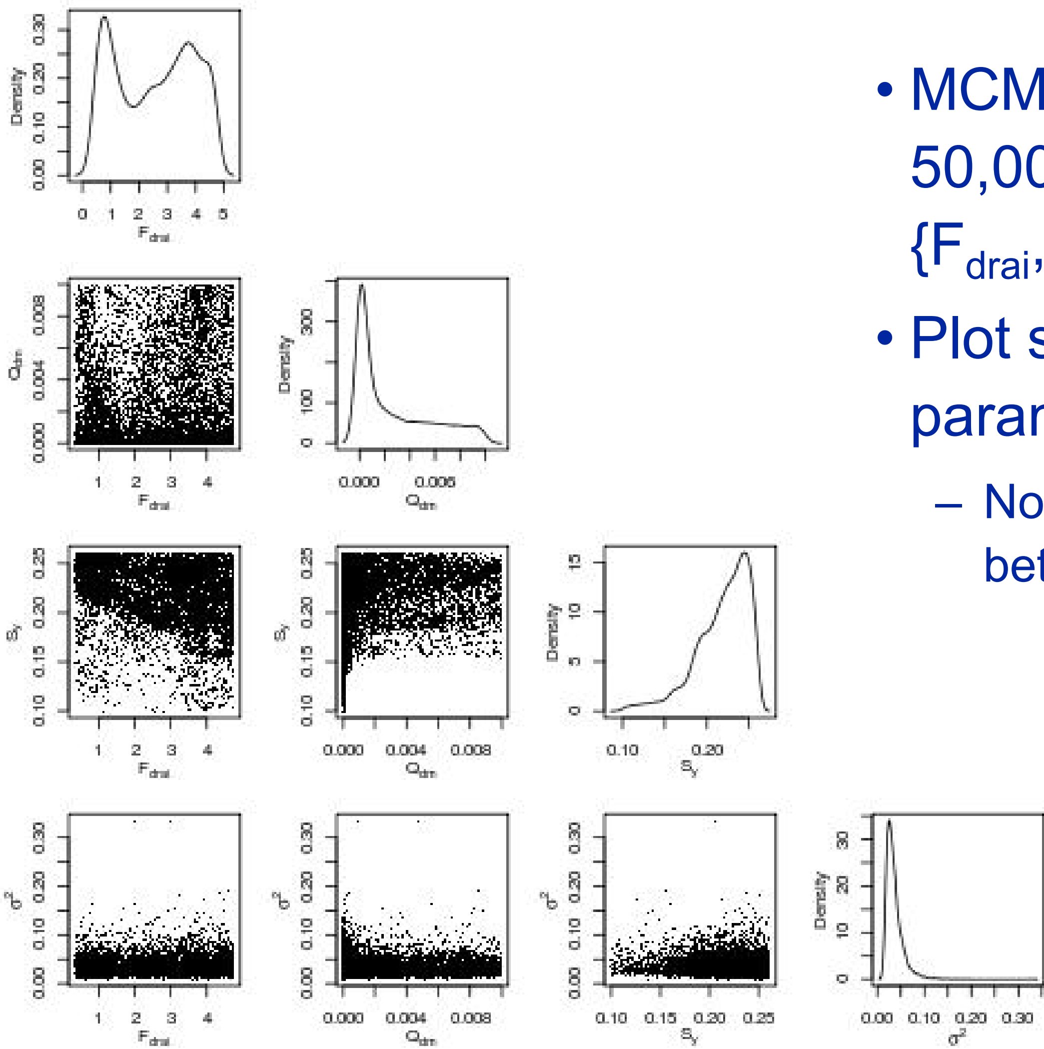
Predictions with calibrated surrogate



CLM predictions with nominal values

- Deterministic fit (w/ surrogate) and "nominal values" look similar
  - But errors sum to zero in the surrogate case

