SAND2011-1174C

Ensemble Emulation

Keith Dalbey, PhD

Sandia National Labs, Dept 1441

Optimization & Uncertainty Quantification

Abani K. Patra, PhD

Department of Mechanical & Aerospace Engineering, University at Buffalo

Matthew D. Jones, PhD

Center for Computational Research, University at Buffalo

Eliza S. Calder, PhD

Department of Geology, University at Buffalo

Feb. 28 – Mar. 4, 2011

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under Contract DE-AC04-94AL85000.

Sandia National Laboratories

Outline

- Emulation
- Bayesian Emulators
- Ensemble Emulators
- Test Problem: Volcanic Hazard Map
  - Approach: 3-Level Hierarchical Emulator
  - Results
- Conclusions

Emulation

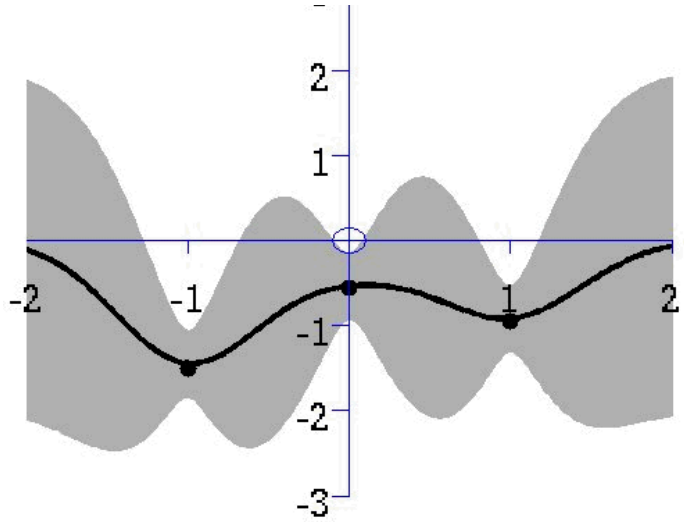
Also known as "**meta-modeling**"

- Process of
  - creating a fast surrogate for simulator or physical system from limited amount of data
  - using surrogate in place of simulator for some purpose (e.g. optimization or uncertainty quantification)
- Can be as simple as a least squares fit
- Can be significantly more complex

Bayesian Emulators

Also known as

- Gaussian Process Emulators
- Bayes Linear Method
- Kriging
- “BLUP” or “BLUE”



Differences among them are minor. All have:

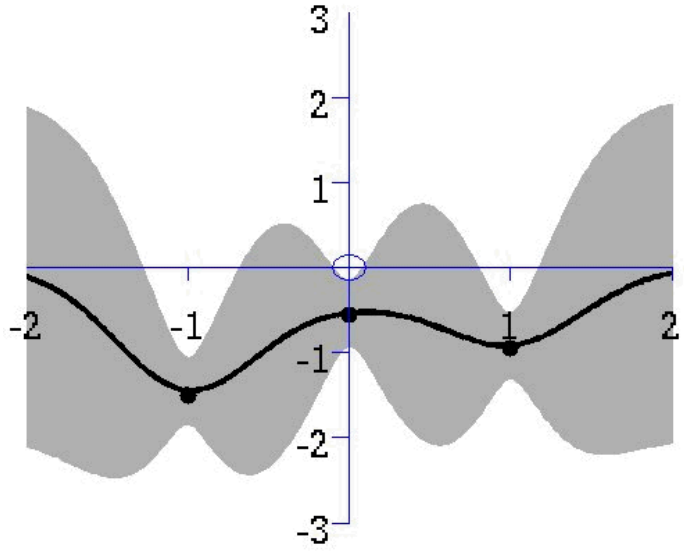
- unadjusted mean (frequently a least squares fit)
- correction/adjustment to mean based on data
- estimated distribution about adjusted mean of possible true surfaces

Sandia National Laboratories

Bayesian Emulators

Also known as

- Gaussian Process Emulators
- Bayes Linear Method
- Kriging
- “BLUP” or “BLUE”

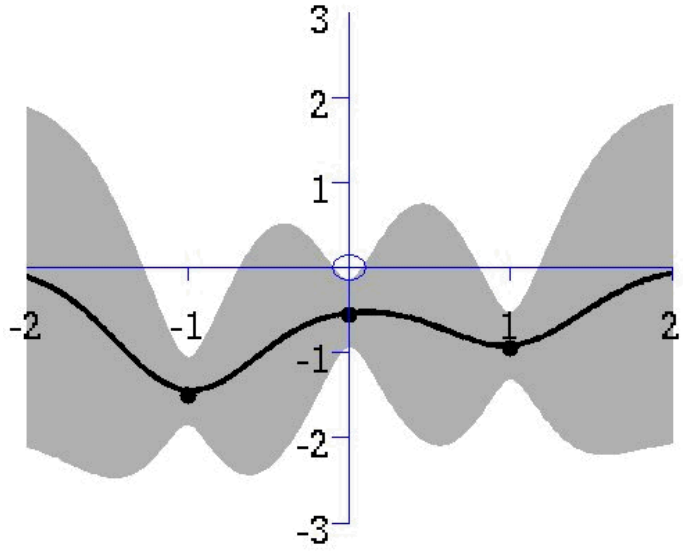


Differences among them are minor & include:

