Wild Seed:

Function Approximation Algorithms for Optimization and Uncertainty Analysis of Multi-modal Computationally Expensive Models with Applications

Christine Shoemaker and Rommel Regis,

Civil and Environmental Engr. &

Operations Research and Industrial Engr.

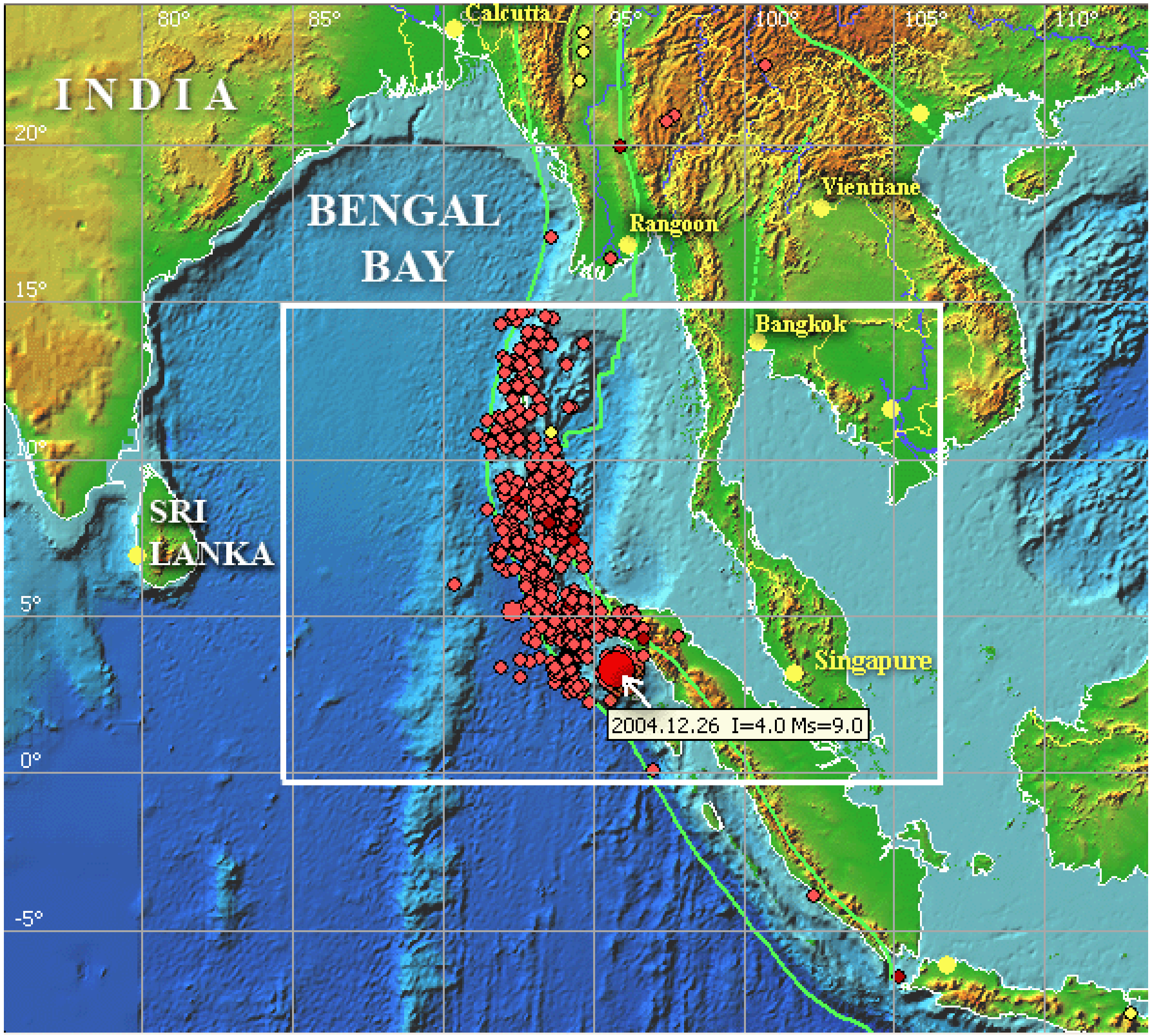
Cornell University

With some results from joint work with others as noted

NSF Workshop on Tsunami

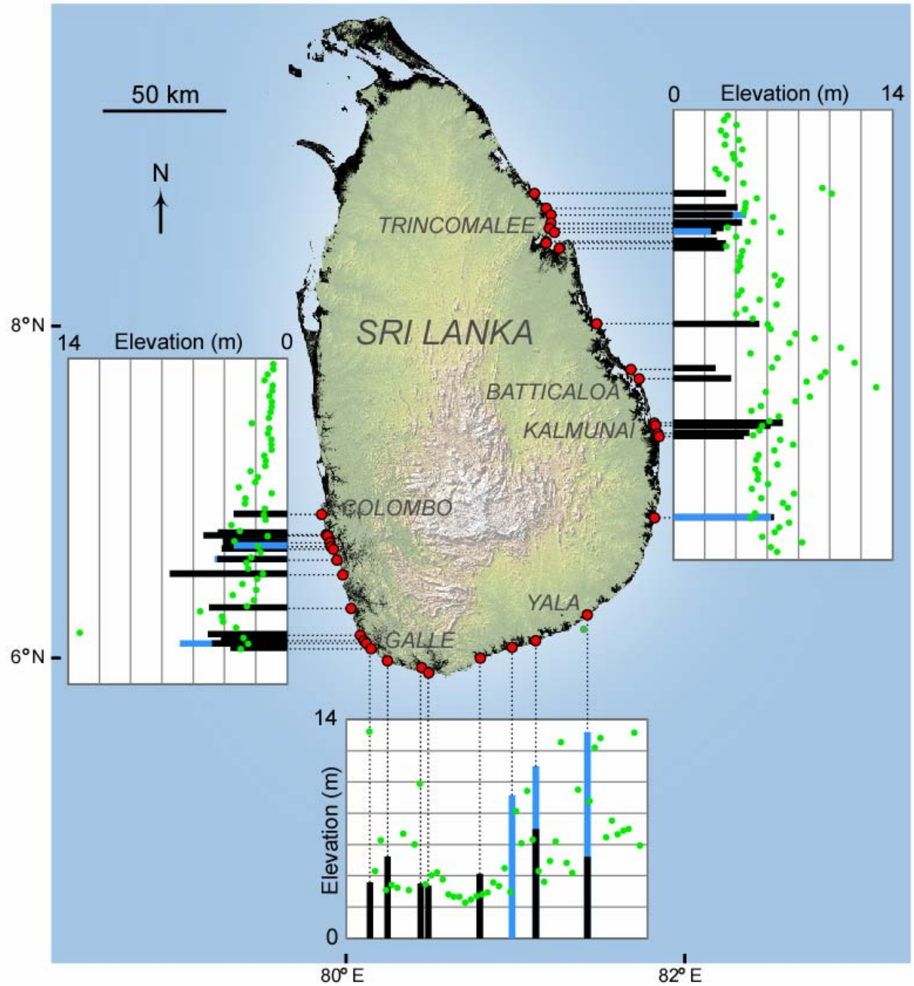
December 26, 2006

Optimization, Statistics, and Uncertainty Analysis can be used to help find best locations for Tsunami Sensors



Comparison between numerical results and field data (II)

Model results depend on Bathymetry



From Liu et al., Science, 2005

Possible Applications of Optimization/Uncertainty Analysis to Tsunami or Other Wave Analysis

1. Estimate spatially distributed bathymetry using limited number of point measurements of water depth plus wave information.
2. Tsunami warning sensors—where to locate tsunami sensors and how many to install
3. Design of breakwaters

Focus of Methods

Nonlinear optimization problem (e.g. bathymetry identification)

- Computationally expensive (costly) simulation models

Optimization for Calibration (e.g for Bathymetry Estimation)

- Our goal is to find the

  minimum of ***f(x)***

  where ***x*** ϵ D

This can be a measure of error between model prediction and observations

X can be parameter values (e.g. bathymetry)

- .
- Let F_{max} be the maximum number of function evaluations (e.g. simulations)
- We want F_{max} to be small because *f(x)* is "costly" to evaluate

Goal: Derivative-Free Optimization of Costly, Black Box, Nonconvex Simulation Models

- For some complex simulations derivatives are unavailable
  - because they cannot be accurately computed sufficiently quickly or
  - because of lack of source code.

Goal: Derivative-Free Optimization of Costly, Black Box, Nonconvex Simulation Models

- Costly functions *f(x)* require a substantial amount of computation (minutes, hours, days) to evaluate the function once.

- Our method seeks to minimize number of costly function evaluations.