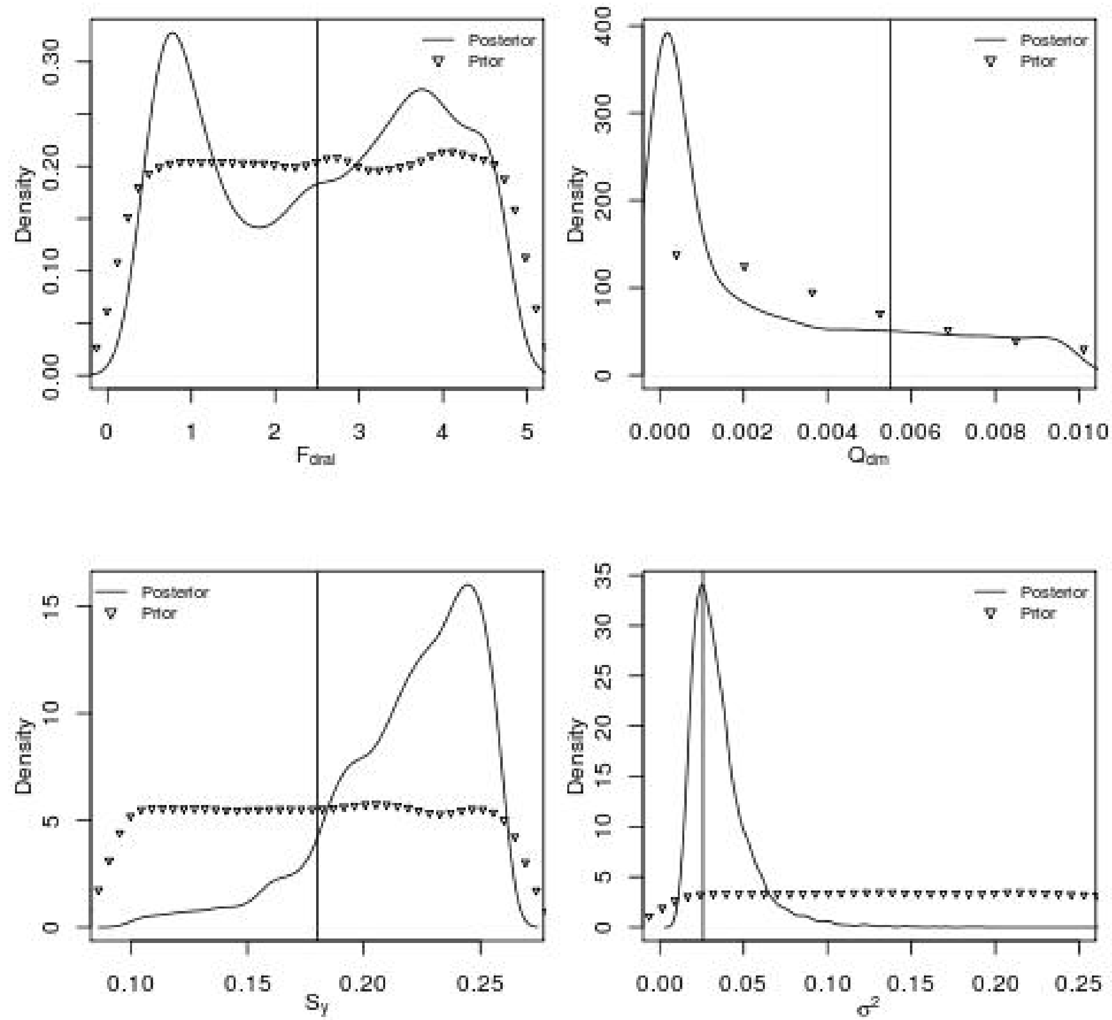
MCMC Results



- MCMC produces 50,000 samples of $\{F_{drai}, \log(Q_{dm}), S_{y,} \sigma\}$
- Plot scatter plots of 2 parameters at a time
  - No correlations between them

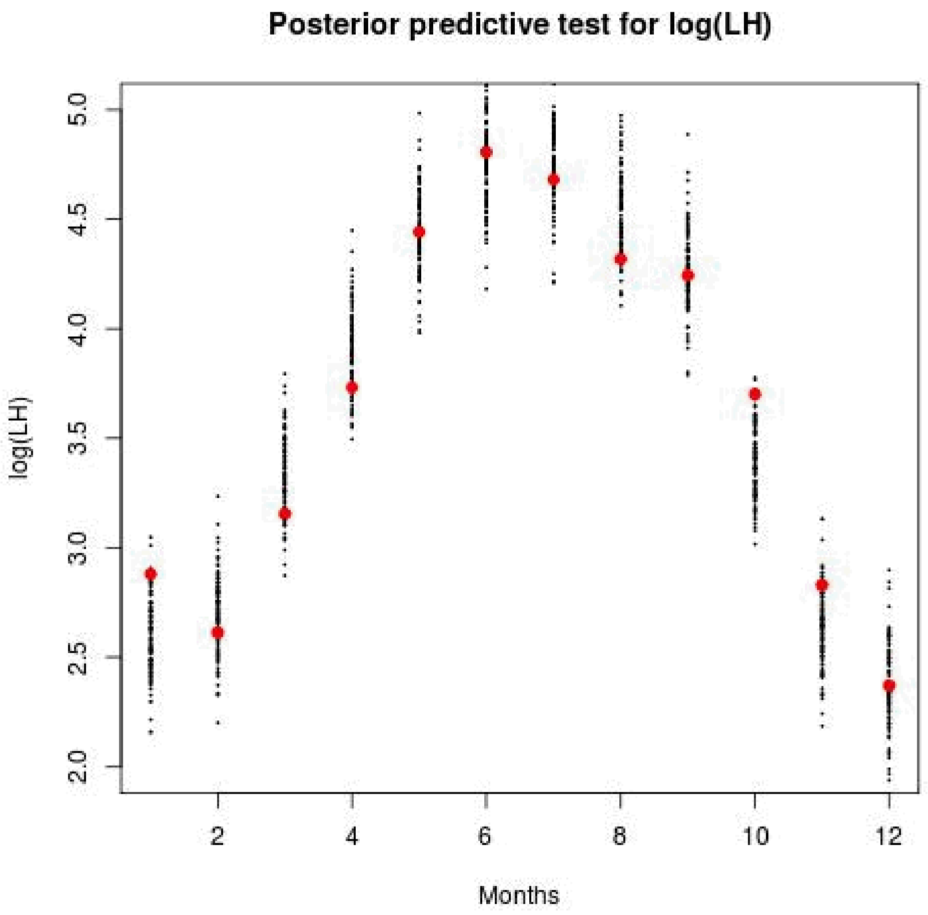
Posteriors and nominal values

- Quite a few problems with being deterministic
- Vertical lines are nominal values
- Nominal value for σ^2 is from the deterministic fit of surrogate



Posterior predictive test

- Sample 100 parameter sets from posterior
- Run forward; add noise using σ
- Plot observations
  - Predictions capture observations, for sure
- Quantify the tightness of the prediction (goodness of calibration)



Gross statistics

Parameter	Nominal Value	Mean Value from US-MOz	Median Value	Inter-quartile range
F_{drai}	2.5	2.56	2.69	1.17—3.77
Q_{dm}	5.5e-3	2.57e-3	1.06e-3	1.05e-4 – 4.63e-3
S_y	0.18	0.20	0.23	0.2 – 0.243

- The mean value from MCMC fit to MOz observations is “close” to nominal values
  - But the skewed distributions mean that the mean/median are not very representations of the high-probability points

Sandia National Laboratories

Interim conclusions – Bayesian inversion

- Bayesian methods allow us to estimate parameters as distributions
  - Distributions narrow and steepen as more data become available or when fit improves
  - Very useful, if we suspect that parameter estimates may be uncertain
  - Allow probabilistic predictions, that enable us to calculate risk of failure / error in prediction
- Can be used with computationally expensive models, if surrogates can be made
  - Often, this is the main challenge
- Can be expanded to spatial / spatio-temporal observations

Problem II – Estimation of fields

- Model parameters/inputs to be estimate can be fields
  - E.g., estimating the fossil-fuel CO_2 (ffCO2) emissions in US
  - The emissions are described on a grid; number of emissions to be estimated = # of grid cells is HUGE!
    - Aka "dimensionality of the problem is large"
  - Nowhere near enough data
- How to do this? Reduce the "effective dimensionality" by regularization
  - If field is smooth, adjacent cells cannot assume arbitrary values
  - If the field has patterns, make a spatial model (with fewer independent parameters)
- General idea – introduce constraints and reduce the # of variables to estimate

