UQ Data Fusion: An Introduction and Case Study

Robert Wolpert
Duke Department of Statistical Science
rlw@duke.edu

MUMS Opening Workshop
SAMSI, RTP NC USA
2018 August 23 9am EDT

What's Data Fusion?

Quoth Wikipedia from *Haghighat*$^{+}$, *2016*, **Data Fusion** is:

> *the process of integrating multiple data sources to produce more consistent, accurate, and useful information than that provided by any individual data source.*

SO,

Suppose we have J data vectors Y_j, each distributed from a parametric distribution

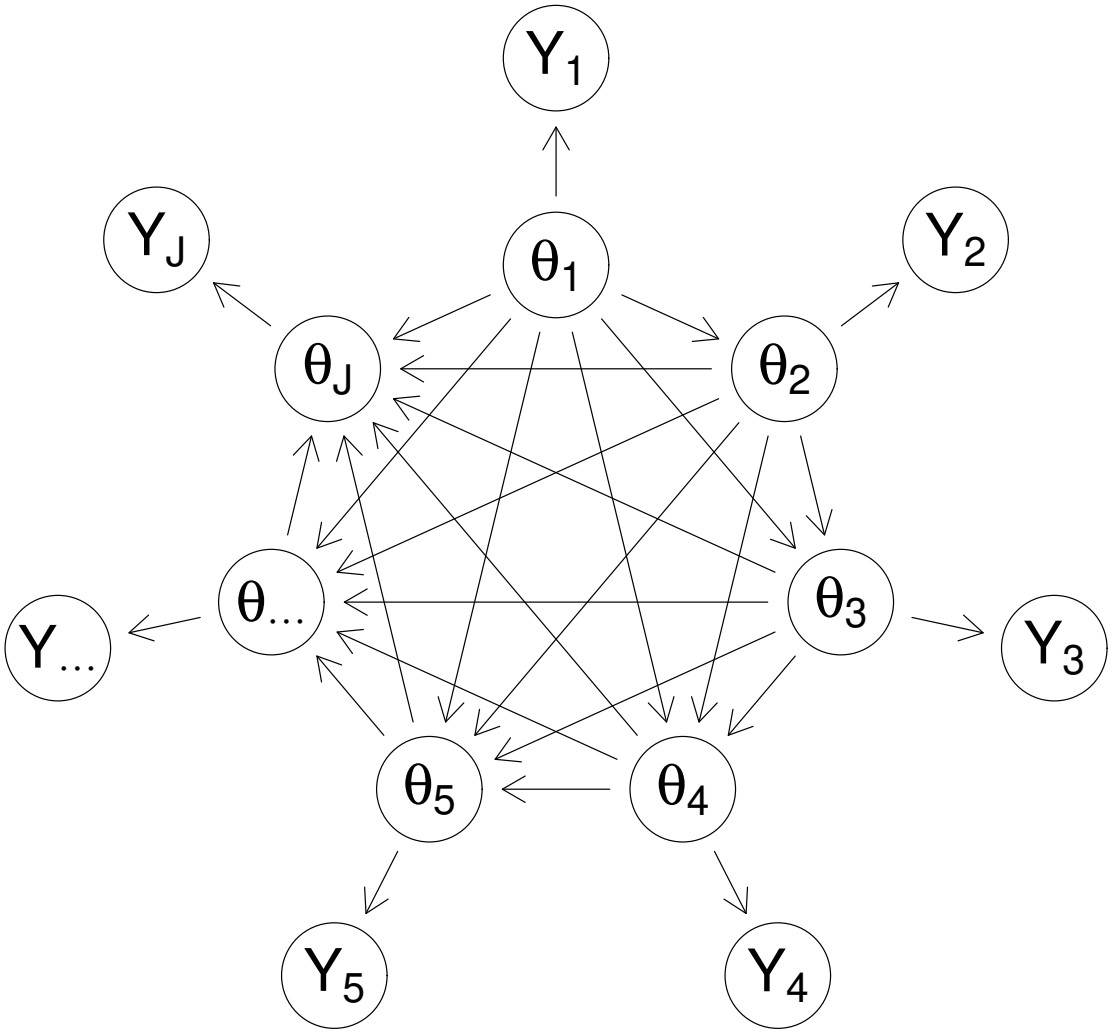
$$Y_j \sim f_j(y_j \mid \theta_j)$$

for some uncertain parameter vector $\theta_j \in \Theta_j$.
In general the joint prior distribution for

$$(\theta_1, \cdots, \theta_J) \in \prod_j \Theta_j$$

could be *anything*...