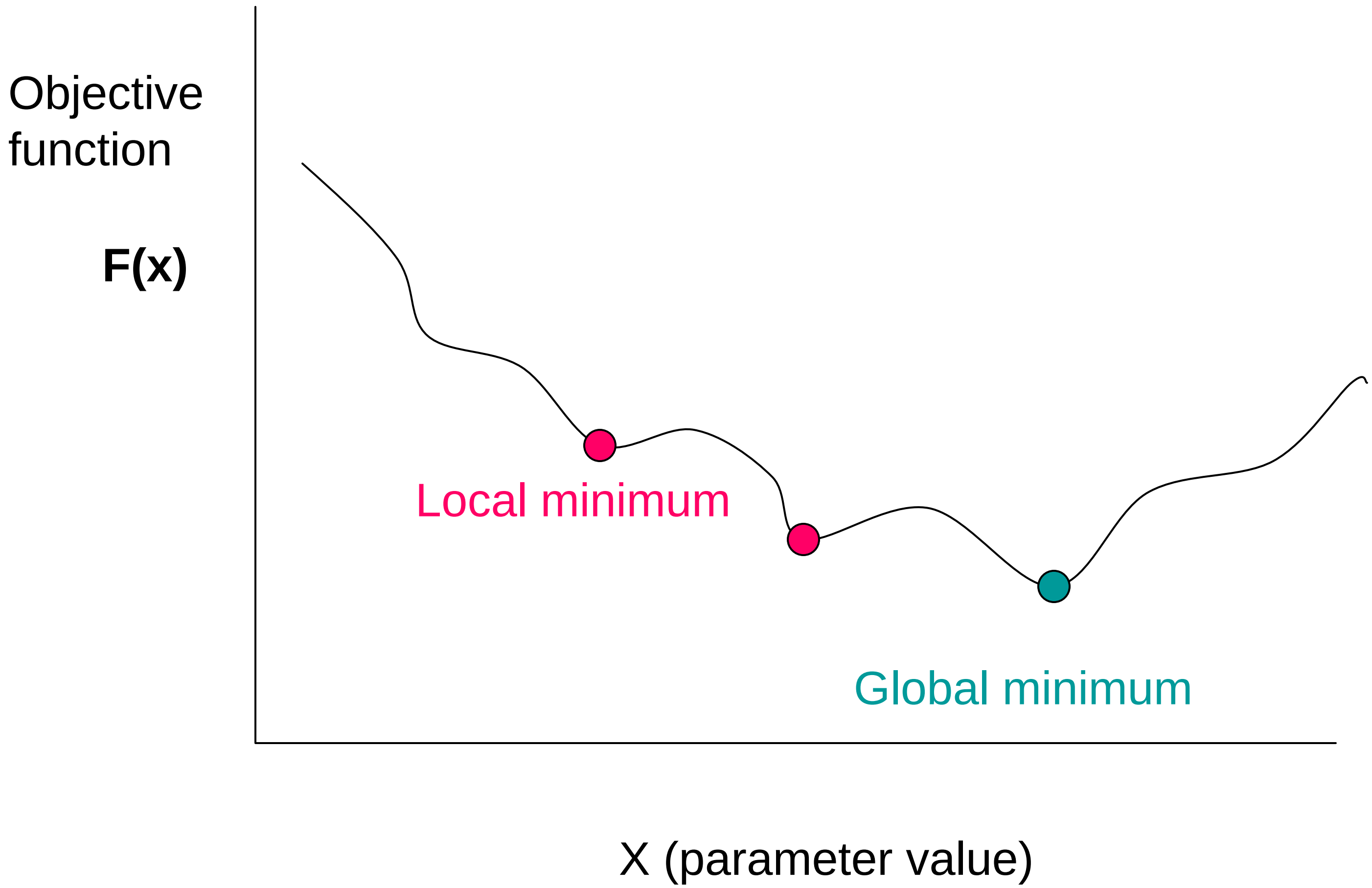
Goal: Derivative-free Optimization of Costly, Black Box, Nonconvex Simulation Models

- Derivative-based optimization methods stop at local minima instead of searching further for the global minimum.

- For black box functions, we don't know if the function is nonconvex or convex.

Global versus Local Minima

Multi-Modal Problems have Multiple local minima



Function Approximation Methods

- A function approximation $R(x)$ to a continuous function $f(x)$ is also called a "response surface model" or a "surrogate model".

- We use radial basis functions for our function approximation, but other methods for non-convex surfaces could also be used.

Why Use Function Approximation Methods?

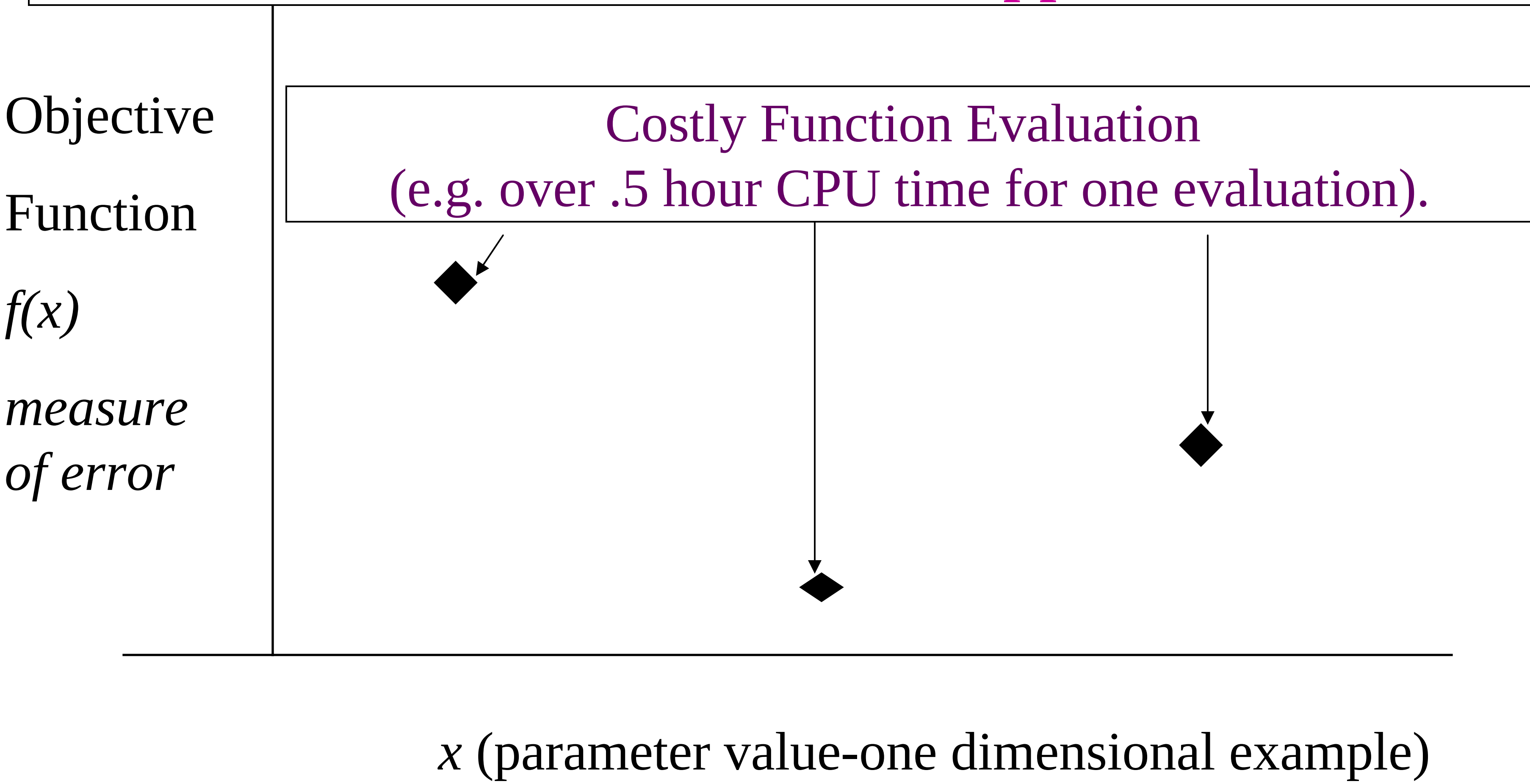
- A function approximation *R(x)* can ***reduce the number of points at which we do a simulation to evaluate f(x),*** and thereby significantly reduce computational cost.

.