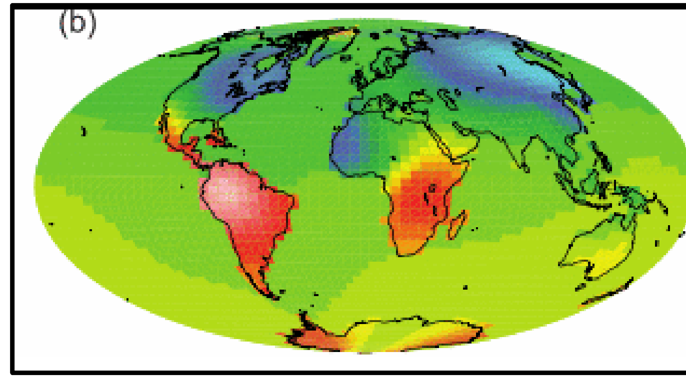
The ffCO2 estimation problem

- **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

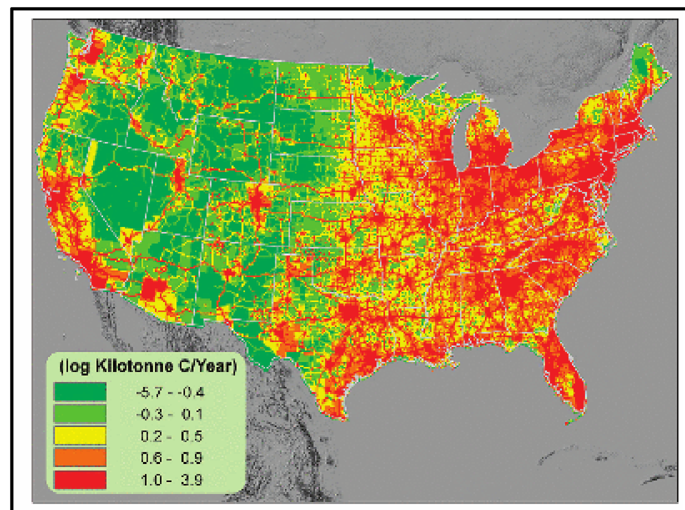
Sandia National Laboratories

CO_2 flux inversions (1/2)

- 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

Sandia National Laboratories

CO_2 flux inversions (2/2)

- NOAA runs a set of towers which measure CO_2 concentrations every 3 hours – main data source
  - Meant for biospheric fluxes (far from cities)
  - About 100 today
- ffCO2 emissions happen
  - Electricity generation (source details at http://carma.org)
  - Where people live (transport, light & heavy industry)
  - Images of lights at nights at night provide a rough spatial pattern
- Simplification – CO2 transport (source – observation linkage) is that of a passive scalar
  - $\mathbf{y}^{(pred)} = \mathbf{[H]e}$, $\mathbf{e}$ = ffCO2 emissions on a grid – a linear problem!
  - **[H]** called the transport matrix- links CO_2 concentrations at sensors with emissions **e**

Technical challenges in inversion

- Atmospheric transport model - largest source of uncertainty
- Limited measurements - second-largest contribution to uncertainty
- Discriminating between anthropogenic and biogenic CO_2 (biogenic is 10x larger)
  – But anthropogenic and biogenic CO_2 and different (and known) proportions of $^{12}CO_2$ and $^{14}CO_2$
- Spatial models for anthropogenic CO_2 emissions
  – Non-stationary distribution in space – what is the spatial model?
  – How to reduce the dimensionality of the spatial model?

Spatial modeling

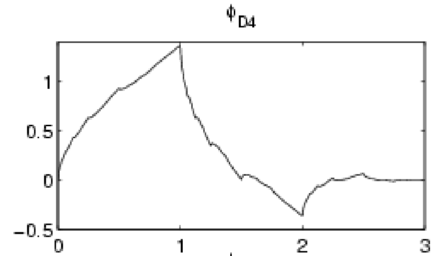
- An emission field on 2^N x 2^N pixels grid
  - Can be decomposed on a wavelet basis, *N* deep
  - Each level *s* has $2^s \times 2^s - (2^{s-1} \times 2^{s-1})$ weights
- Spatial model for emissions

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

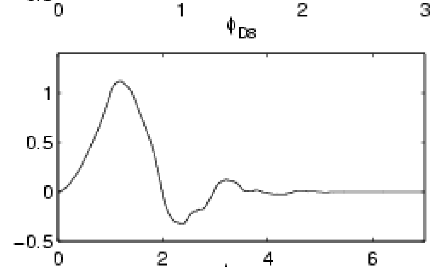
  - ϕ are orthogonal bases (wavelet basis) of different resolution (scale)
  - A priori, the model is not low-dimensional (**w** is large)
- Conjecture
  - $w_{s,i,j}$ are mostly zero (i.e., is sparse)
  - Most can be removed by comparing to a wavelet transform of nightlights
  - Of the remaining, a fraction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

How does one represent emissions with wavelets?

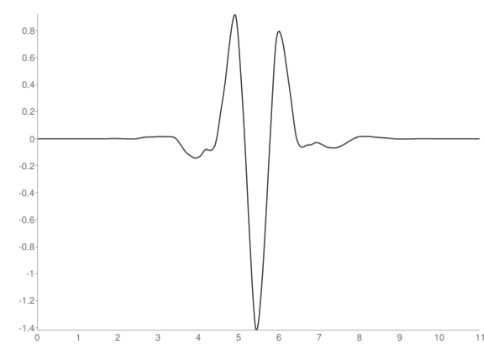
- Propose $$E(x) = \sum_{s,l} w_{s,l} \phi_{s,l}(x)$$
  - $\phi_{s,l}(x)$ is a wavelet basis; *s, l* are its *scale* and *location* indices
  - $w_{s,l}$ are weights
- So what are wavelets?
  - Basis set with compact support
  - Belong to different families
  - Within a family, can have different orders (high order ~ smoother)
  - One chooses a family and an order, to expand E(x)
  - The expansion consists of varying
    - s, to get different frequency content
    - l, to shift in space (location)