In pictures, as a DAG...



rtn

Kind of a mess.

We need:

1. More structure, i.e., fewer arrows;
2. Clearer link to whatever's important for us.

A Bayesian Solution, in Words & Formulas:

- Identify just what it is that we care about from these experiments, the Quantity of Interest (QoI)— Average eruption rate? Mean number λ_J of pulses in GRBs? Something else? Let's call it "ε".
- Identify a vector θ of whatever is

  uncertain and **common to two or more** $\{\theta_j\}$,

  in the sense that the collection

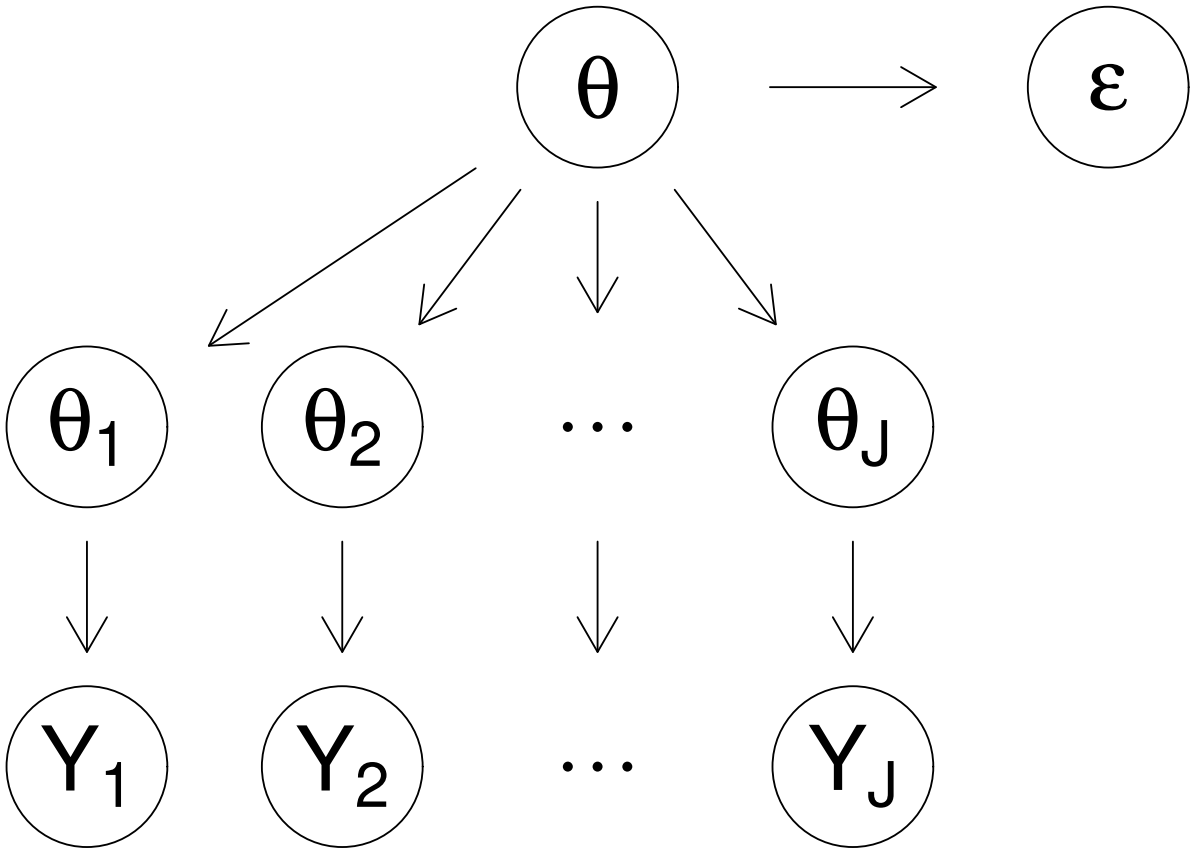
  $\{\theta_j\}$, ε are **conditionally independent** *a priori* given θ.