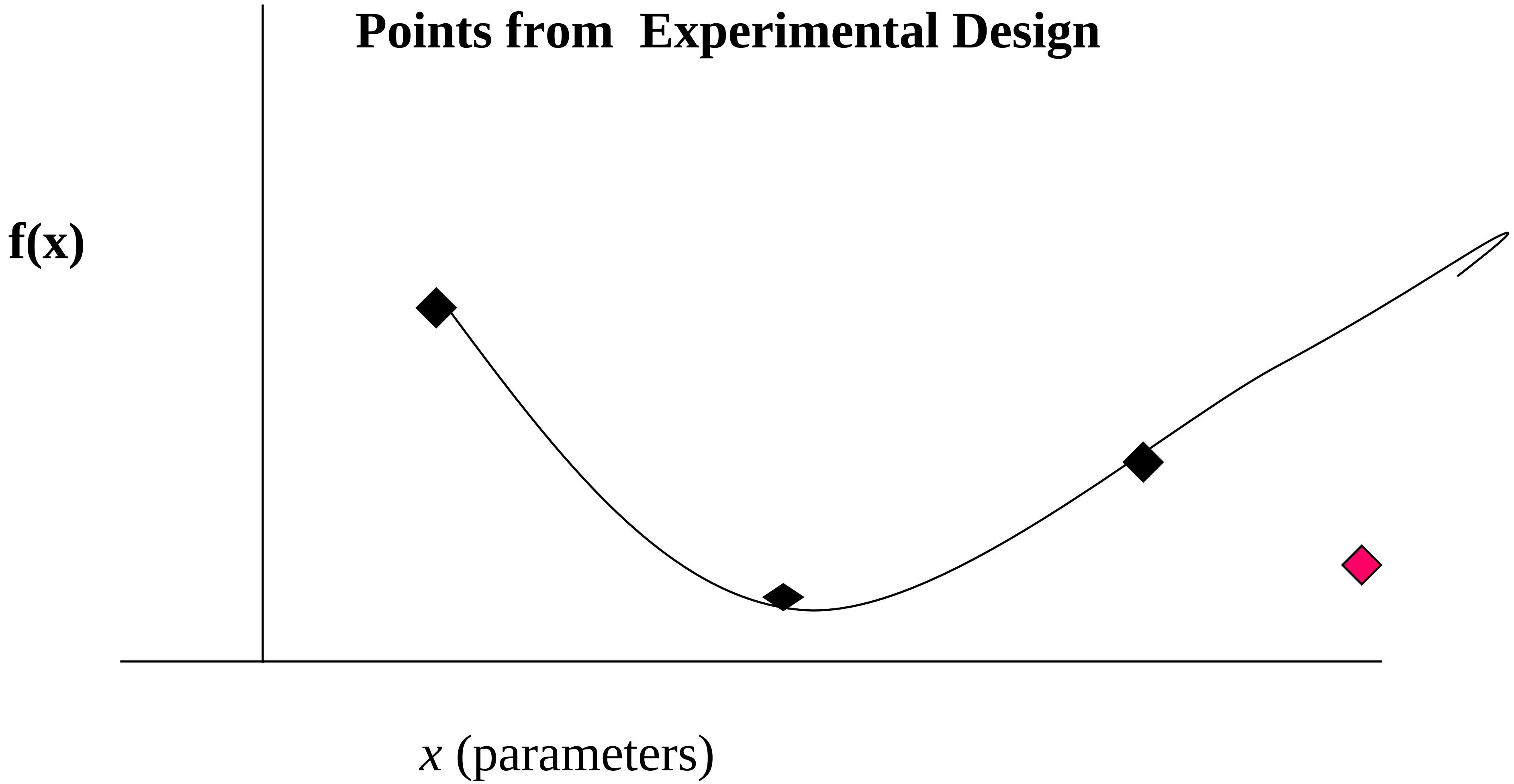
Experimental Design with Symmetric Latin Hypercube (SLHD)

- To fit the first function approximation we need to have evaluated the function at several points.
- We use a symmetric Latin Hypercube (SLHD) to pick these initial points.

One Dimensional Example of Experimental Design to Obtain Initial Function Approximation



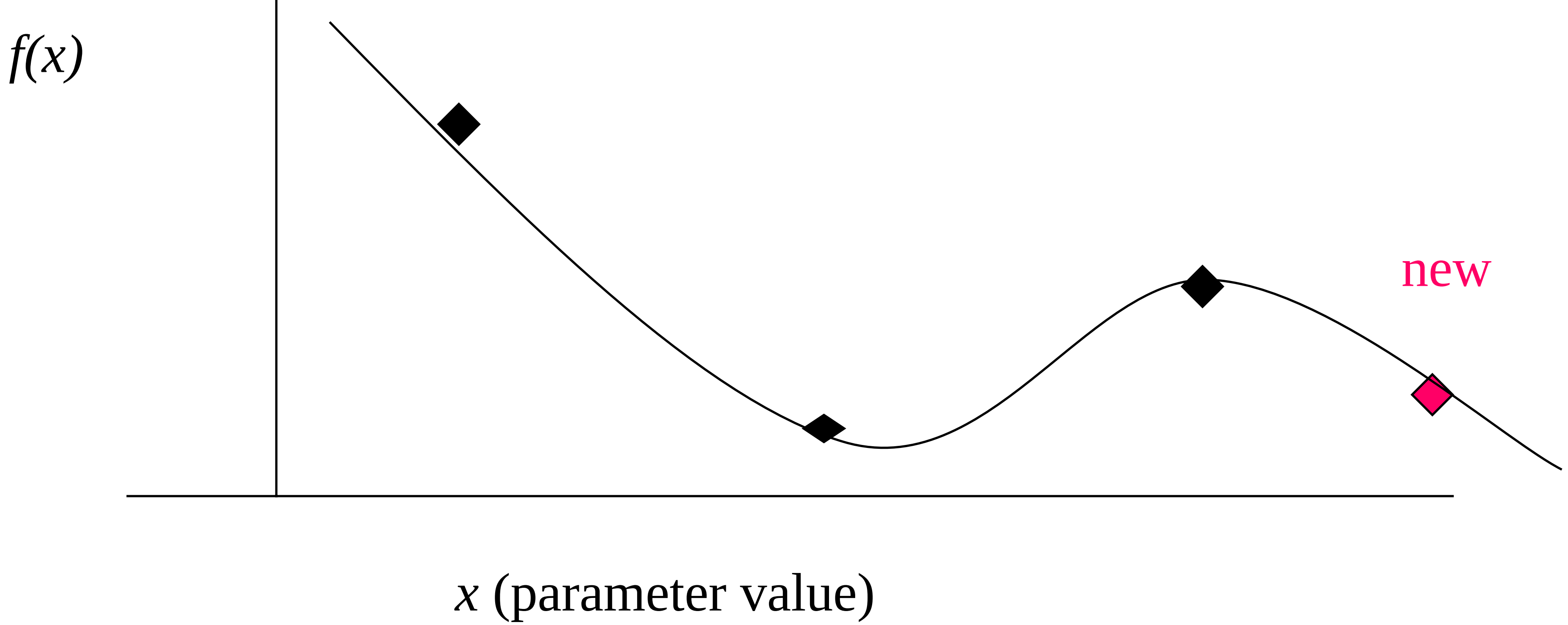
Function Approximation with Initial Points from Experimental Design



In real applications *x* is multidimensional since there are many parameters (e.g. 10).

Update in Function Approximation with New Evaluation

Update done in each iteration for function approximation for each algorithm expert.



Function Approximation is a guess of the function value of *f(x)* for ***all*** *x*.

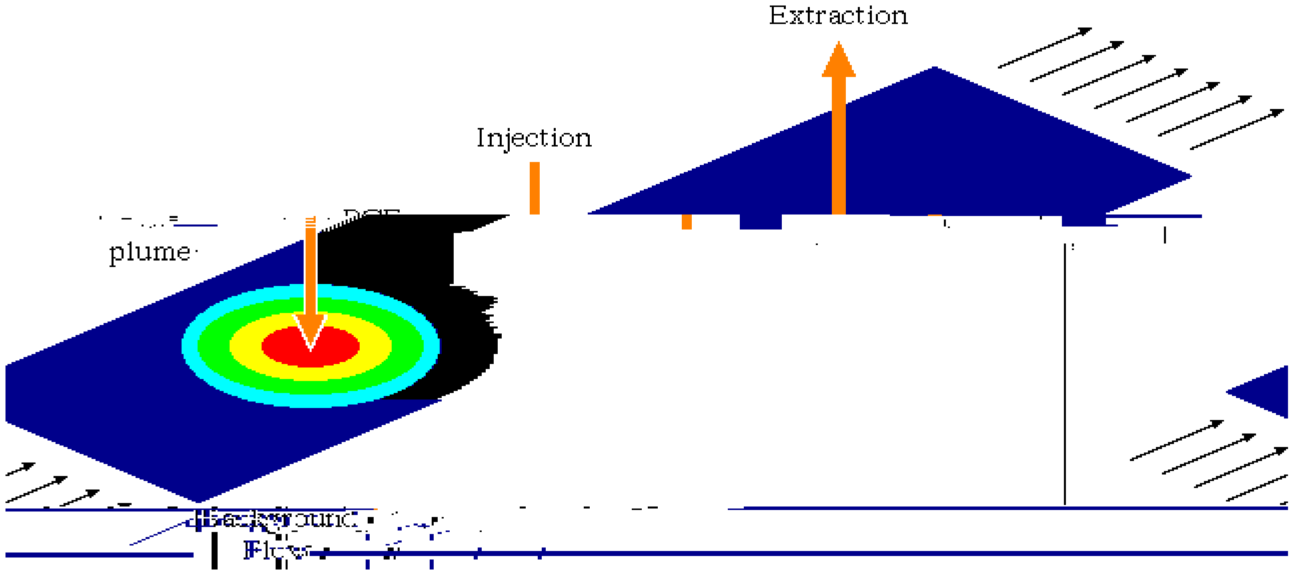
Example Application: Optimization of Calibration of Groundwater Bioremediation Model

Pradeep Mugunthan,
Christine Shoemaker
Rommel Regis

Published in Dec. 2005 in
Water Resources Research

from my NSF grant “Improving Calibration, Sensitivity and Uncertainty Analysis of Data-Based Models of the Environment” from Engineering Directorate

Engineered Bioremediation of Groundwater Contamination by Injection of Hydrogen Donor and Extraction



Injected Donor promotes degradation of chlorinated ethenes by providing hydrogen.