- Choice of “error model” e.g. whether to restrict (“vertical”) distribution about adjusted mean to the normal distribution
- Method of parameter selection

Bayesian Emulators

The equations for the most common formulation are:



$$\mathrm{E}\left[\hat{y}(\underline{x})\middle|\underline{Y}\right] = \underline{g}(\underline{x})^T\,\underline{\beta} + \underline{r}(\underline{x})^T\,\underline{\underline{R}}^{-1}\left(\underline{Y} - \underline{\underline{G}}\,\underline{\beta}\right)$$

$$\mathrm{Var}\left[\hat{y}(\underline{x})\middle|\underline{Y}\right] = \hat{\sigma}^2\left(1 - \underline{r}(\underline{x})^T\,\underline{\underline{R}}^{-1}\underline{r}(\underline{x}) + \ldots\right.$$

$$\left.\left(\underline{g}(\underline{x})^T - \underline{r}(\underline{x})^T\,\underline{\underline{R}}^{-1}\underline{\underline{G}}\right)\left(\underline{\underline{G}}^T\,\underline{\underline{R}}^{-1}\underline{\underline{G}}\right)^{-1}\left(\underline{g}(\underline{x})^T - \underline{r}(\underline{x})^T\,\underline{\underline{R}}^{-1}\underline{\underline{G}}\right)^T\right)$$

$$r_i(\underline{x}) = r\left(\underline{x}, \underline{X}_i\right) = \exp\left(-\sum_{d=1}^{D}\theta_d\left(x_d - X_{i,d}\right)^2\right)$$

$$R_{i,j} = R_{j,i} = r\left(\underline{X}_i, \underline{X}_j\right) \qquad G_{i,j} = g_j\left(\underline{X}_i\right)$$

Sandia National Laboratories

Bayesian Emulator Parameter Selection

- Always involves repeated inversion of error model's "correlation matrix," $\underline{\underline{R}}$
- $\underline{\underline{R}}$ is an N x N matrix, where N is the number of data points
- Requirement of matrix inversion restricts emulators to small amounts of data because, for "Large" N:
  - $\underline{\underline{R}}$ is poorly conditioned (numerically singular)
  - Cost of inverting matrix is $\mathcal{O}(N^3)$ operations

Ensemble Emulation[1,2]

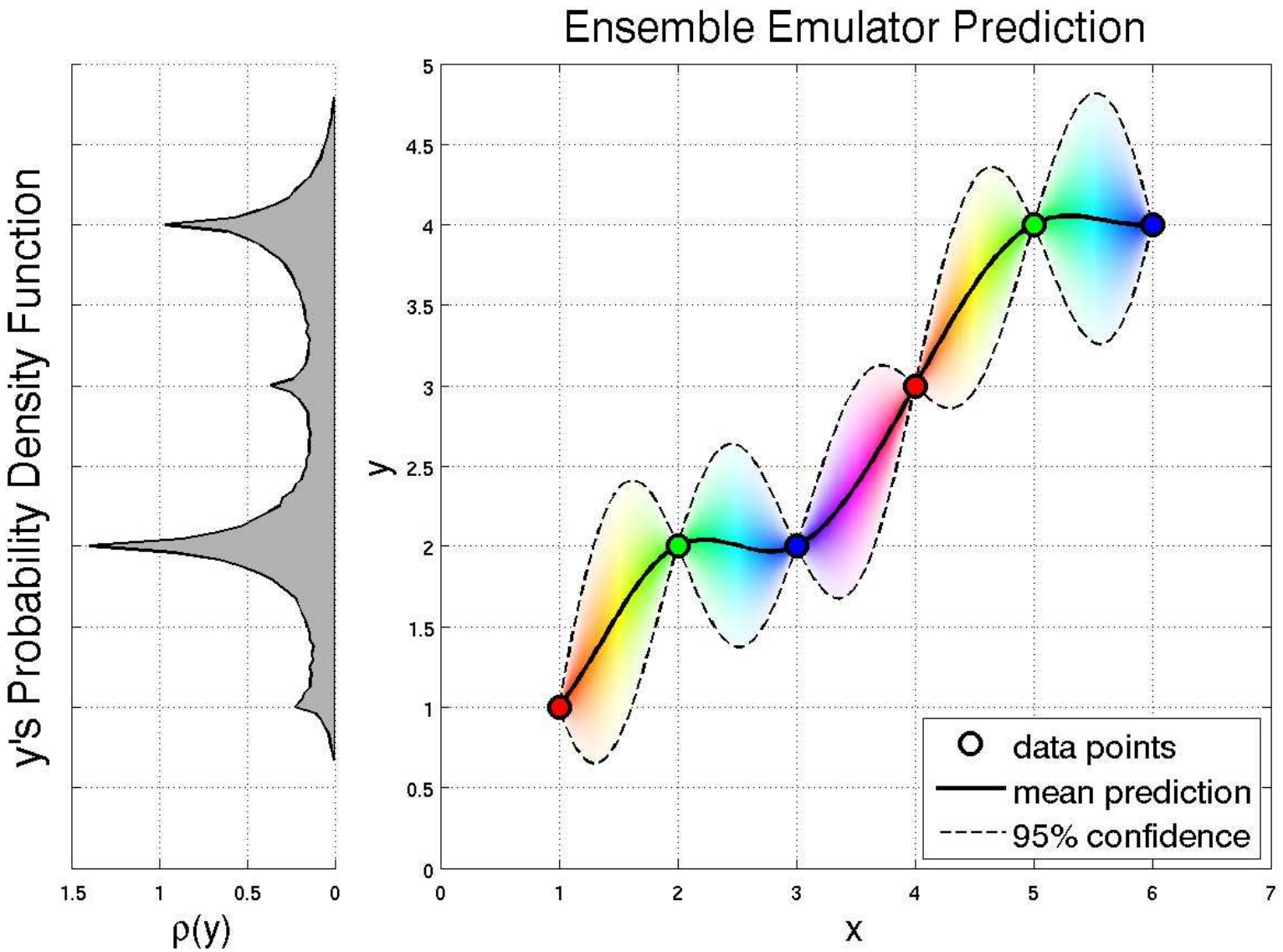
- Uses an ensemble of many small component emulators instead of 1 large emulator
- Component emulators use small subsets of data