  That means the conditional prior distribution will factor as:

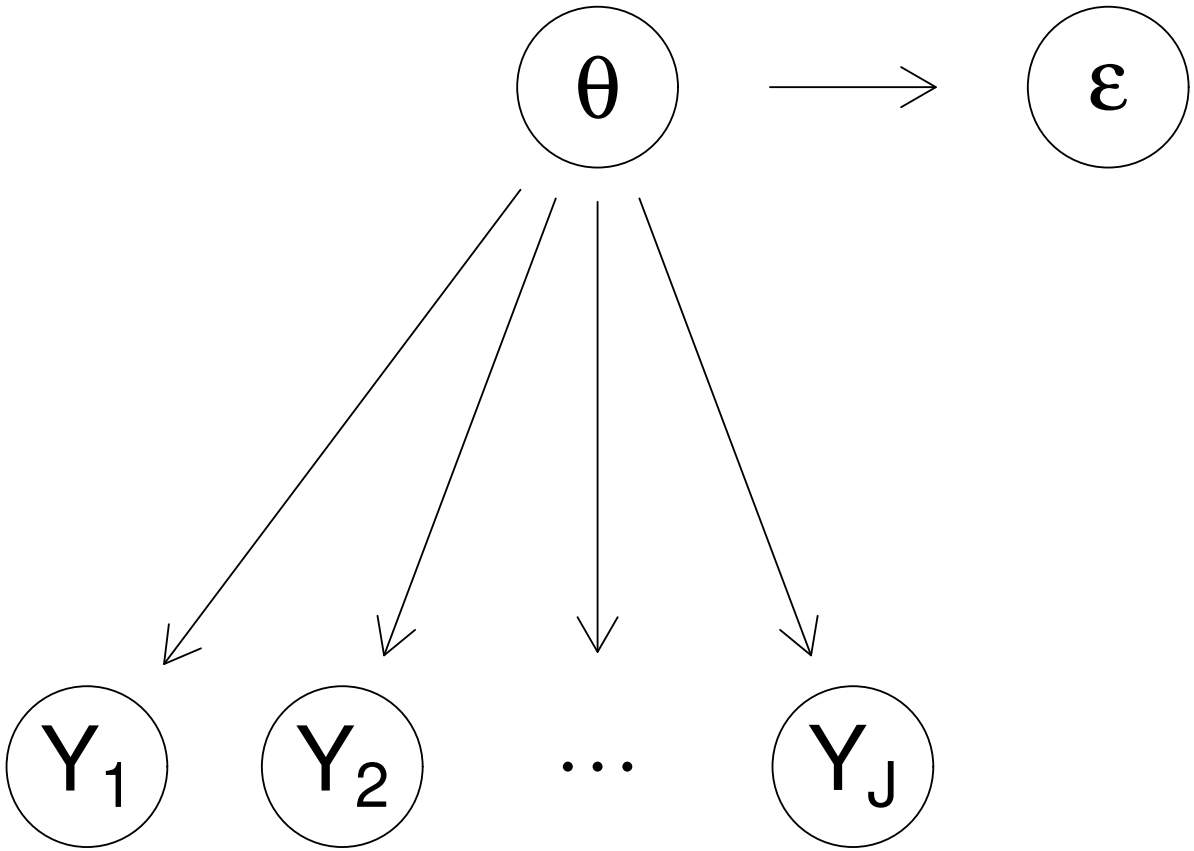
$$\pi(d\theta_1 \cdots d\theta_J \, d\varepsilon \mid \theta) = \pi(d\varepsilon \mid \theta) \prod_{j \le J} \pi(d\theta_j \mid \theta)$$

This entails some thoughtful modeling and some compromises and approximations, in the hope of finding a **low dimensional** feature θ that "separates" ε and $\{\theta_j\}$, with **simple and tractable** distributions $\pi_j(d\theta_j \mid \theta)$.