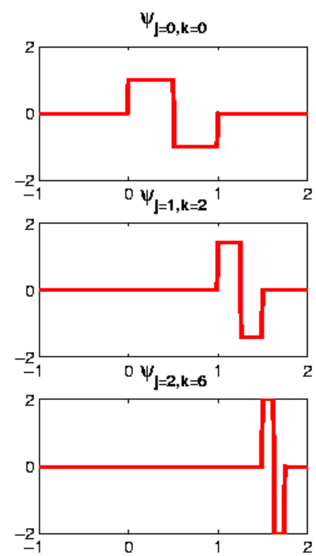
Daubechies, order 4



Daubechies, order 6



Symlet, order 6



Haars at different scales and locations

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

Detour – sparsity enforced reconstruction

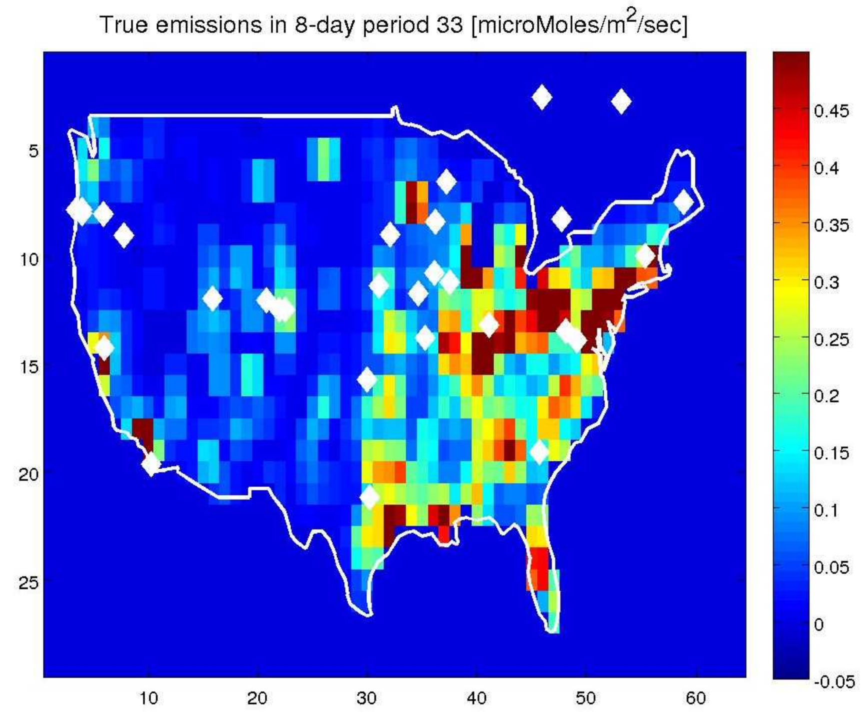
- Let **e** be a signal of length *N*– can be sampled as
  - $\mathbf{y}^{(samp)} = [\mathbf{A}]\mathbf{e}$; lossless reconstruction of **e** requires $\mathbf{y}^{(samp)}$ to be *N* long
  - **[A]** is usually random
- Suppose $\mathbf{e} = [\Phi]\, \mathbf{w}$, where Φ is an orthogonal basis set
  - And **w** is sparse; i.e., only $k \ll N$ elements of w are non-zero (don't know which)
  - To estimate the k non-zero elements of w, one needs $O(k \log_2(N/k))$ elements (samples) in $\mathbf{y}^{(samp)}$
    - Theory of compressive sampling
- Reconstruction from noisy samples posed as
  - $\mathbf{y}^{(obs)} = [\mathbf{A}][\Phi]\mathbf{w} + \varepsilon$, **w** is sparse and ε is noise

Reconstruction – via penalized optimization

- Typically, when fitting, we would solve
  - minimize | $\mathbf{y}^{(obs)} - [\mathbf{A}][\mathbf{\Phi}]\mathbf{w}$ $|_2$ wrt **w**
- Sparsity-enforced (we want a sparse w)
  - minimize | $\mathbf{y}^{(obs)} - [\mathbf{A}][\mathbf{\Phi}]\mathbf{w}$ $|_2$ + $|\mathbf{w}|_1$
  - 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 | $\mathbf{y}^{(obs)} - [\mathbf{A}][\mathbf{\Phi}]\mathbf{w}$ $|_2 < \varepsilon_s$
  - 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 | $\mathbf{y}^{(obs)} - [\mathbf{H}][\mathbf{\Phi}]\mathbf{w}$ $|_2$ + $|\mathbf{w}|_1$

Setting up the synthetic data inversion

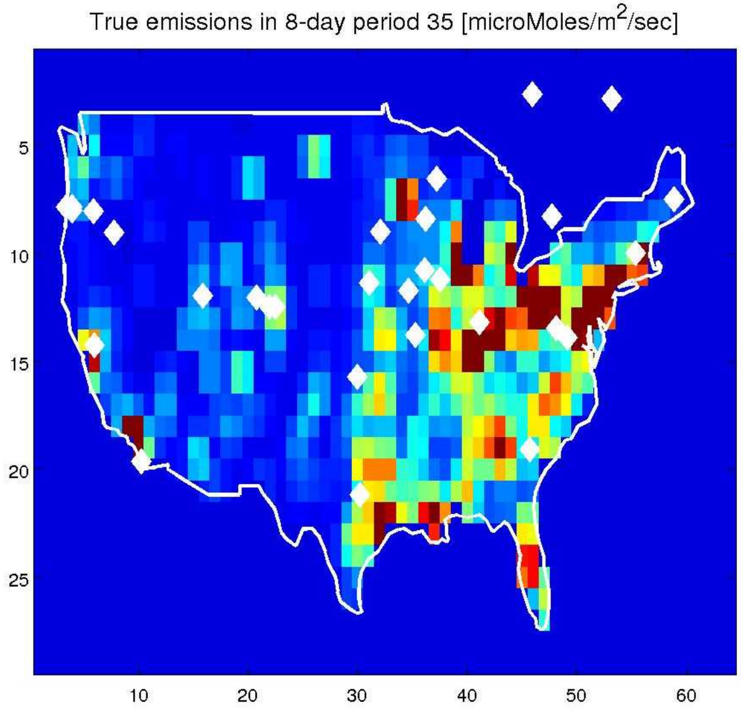
- True emissions – Vulcan database for US, 2002
  - Used to generate CO_2 concentrations at towers
  - 3 hr temporal resolution
- Nightlight images (for 1997) – used to remove wavelets from "dark" areas
- Emissions discretized on a grid
  - 1 degree spatial resolution Fluxes assumed to be constant over 8-day periods ("a week")