Groundwater decontamination can cost hundreds of millions of dollars at a single site so optimization is important.

Objective Function

- Objective function for calibration to be optimized is
  - sum of squared errors or a related function
  - based on the difference between data and the corresponding model prediction.

Optimization of Calibration of Groundwater Model

- Evaluating the objective function involves numerical solution of a system of partial differential equations by finite difference methods.

- Optimization applied to two cases:
  - hypothetical- at least 8 minutes/simulation
  - field data application- 3 hours/simulation

Algorithms Used for Comparison of Optimization Performance on Calibration

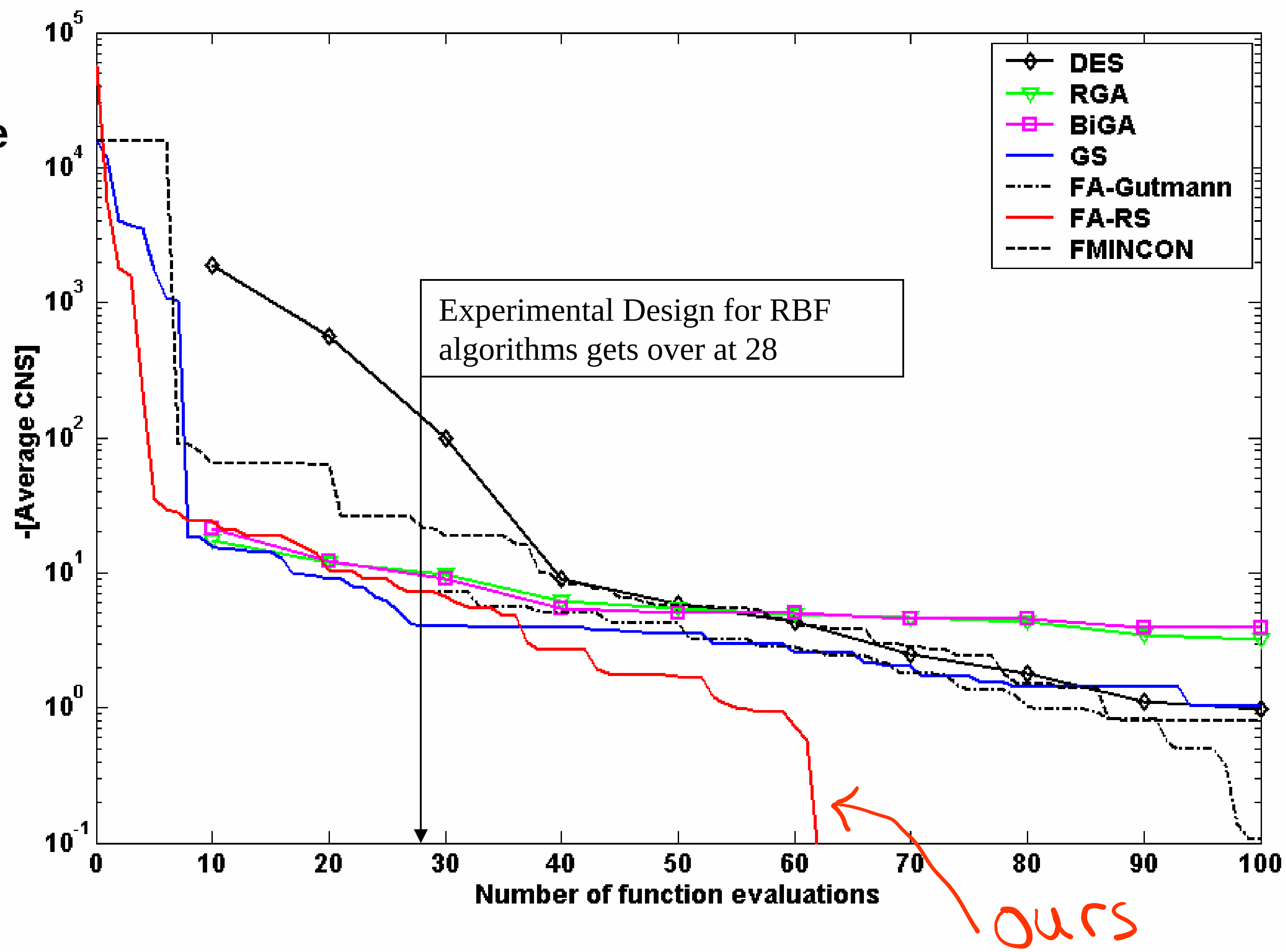
- **Stochastic Greedy Algorithm**
  - Neighborhood defined to make search global
  - Neighbors generated from triangular distribution around current solution. Moves only to a better solution.
- **Evolutionary Algorithms**
  - Derandomized evolution strategy **DES** with lambda = 10 and b_1 = 1/n and $b_2 = 1/n^{0.5}$ (Ostermeier et al. 1992)
  - Binary or Real Genetic algorithm **GA**, population size 10, one point cross-over, mutation probability 0.1, crossover probability 1
- **RBF Function Approximation Algorithms**
  - **RBF Gutmann-** radial basis function approach, with cycle length five, SLH space filling design

    Global Stochastic **RBF-Cornell** radial basis function approach ← ours
- **FMINCON**
  - derivative based optimizer in Matlab with numerical derivatives

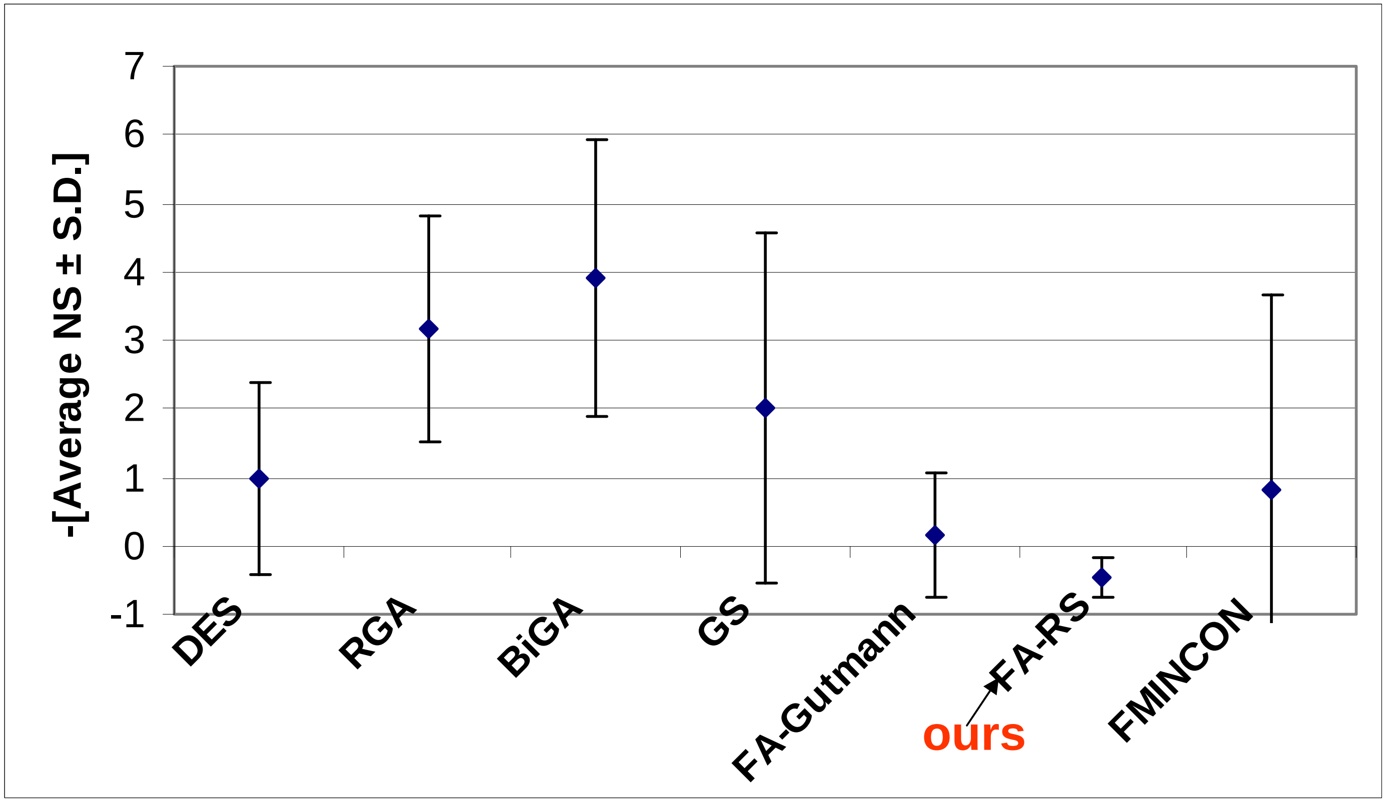
- ***10 trials of 100 function evaluations were performed for heuristic and function approximation algorithms for comparison***