Benefits:

- Avoids problem of ill conditioning
- Can greatly reduce computational cost
- Allows concurrent construction & concurrent evaluation of component emulators
- Macro emulator is non-stationary

1. Gramacy et al 2004 2. Dalbey, PhD 2009

Ensemble Emulation: 1D Example

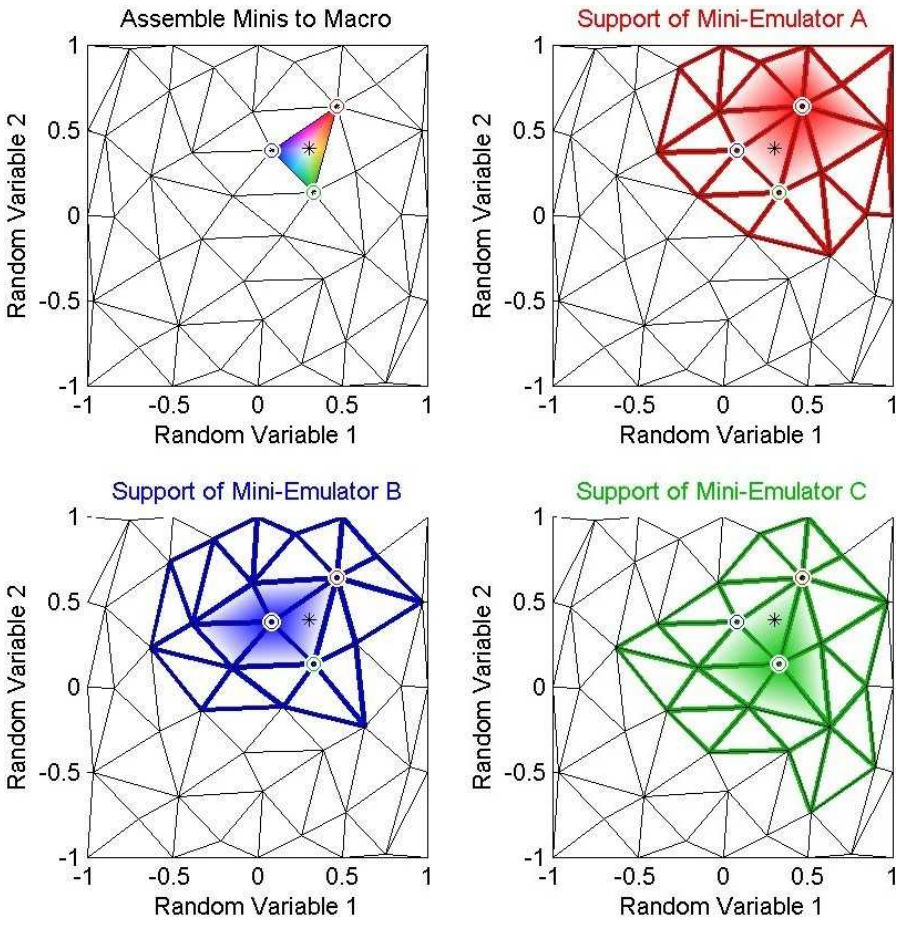


Sandia National Laboratories

Ensemble Emulation

$\mathcal{O}(N^3) \rightarrow \mathcal{O}(N\ M^3)$

- Tessellate sample inputs & generate 2 hop neighborhood for each sample
- Concurrently build mini-emulator for each sample's 2 hop neighborhood
- Concurrently evaluate mini-emulator nodes of "triangles" containing re-sample points
- The (non-stationary) macro-emulator's output is the weighted (by barycentric coordinates) sum of mini-emulator outputs