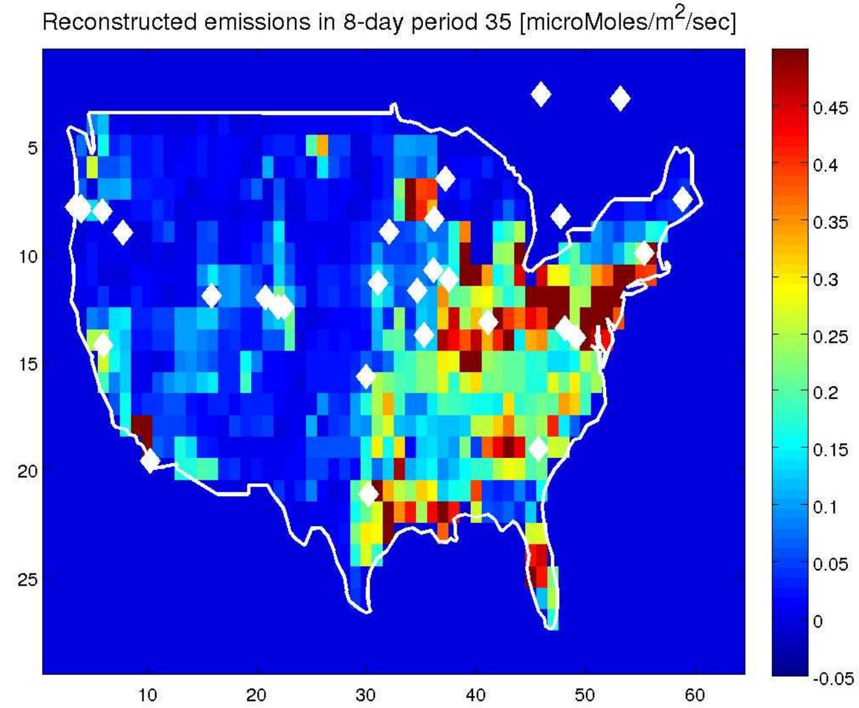
Emissions for a week in August 2002 (Vulcan database, 1 deg resolution)

Sandia National Laboratories

How good is the reconstruction?