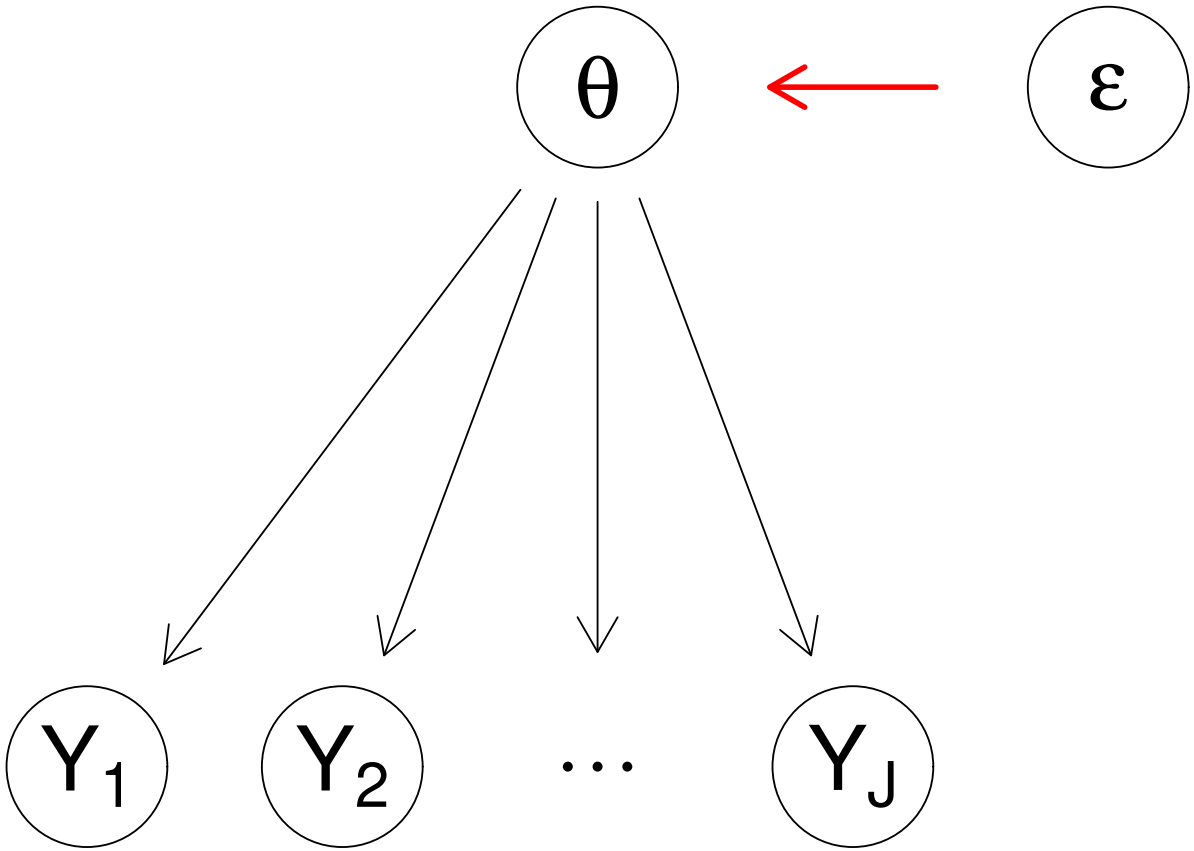
And in pictures, as a DAG...



And in pictures, as a DAG...



And in pictures, as a DAG...



Leading to "Marginal Likelihood" for ε:

$$\mathcal{L}(\varepsilon) := \int_\Theta \left\{ \prod \mathcal{L}_j(\theta) \right\} \pi(d\theta \mid \varepsilon)$$

based on the "**adjusted likelihood functions**" for θ:

$$\mathcal{L}_j(\theta) := \left\{ \int_{\Theta_j} f_j(y_j \mid \theta_j) \pi(d\theta_j \mid \theta) \right\}$$

Now, it's your choice— find $\hat{\varepsilon} := \operatorname{argmax}_\varepsilon \mathcal{L}(\varepsilon)$, plot and explore $\mathcal{L}(\varepsilon)$, or use it to find posterior probabilities and expectations.

Examples:

1. **Multiple trials:** Evidence in J different trials of same (or similar) treatment. Treatment details, patient populations, and subject selection may differ. Interest centers on treatment efficacy ε.
2. **Gamma Ray Bursts:** Evidence from J compound GRBs, in the form of binned photon counts from BATSE satellite. Model counts as increments over bins of continuous-time inhomogeneous Poisson process