Comparison of Algorithm Performance on Hypothetical Aquifer – CNS

Objective Function



Mean and Standard deviation of best solution produced after 100 function evaluations – Hypothetical Example



•Algorithm with a lowest mean and lowest standard deviation is desirable

•Based on 10 trials

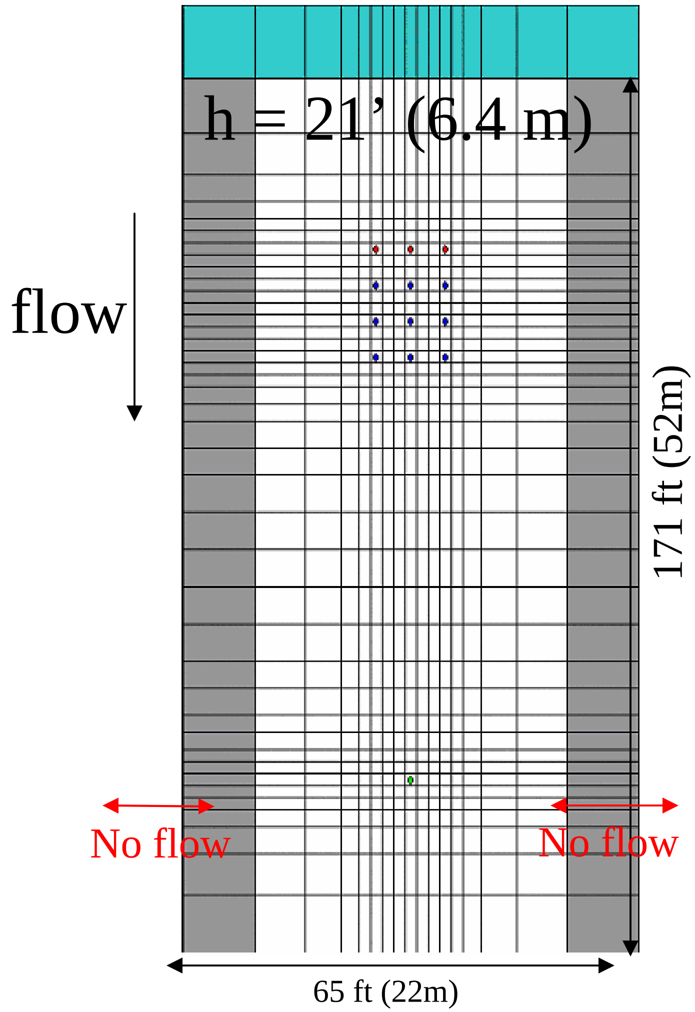
Real Field Site: Alameda Field Data

- The next step was to work with a real field site with DOD data.

- Running the simulation model takes about 2.5 hours for one run of the chlorinated ethene model at this site because of the nonlinearities in the kinetics equations.

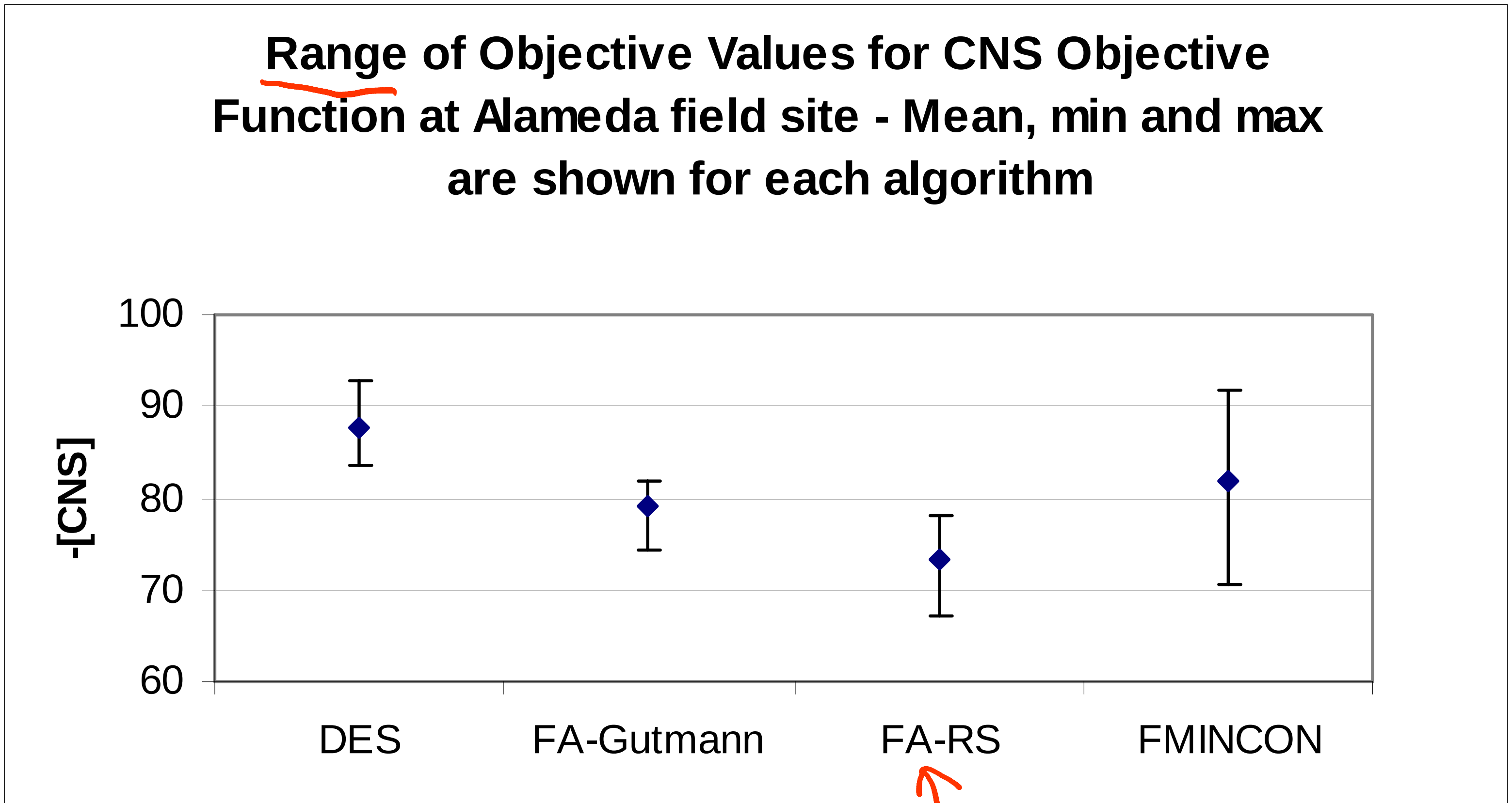
Numerical Set-up for Simulation

- 3-D problem with 8 layers
- Finite difference grid with 17Cx41R
- Constant head of 21' up gradient
- No flux boundaries along columns on either side
- Active boundary down gradient
- About 4800 nodes



Numerical set-up for simulation

Mean, max, and min of best solution produced after 100 function evaluations – CNS



ours

•Based on 3 trials

•Algorithm with a lowest mean and least spread is desirable

Conclusions on Bioremediation Example

- Our function approximation algorithm generally outperformed the alternative algorithms considered.
- This performance was based on a limited number of function evaluations.
- The function approximation algorithm was robust in that it had very few bad results out of 3 or10 trials.