Sandia National Laboratories

Test Problem: Volcanic Hazard Map

Objective: in <24 hours use 1024 processors to generate map of probability that a (volcanic landslide) hazard criteria will be exceeded within 10 years for the island of Montserrat.

2 uncertain input dimensions (volcanic flow volume and preferred initial direction)

\+

2 spatial dimensions (East, North)

= 4 input dimensions

Needs hundreds to thousands of simulations; each will produce a field variable ($\mathcal{O}$(10^5) data points) as output.

Each simulation takes $\mathcal{O}$(10) processor hours

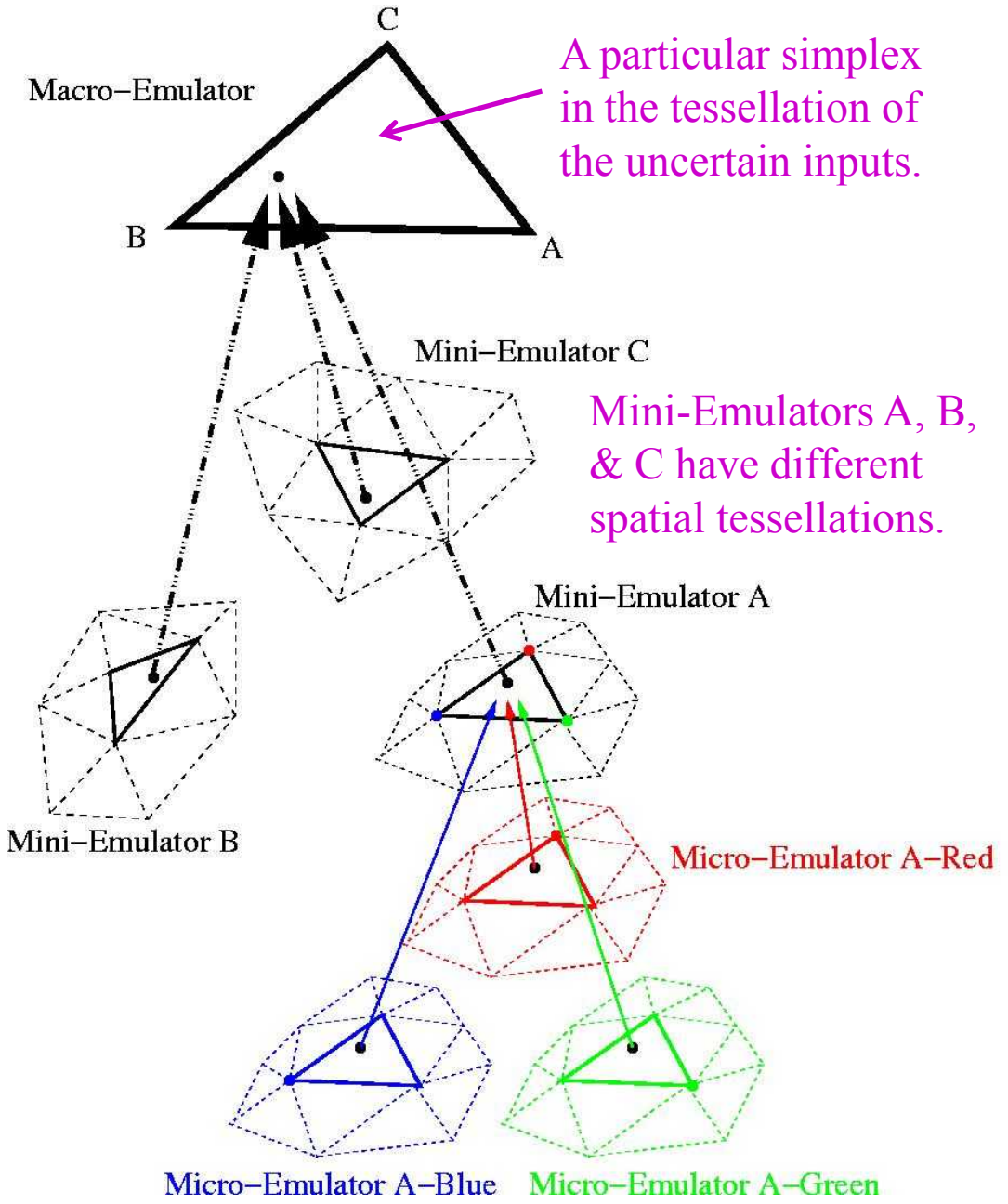
Sandia National Laboratories

Approach

- Used “top down” 3-level hierarchical ensemble emulator
- Replaced global N-by-N $\underline{\underline{R}}$ matrix with N local M-by-M $\underline{\underline{R}}$ matrices, N is in millions, M is $\mathcal{O}(100)$ … This reduced cost from $\mathcal{O}(N^3)$ to $\mathcal{O}(N\,M^3)$
- Distributed work to nodes of supercomputer
- Generated hazard map in under 9 hours using 1024 processors; goal was 24 hours