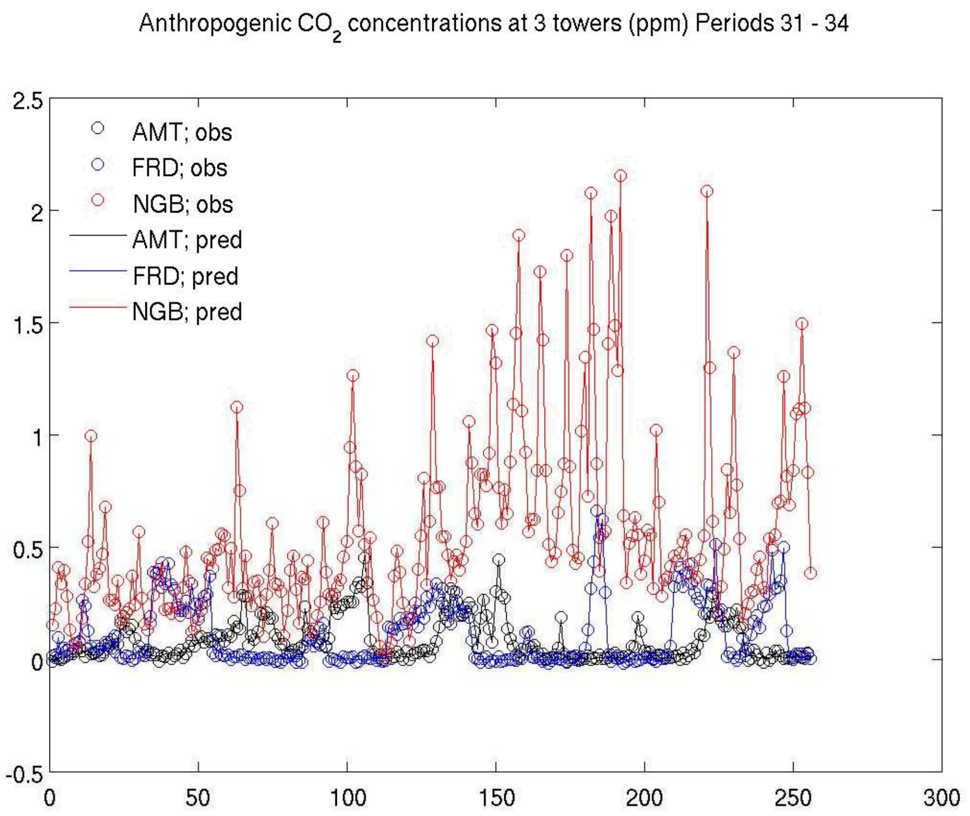
True emissions



Reconstructed emissions

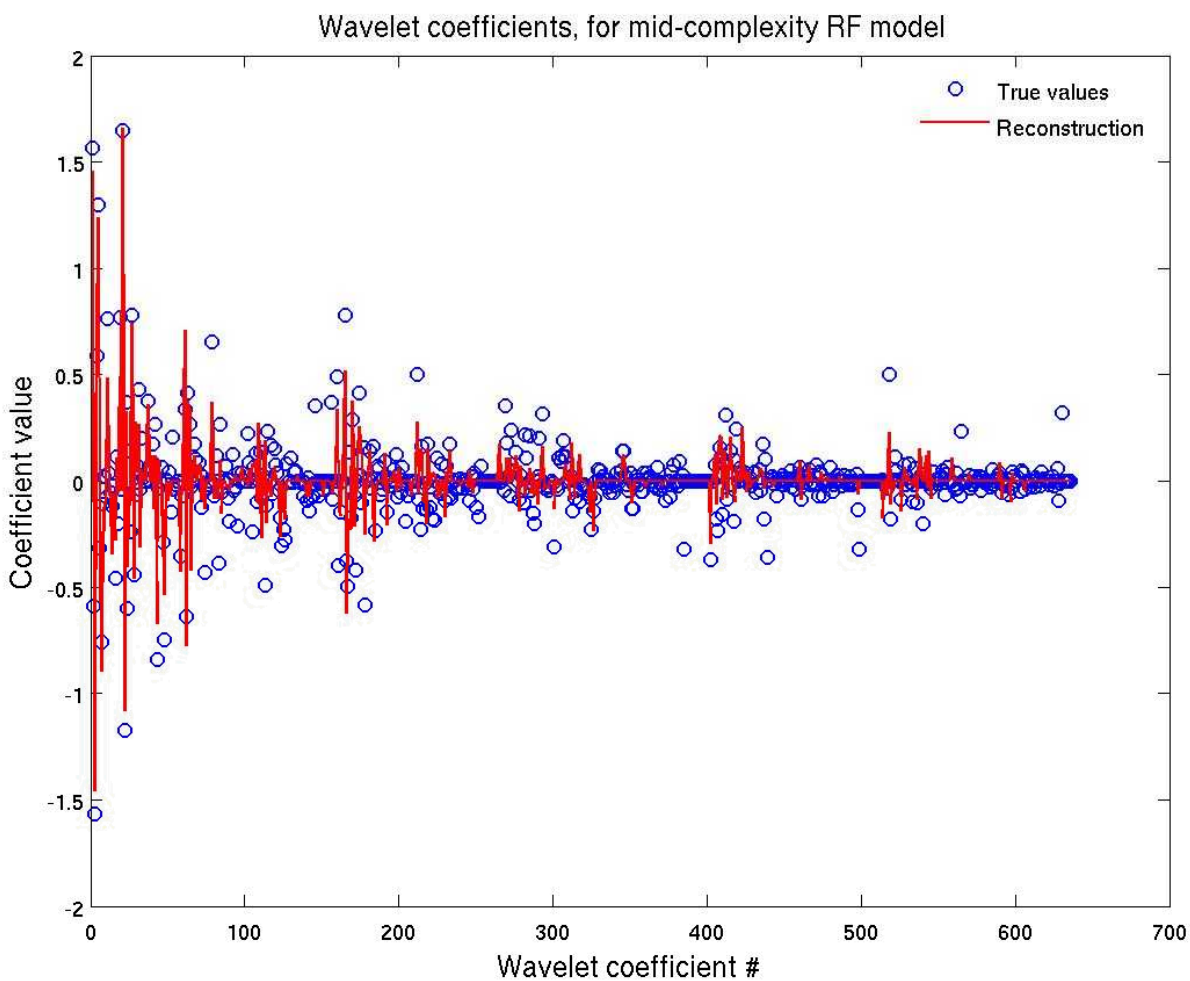
- A week in September 2002

Can we reproduce tower observations?



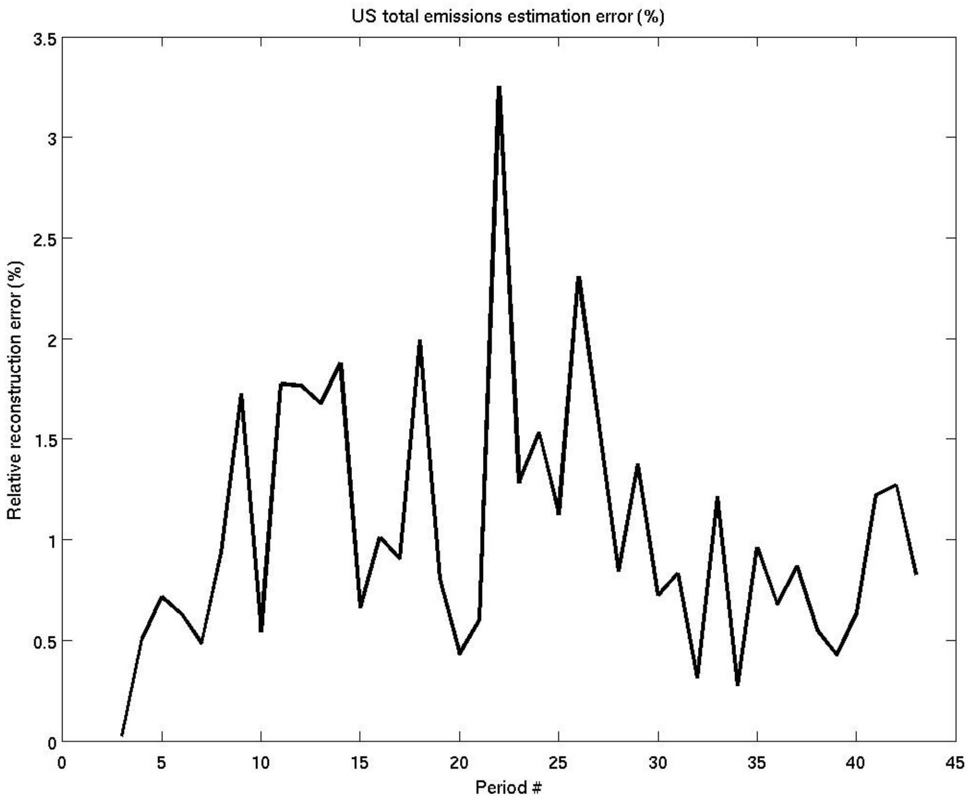
- Tower concentration predictions with reconstructed fluxes (only 3 weeks)
  - Symbols : observations used in the inverse problem.

Did sparsification work?