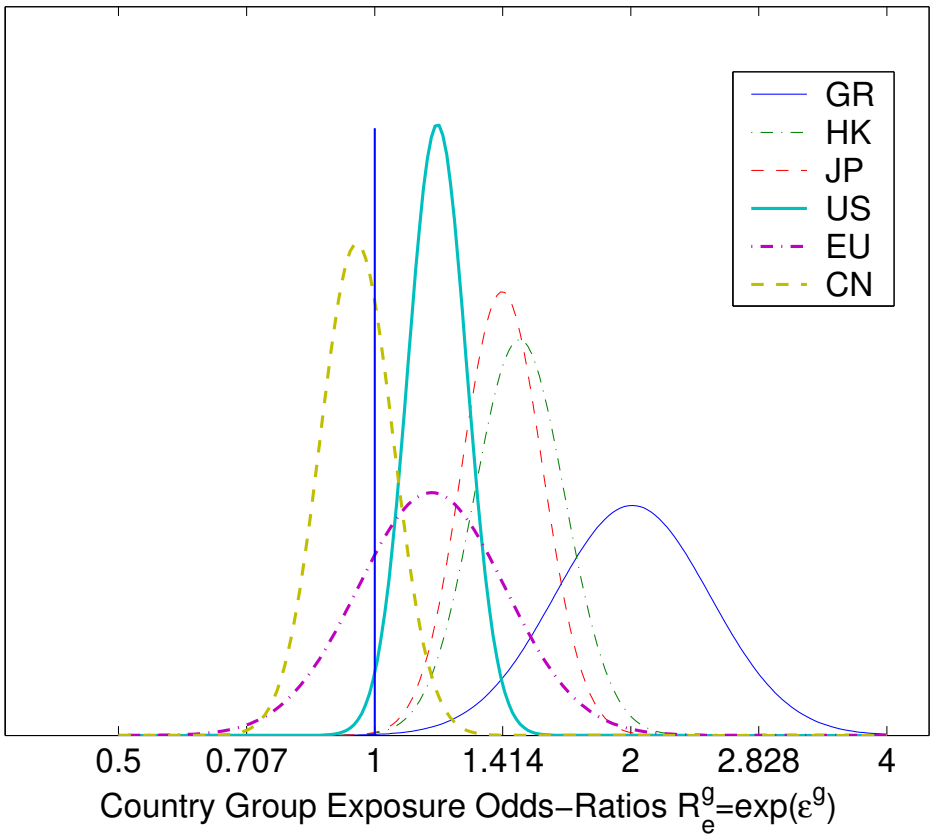
$$Y_t \sim \mathrm{Po}\Big(f_{\theta_j}(t)\,dt\Big)$$

   with some semi-parametric structure on $\{f_{\theta_j}\}$. Interest centers on features ε of the distribution of the number and amplitudes of individual pulses that comprise a burst.
3. **Volcanic Hazard:** Evidence from historical eruption records from multiple volcanoes, from measurement of seismic, atmospheric, and other sources must be synthesized or "Fused" to generate forecasts of hazard at specific locations near an active volcano.

Example 1: Multiple Trials of 2HT Exposure

Data from thirty-one trials from six country groups studying the risk of second-hand tobacco exposure were fused in Wolpert & **Kerrie Mengersen** (2004) in an effort to see the relative risk in women who are correctly classified as to their

- **eligibility** (as non-smoker),
- **diagnosed** as cancerous or not at death, and
- **exposure** to (typically spousal) 2nd-hand smoke.



Example 1: Multiple Trials (cont)