Part II: Uncertainty Analysis

- All models make predictions that are less than perfect.
- Uncertainty analysis seeks to quantify the uncertainty in model predictions
- Examples would include
  - giving confidence limits or
  - giving the probability that the predicited quantity will exceed a threshold

Bayesian Uncertainty Analysis Using Function Approximation for Computationally Expensive Simulation Models

C. Shoemaker and D. Ruppert- PIs
N. Blizniouk (lead author)
R. Rommel, S. Wild, & P. Mugunthan
(submitted paper)

NSF Grant from Mathematical Sciences

Calibration and Uncertainty Analysis

- We are interested in the relationship between data used to calibrate a model and uncertainty.

- Goal of our project is to assess the uncertainty in calibrated parameter values and in model outputs.

- The procedure combines our optimization and function approximation procedures with Bayesian Analysis.

How Do HydrologistsTypically Calculate Uncertainty?

Most widely used method is “GLUE” which:

- Is more computationally demanding than the method we will propose here.
- Is not based on rigorous statistical theory.

Markov Chain Monte Carlo (MCMC)

- This is a statistically rigorous way of computing uncertainty.
- It generates “posterior" multivariate probability density functions (pdf) for each parameter and for model output.
- It requires at least 10,000 simulations to get “convergence” and hence is not feasible for computationally expensive functions.

Our Method for Uncertainty Analysis

- Is based on statistically rigorous theory.
- Is much less computationally demanding than traditional MCMC or GLUE
- Is based on using a function approximation of the likelihood function to do the MCMC.

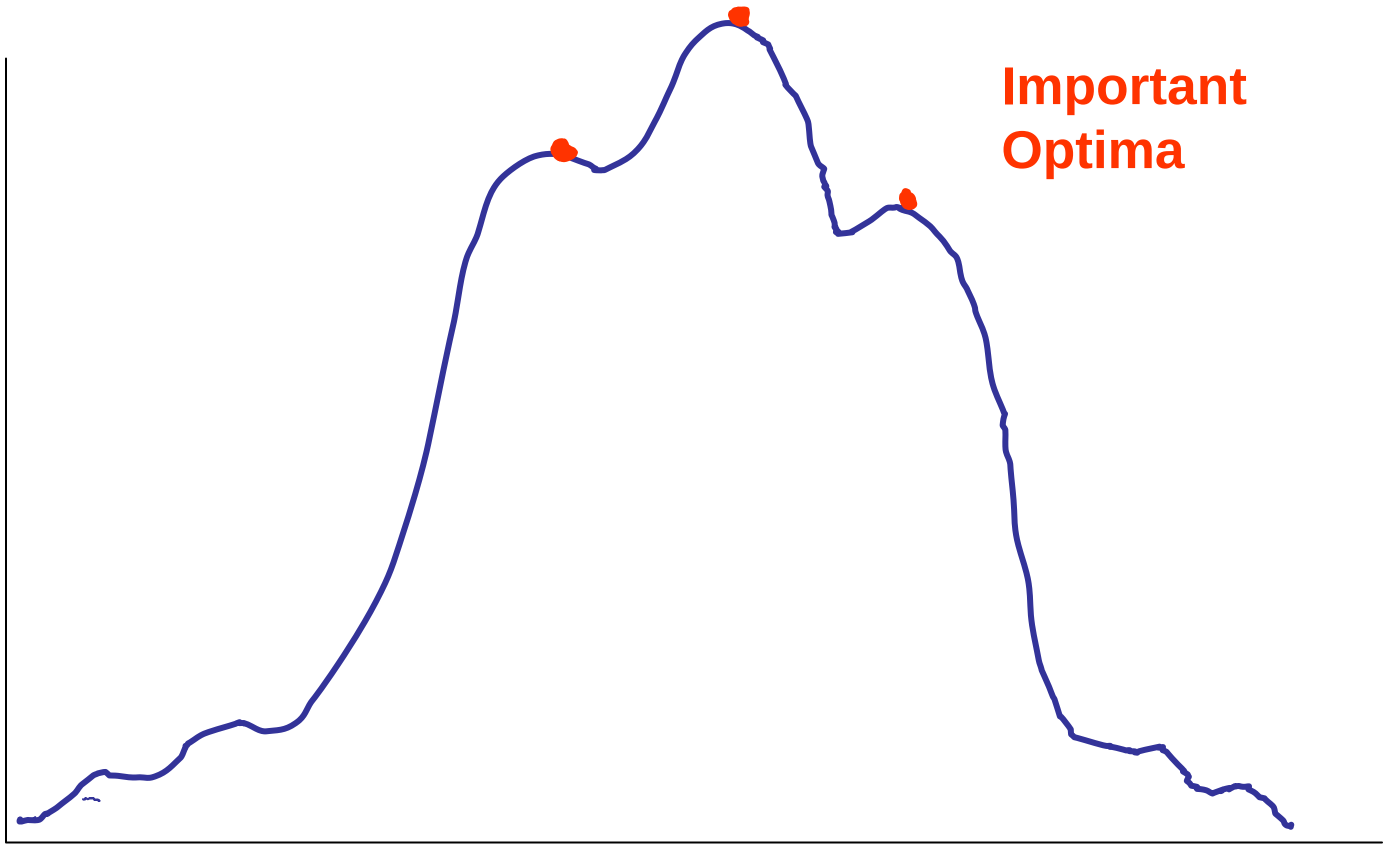
Our Objective Function

- The optimization objective is the likelihood function.
- The likelihood function includes the basic parameters as well as transformations to convert non normal random variables into normal.

We approximate the Likelihood Function

Objective

(Likelihood)



Parameter value (showing one dimension only)

Role of Optimization

We use our optimization search to find the global maxima and important local maxima of likelihood function.

- We build a function approximation of the objective function based on the simulations done in the optimization search plus additional simulations.

Use of Function Approximation in MCMC

- We then apply MCMC (Markov Chain Monte Carlo) to the function approximation of the likelihood function
  - to obtain the joint pdf (probability density function) of the parameter values
  - requires $10{,}000^{+}$ function evaluations of the **function approximation** so takes little computational time

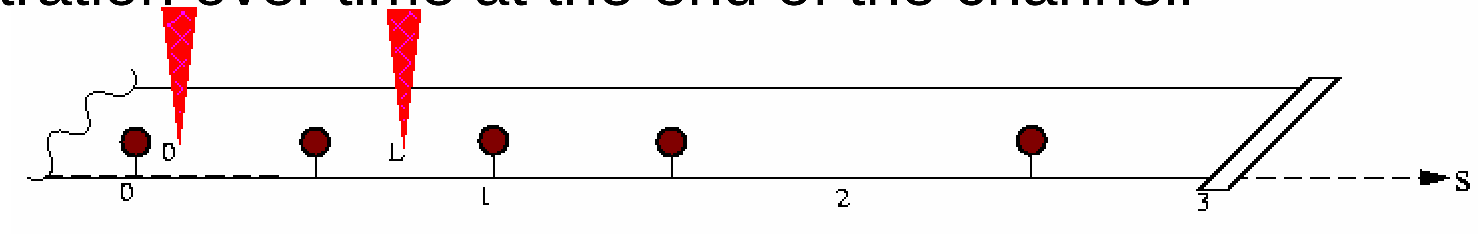
Application: Diffusion of Chemical Spill in Channel

•There is a chemical spill of mass M into a long narrow channel of water at both locations marked in red.

•The system is described by advection diffusion equation.

•We want to estimate the mass **M**, the time **t** and location **L** of the second spill, and the diffusion coefficient **D,** which are the model parameters.

•The **output** from the model we want is average pollutant concentration over time at the end of the channel.