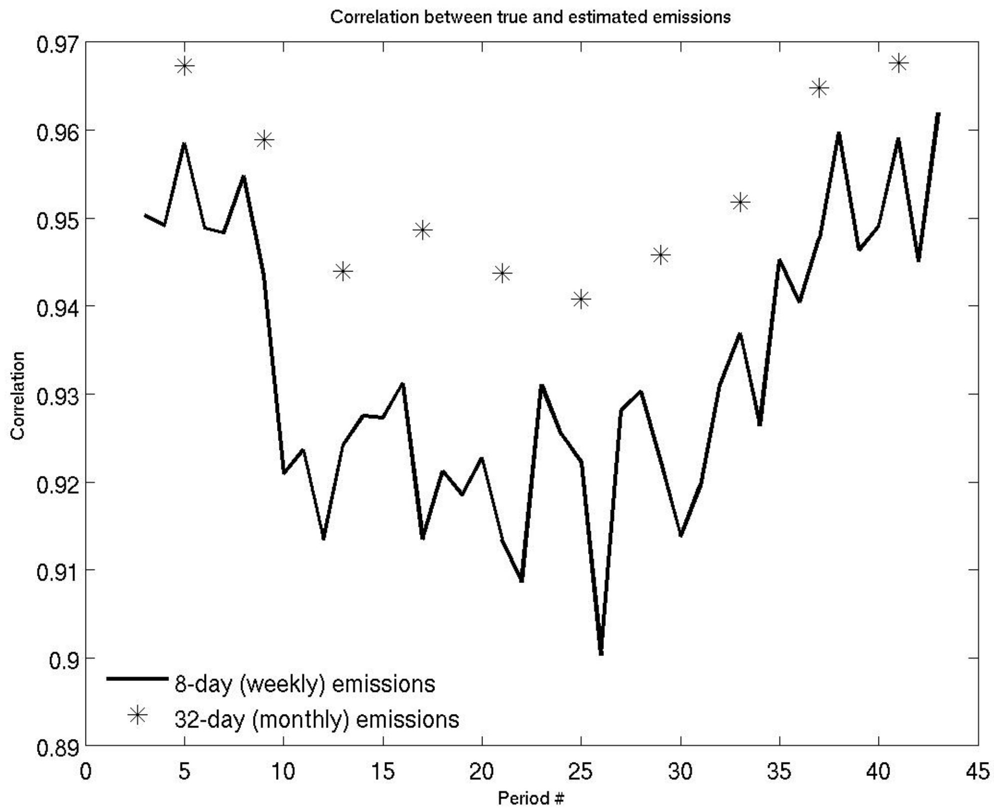
- Only about half the wavelets could be estimated
- We are probably not over-fitting the problem
  - Data-driven sparsification works

Reconstruction error in total US emission



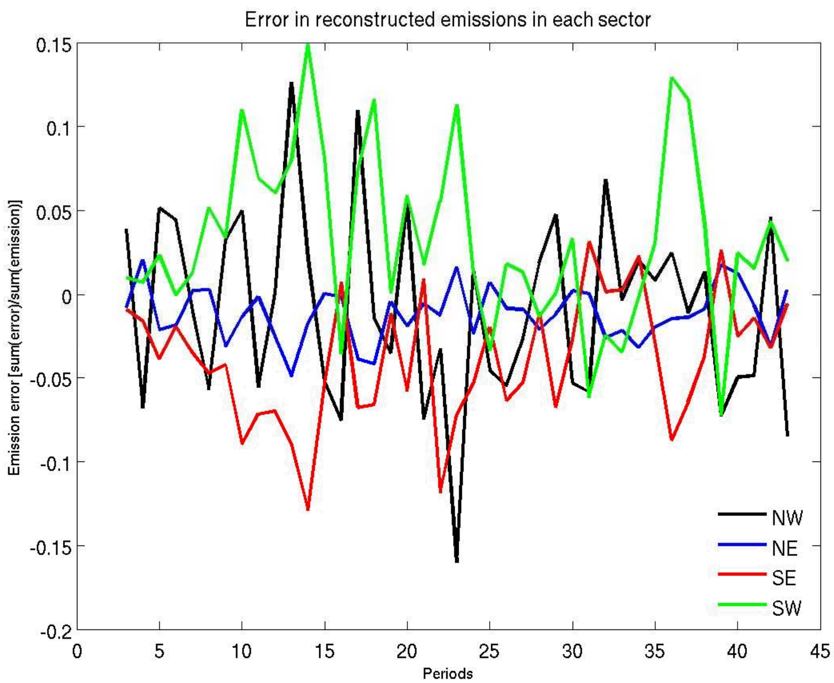
- We get about 3.5% error, worst case

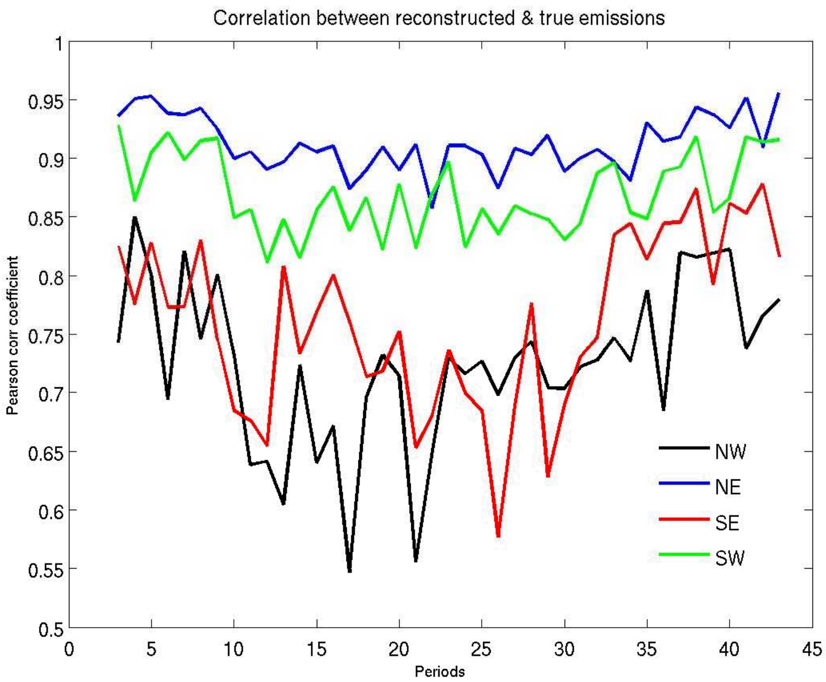
Is the spatial distribution correct?