Sandia National Laboratories

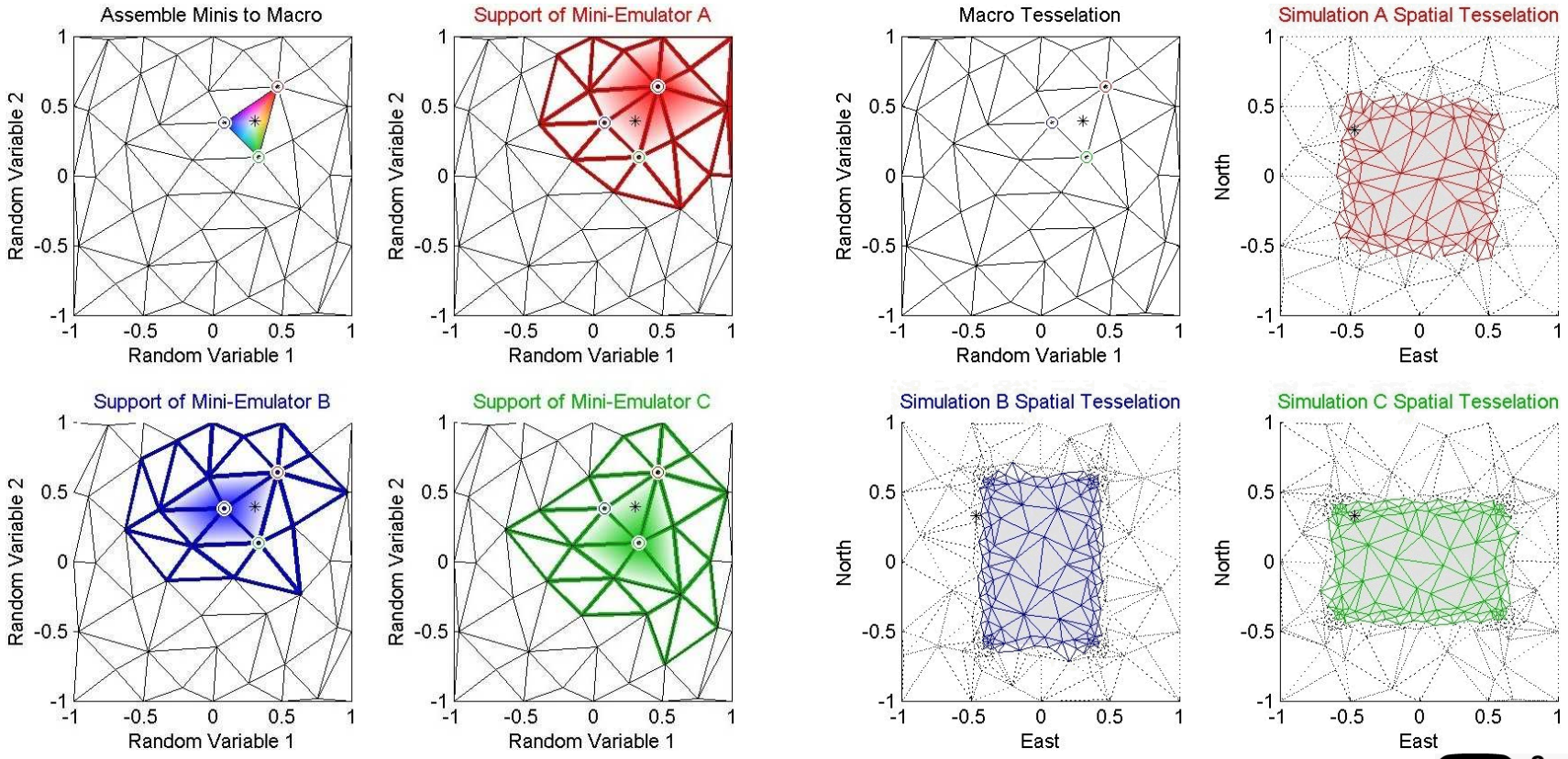
3-Level Hierarchical Emulator



Sandia National Laboratories

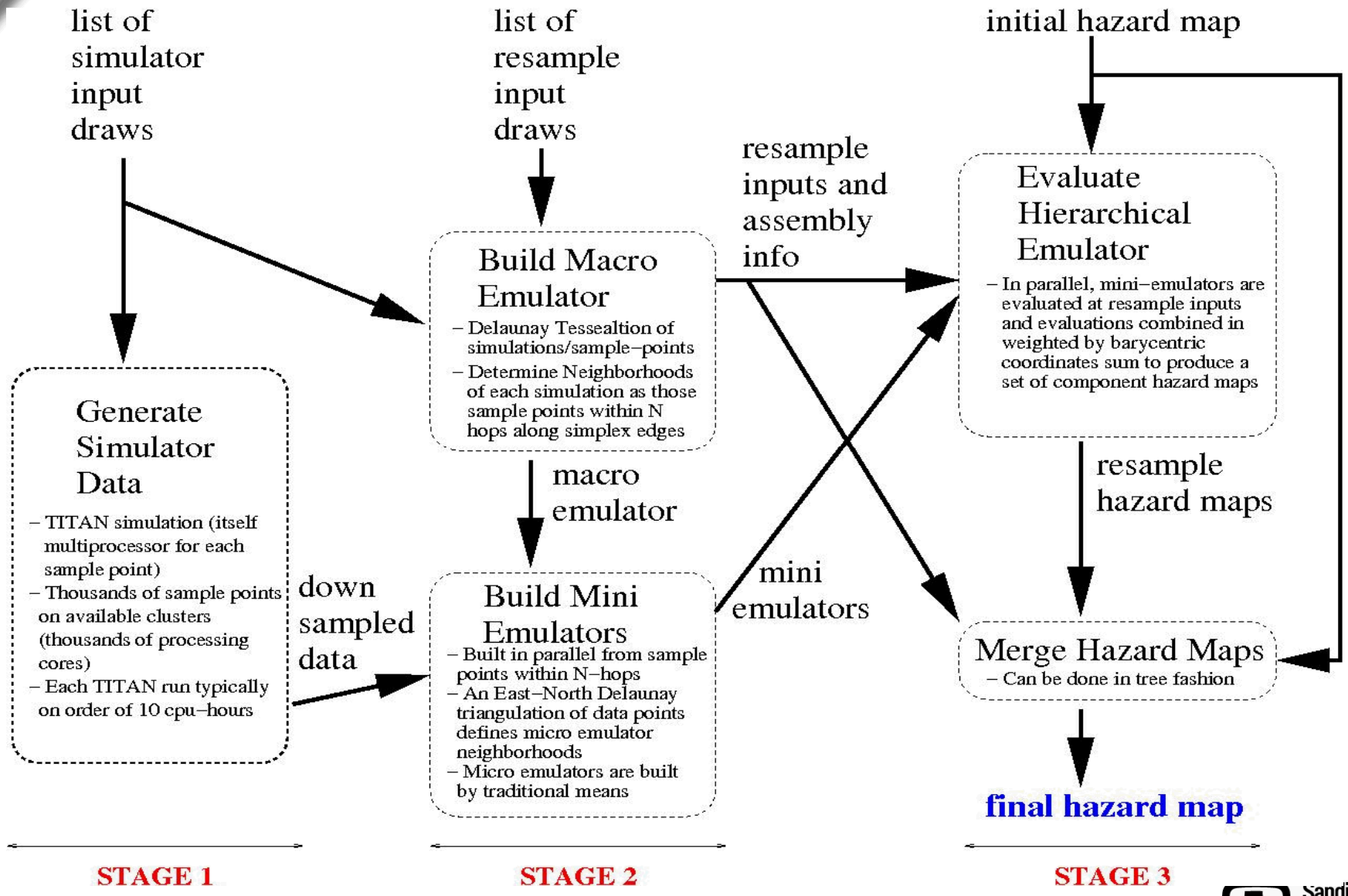
3-Level Hierarchical Emulator

- Emulator's inputs are the tensor product of simulation output's physical spatial dimensions & stochastic inputs
- Error model is correlated through **all** emulator inputs