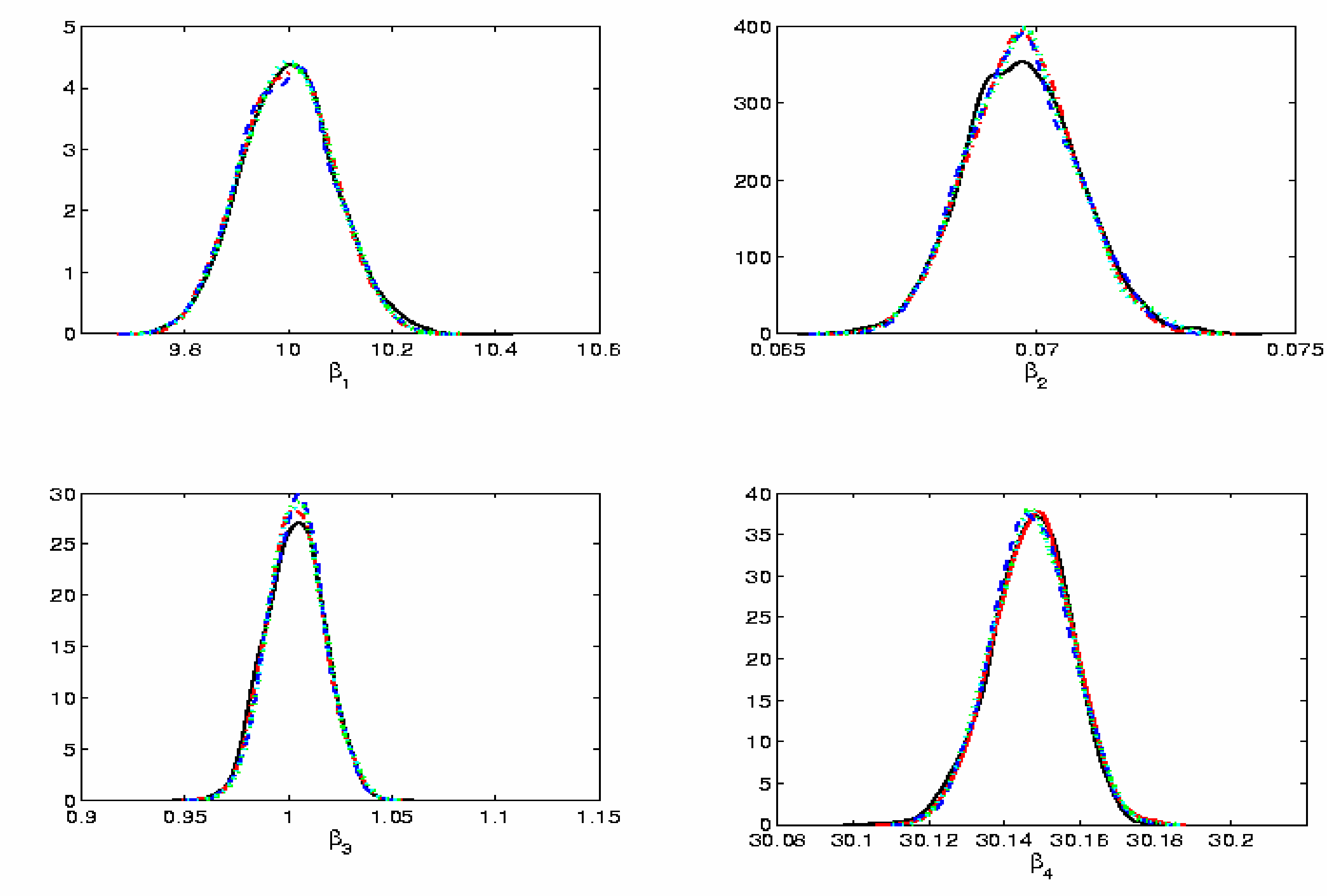
Numerical Results on Chemical Spill Problem

- The following slides show numerical results.
- The solid line is the true value of the marginal pdf of the parameter distribution.
- The colored line shows the marginal pdf obtained by our method (MCMC on the function approximation of the likelihood function).
- There are multiple RBF dashed lines for different approximations.

Estimates of the marginal posterior densities (pdf) obtained by

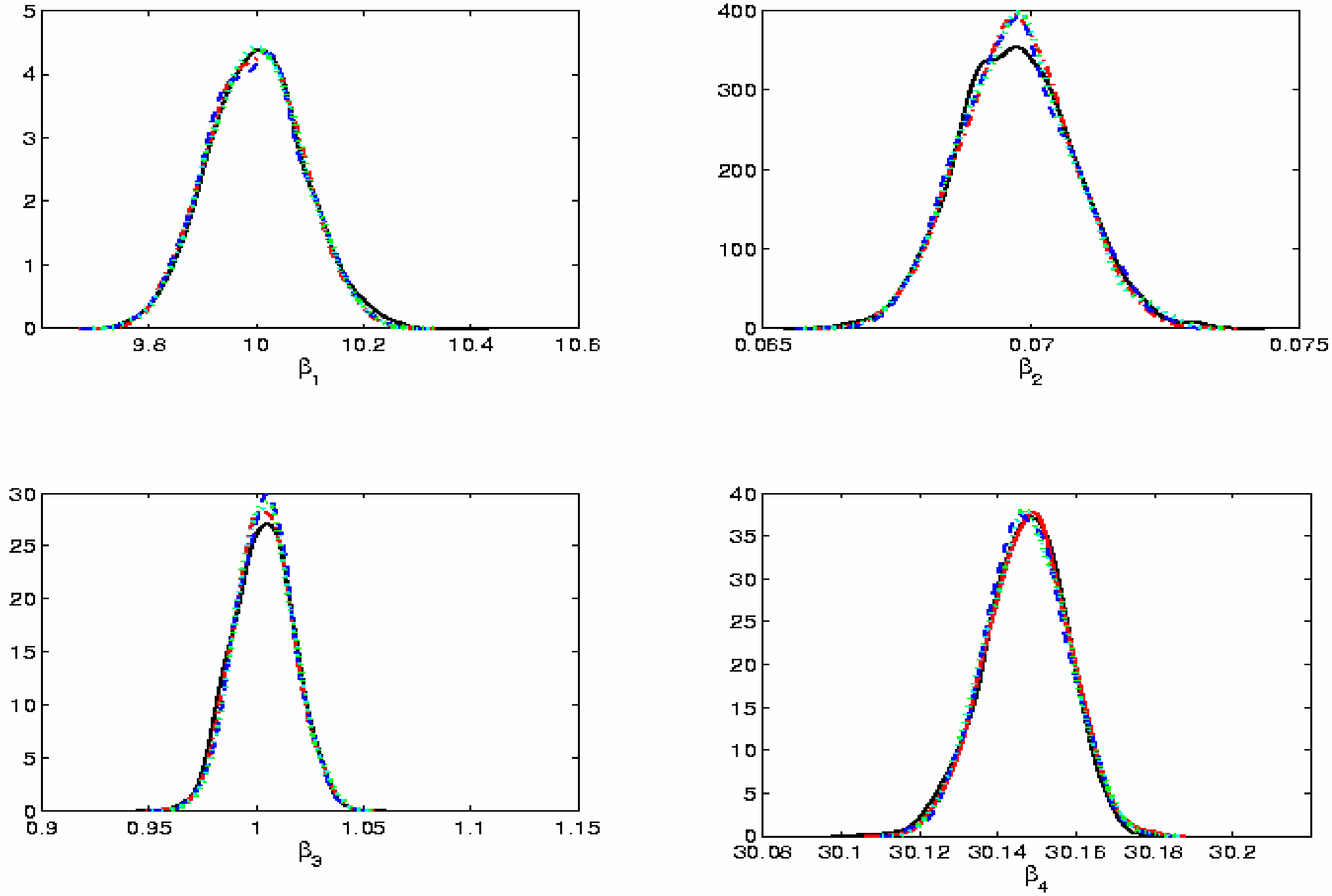
a) (solid line) exact joint posterior obtained from conventional MCMC Analysis with **10,000** function evaluations and

b) (dashed lines) with our function approximation method with **150** function evaluations. *One graph for each parameter*.



Hence we obtained with 150 function evaluations and with function approximation densities that are very similar to those obtained with 10,000 function evaluations and conventional Bayesian statistics which is over a

60 fold reduction in computational demands.