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

Which parts of US are well estimated?





- The NE has the lowest errors and best correlations
- The NW is generally the worst estimated

Interim conclusions – field estimation

- Sparsity-enforced estimation can deal with high-dimensional spatial random field models
  - Of use when estimating complex, multiscale field
  - For smooth fields, much simpler methods exist
- Not discussed here – non-negativity enforcement
  - The emissions estimated by sparsity enforcement can sometimes be negative
  - A post-processing step (non-sparsity enforcing) corrects it
  - Simple and works only because we start with a very good guess

How far to engineering practice?

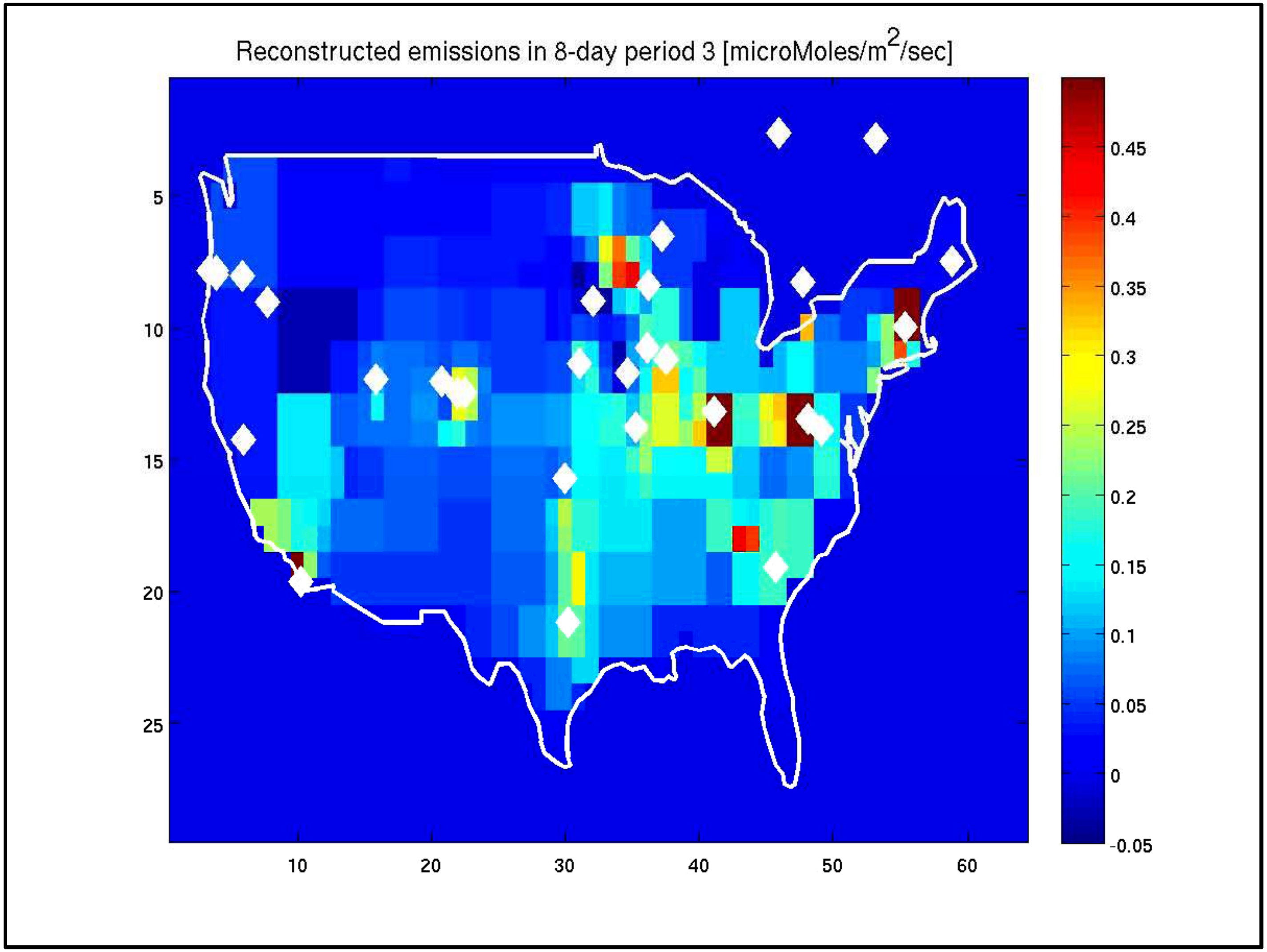
- These are NOT hero HPC codes
  - All done in Matlab and R
  - Sophisticated utilities (wavelets, sparsity-enforced optimization etc.) available as open-source toolboxes and packages
- Largest computational challenge – running ensemble of runs on clusters to generate data for surrogate models
  - Naively, a book-keeping nightmare, but …
  - DAKOTA (http://dakota.sandia.gov) does the sampling, running, batch-job submission and data collation for you
  - Indispensable for $O(10^4)$ runs if $O(10)$ parameters have to be addressed

Sandia National Laboratories

The summing up

- Bayesian inverse problems are close to being used in regular engineering practice
  - Certainly escaped from the math labs into science labs
  - Immense possibilities for quantification of margins and failure-risk estimation
  - Limited to about 10-40 variables
- Sparsity-enforced reconstruction good for field estimation
  - Simplifies / reduces dimensionality of inverse problem, based on info content of observations
  - Can be done probabilistically too (error bars on each grid cell)
    - Called Bayesian compressive sensing / relevance vector machines
    - Can be done for nonlinear problems too

Questions?



Reconstructed emissions in 8-day period 3 [microMoles/m2/sec]
5
10
15
20
25
10
20
30
40
50
60
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-0.05