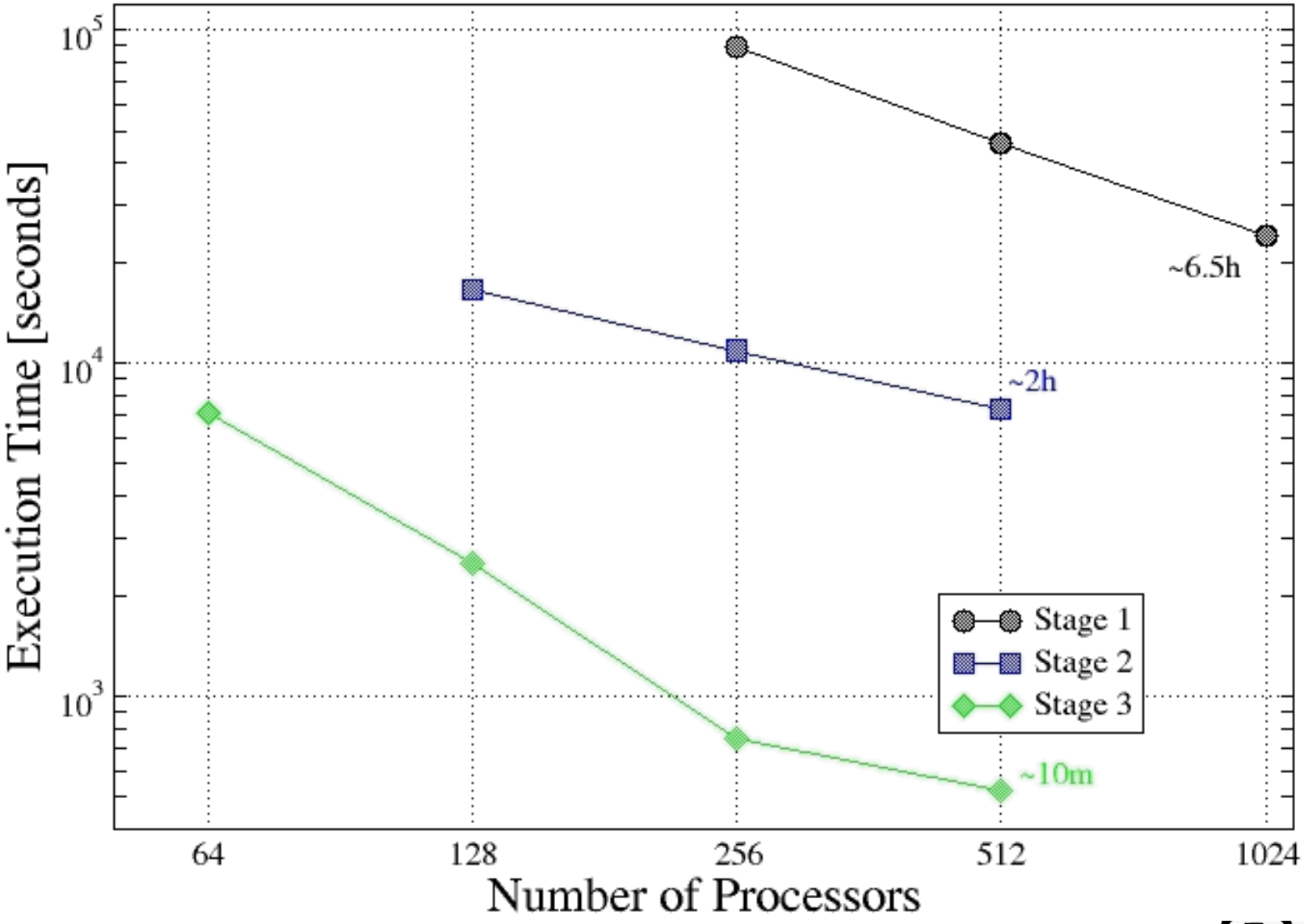
Sandia National Laboratories

Work Flow: 3 Stages



Sandia National Laboratories

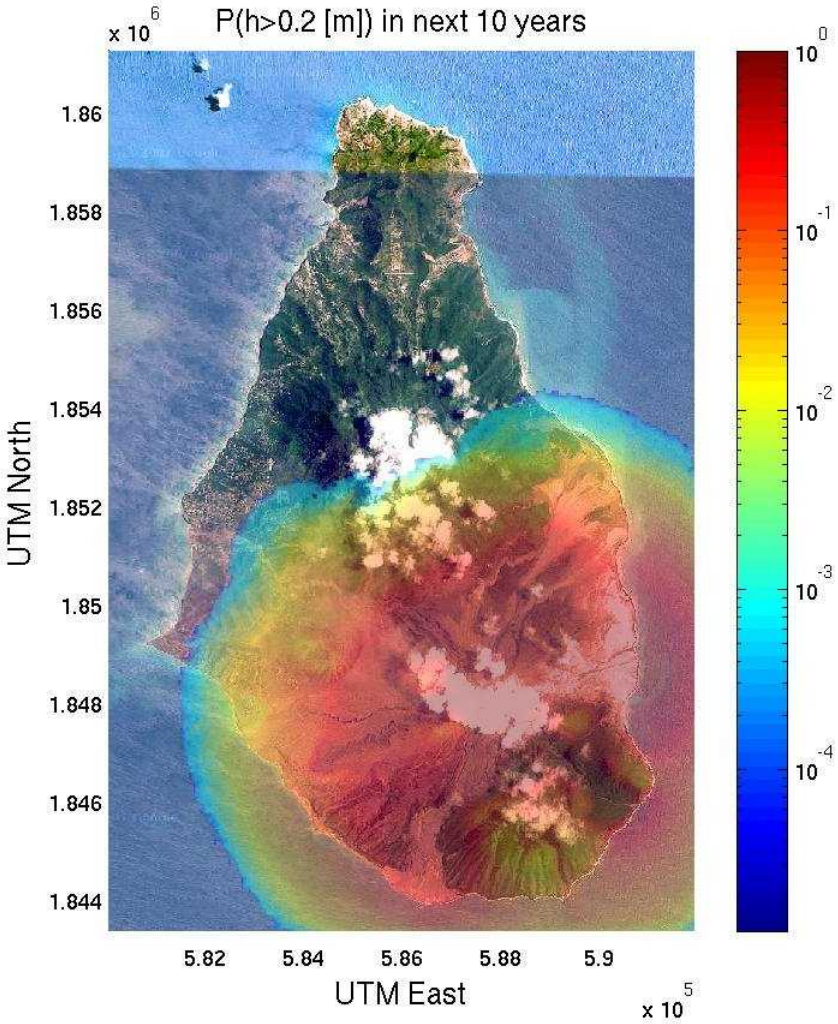
Hierarchical Emulator Results

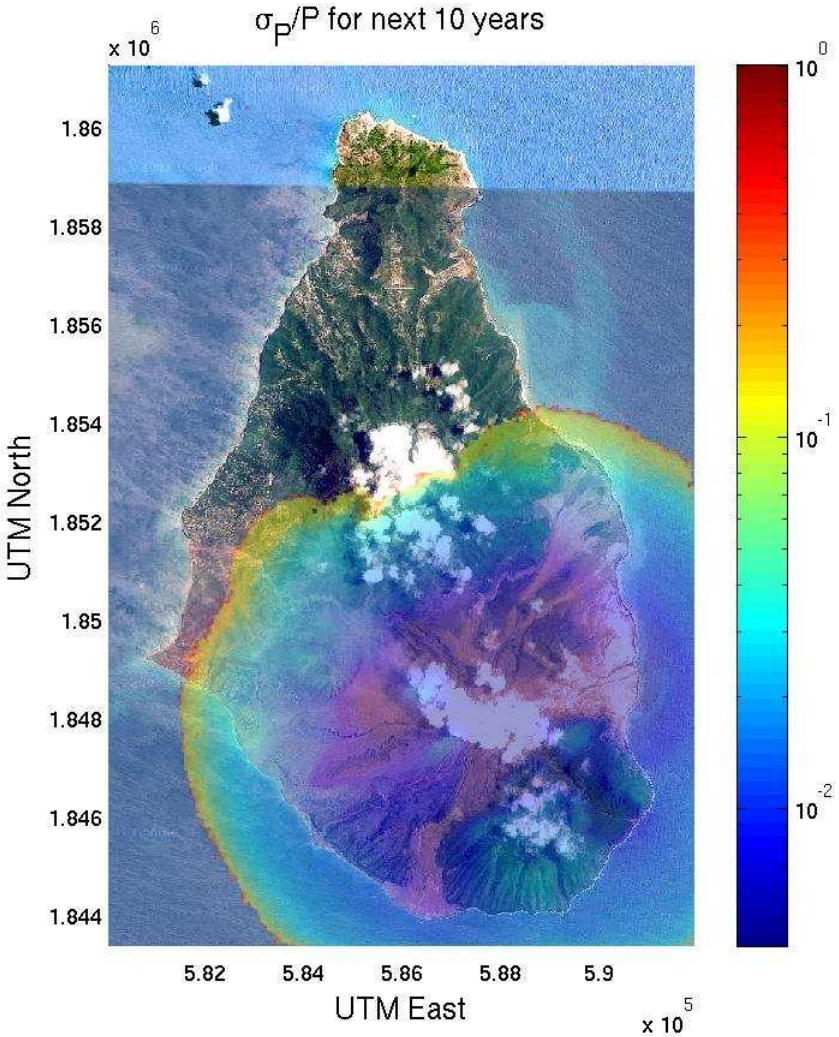


Sandia National Laboratories

Hierarchical Emulator Results

Hazard Map: Volcanic Island of Montserrat





Sandia National Laboratories

Conclusions

Replacing single global emulator built from N points with ensemble of N component emulators built from M points

- Changes build cost from $\mathcal{O}$(N^3) to $\mathcal{O}$(N M^3) operations, if N=$\mathcal{O}$(10^6) & M=$\mathcal{O}$(100) this is $\mathcal{O}$(10^6) reduction
- Avoids problem of ill-conditioned correlation matrix
- Allows ensemble “macro-emulator” to be non-stationary
- Allows for concurrent construction & concurrent evaluation of component emulators (embarrassingly parallel)
- Allows data storage requirements to be distributed among nodes of commodity cluster supercomputer
- Has the same degree of smoothness/continuity