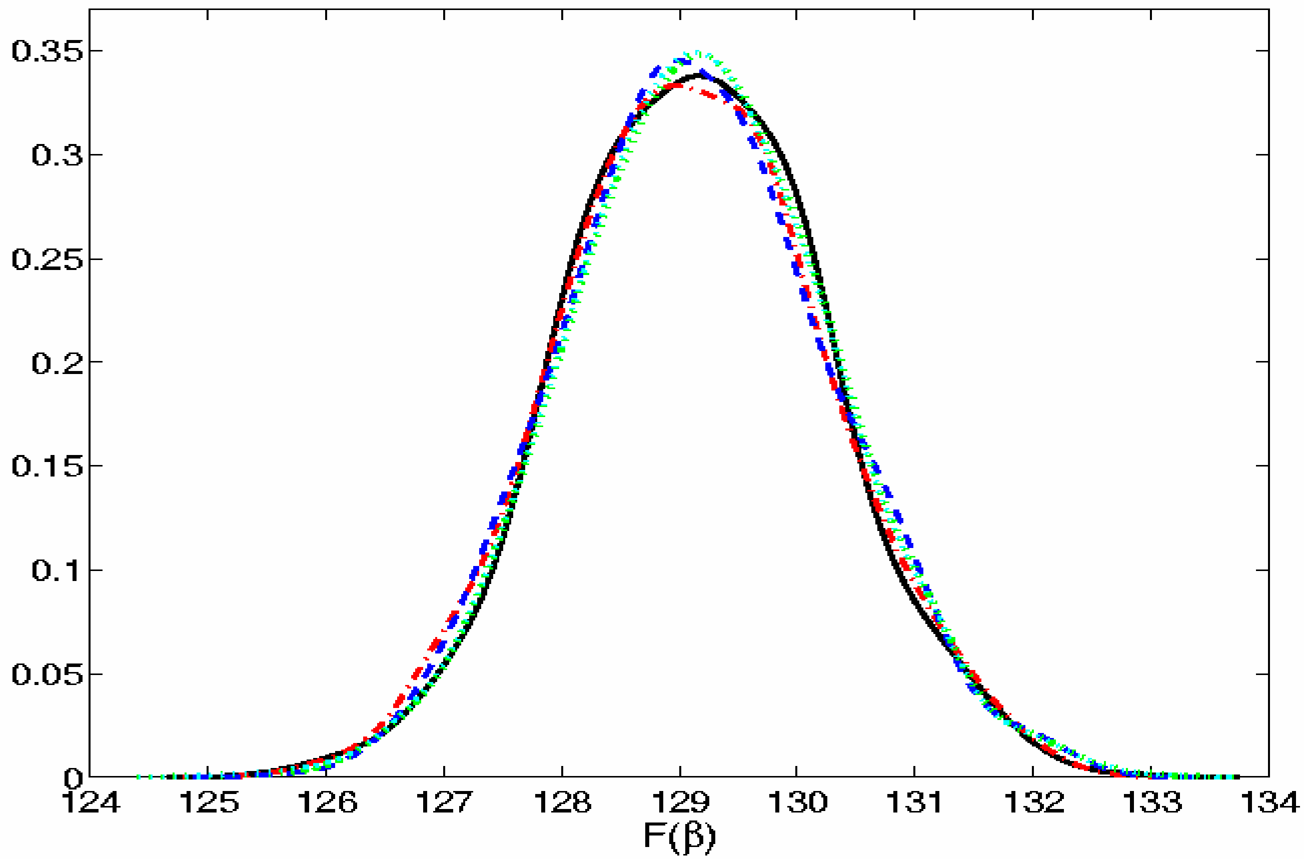


Uncertainty Analysis of Model Output

- Typically what we really want to know is the uncertainty in the model predictions (e.g. **"output")**
- The following shows the comparison of the uncertainty of the prediction of average pollutant concentrations using
  - a) our function approximation method with 150 function evaluations and
  - b) conventional MCMC Bayesian Analysis.

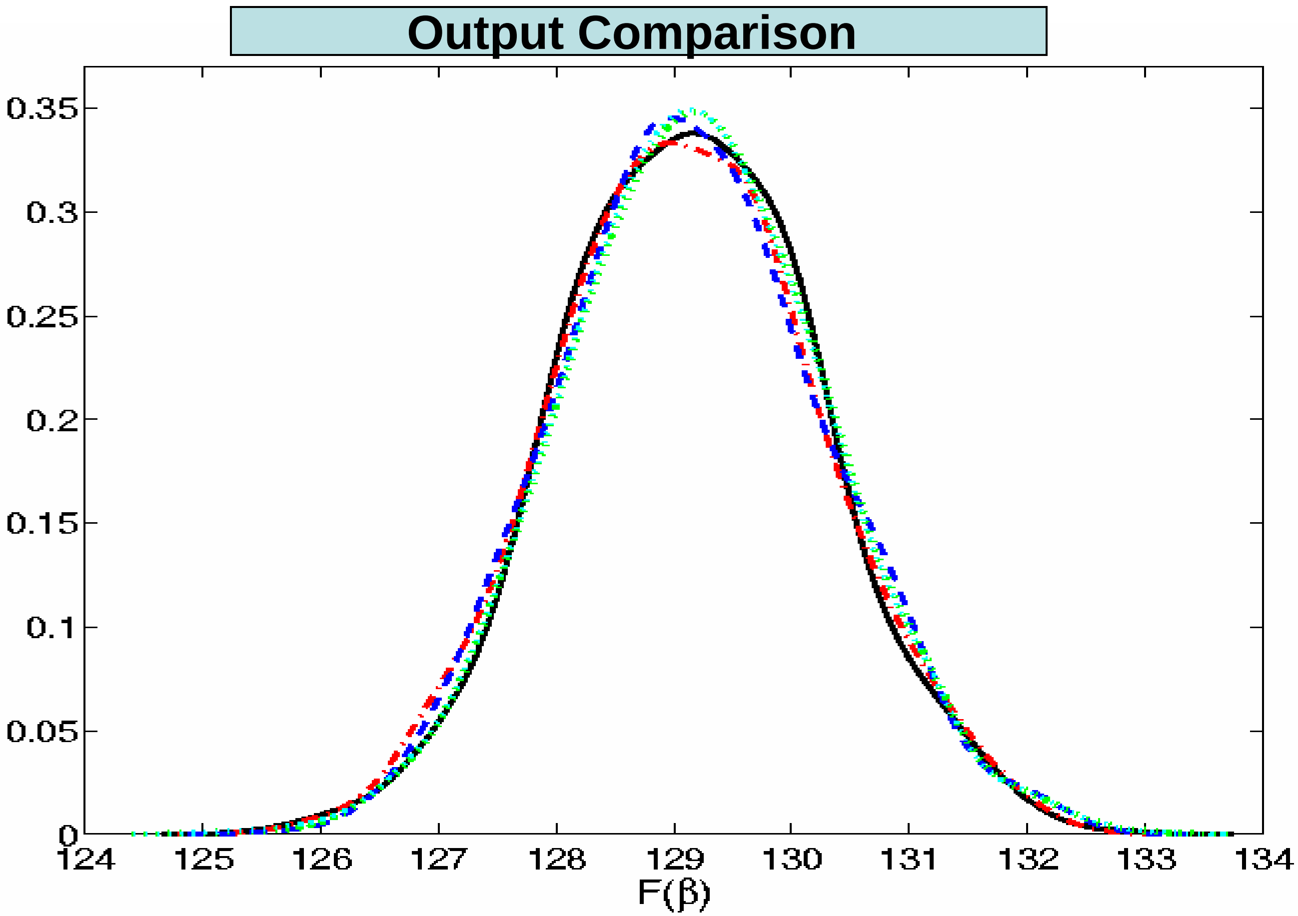
Estimates of the posterior density of the OUTPUT, which is average concentration of the pollutant at fixed location over time .

Comparison



Once you have the pdf, you can compute mean, variance, confidence intervals, probability of exceeding a threshold, etc.

Again we got excellent agreement between our approach with 150 evaluations and the conventional approach with 10,000 evaluations.



Significance of Results

- To obtain the pdf by traditional MCMC requires 10,000 or more “costly function evaluations.
- These results indicated we were able to get good results with much less computational effort by
  - Doing 150 “costly” function evaluations and then fitting a function approximation to the likelihood function and
  - Evaluating 10,000 points on the function approximation surface with the MCMC

What Have We Achieved for Uncertainty Analysis?

- Applied modern statistical tools to calibration of environmental engineering models, including transformations.
- Implemented a Bayesian method of uncertainty analysis
- Substantially reduced (by factor of 60) the number of evaluations of the computationally expensive environmental model.

Overall Talk Summary

- Computationally expensive, multimodal functions are an important class of optimization problems.
- Function Approximation Optimization (FAO) appears to be promising for global optimization of computationally expensive functions.
- FAO can potentially alo8tOlObutears)

Review: Possible Applications of Optimization/Uncertainty Analysis to Tsunami or Other Wave Analysis

1. Estimate spatially distributed bathymetry (optimization for calibration)
2. Tsunami warning sensors —where to locate tsunami sensors and how many to install (design optimization)
3. Design of breakwaters (design optimization)

Selected Recent Papers

- Mugunthan, P., C.A. Shoemaker, R. G. Regis "Comparison of Function Approximation, Heuristic and Derivative-based Methods for Automatic Calibration of Computationally Expensive Groundwater Bioremediation Models," **Water Resources Research** Vol. 41, W11427,doi:10.1029/2005WR004134, Dec. 2005

- Regis, R.G., C.A. Shoemaker, "A Stochastic Radial Basis Function Method for the Global Optimization of Expensive Functions", **INFORMS Journal of Computing**, *in press*

- Regis, Rommel and C. Shoemaker, "Parallel Radial Basis Function Methods for the Global Optimization of Expensive Functions," **European Journal of Operations Research**, *in press*

- Mugunthan, P., C.A. Shoemaker, "Assessing the Impacts of Parameter Uncertainty for Computationally Expensive Groundwater Models," **Water Resources Research**, in press

- Tolson, B. and C.A. Shoemaker, "The Dynamically Dimensioned Search Algorithm for Computationally Efficient Automatic Calibration of Environmental Simulation Models," **Water Resources Research**, in press.

- Bliznioukk, N., D. Ruppert, C. Shoemaker, R. Regis, S. Wild, P. Mugunthan, "Bayesian Calibration of Computationally Expensive Models Using Optimization and Radial Basis Function Approximation," **Computational and Graphical Statistics,** submitted paper

The End