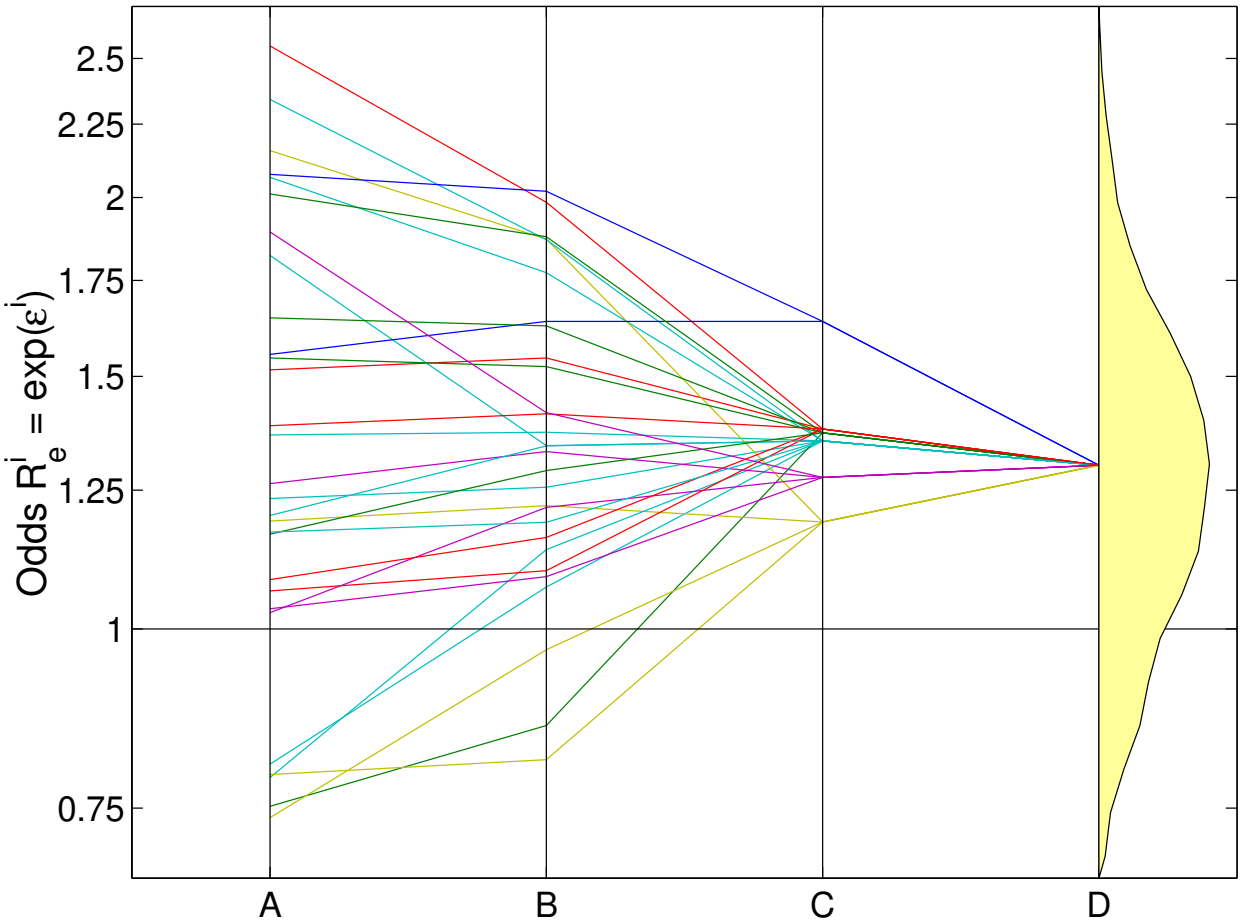


Figure: Individual exposure odds ratios $R_e^i = \exp(\varepsilon^i)$: (A) MLE, (B) Posterior mean, (C) posterior group mean, and (D) overall effect (with posterior pdf) in hierarchical model without quality adjustment, all on log scale.

Example 2: Gamma Ray Bursts

Here we model Gamma Ray Burst photon arrivals as a Cox process:

$$Y_t \sim \mathrm{Po}\Big(f_\theta(t)\,dt\Big)$$

for some structured random mean function $f_\theta(t)$. Below we will take

$$f_\theta(t) = B + \sum_{j=1}^{J} A_j k_j(t \mid \theta)$$

to be a background rate plus a weighted sum of kernels of "Norris" form

$$k_j(t \mid \theta) = \exp\left(2\sqrt{\tau_{1j}\tau_{2j}} - \frac{\tau_{1j}}{(t-t_{0j})} - \frac{(t-t_{0j})}{\tau_{2j}}\right)\mathbf{1}_{\{t>t_{0j}\}}$$

with uncertain polydimensional parameter

$$\theta = \left(B, J, \vec{A}, \vec{t_0}, \vec{\tau}_1, \vec{\tau}_2\right)$$

Example 2: Gamma Ray Bursts

Here we model Gamma Ray Burst photon arrivals as a Cox process:

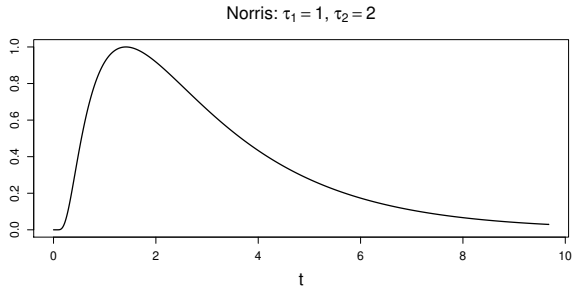
$$Y_t \sim \mathrm{Po}\Big(f_\theta(t)\,dt\Big)$$

for some structured random mean function $f_\theta(t)$. Below we will take

$$f_\theta(t) = B + \sum_{j=1}^{J} A_j k_j(t \mid \theta)$$

to be a background rate plus a weighted sum of kernels of "**Fred**" form

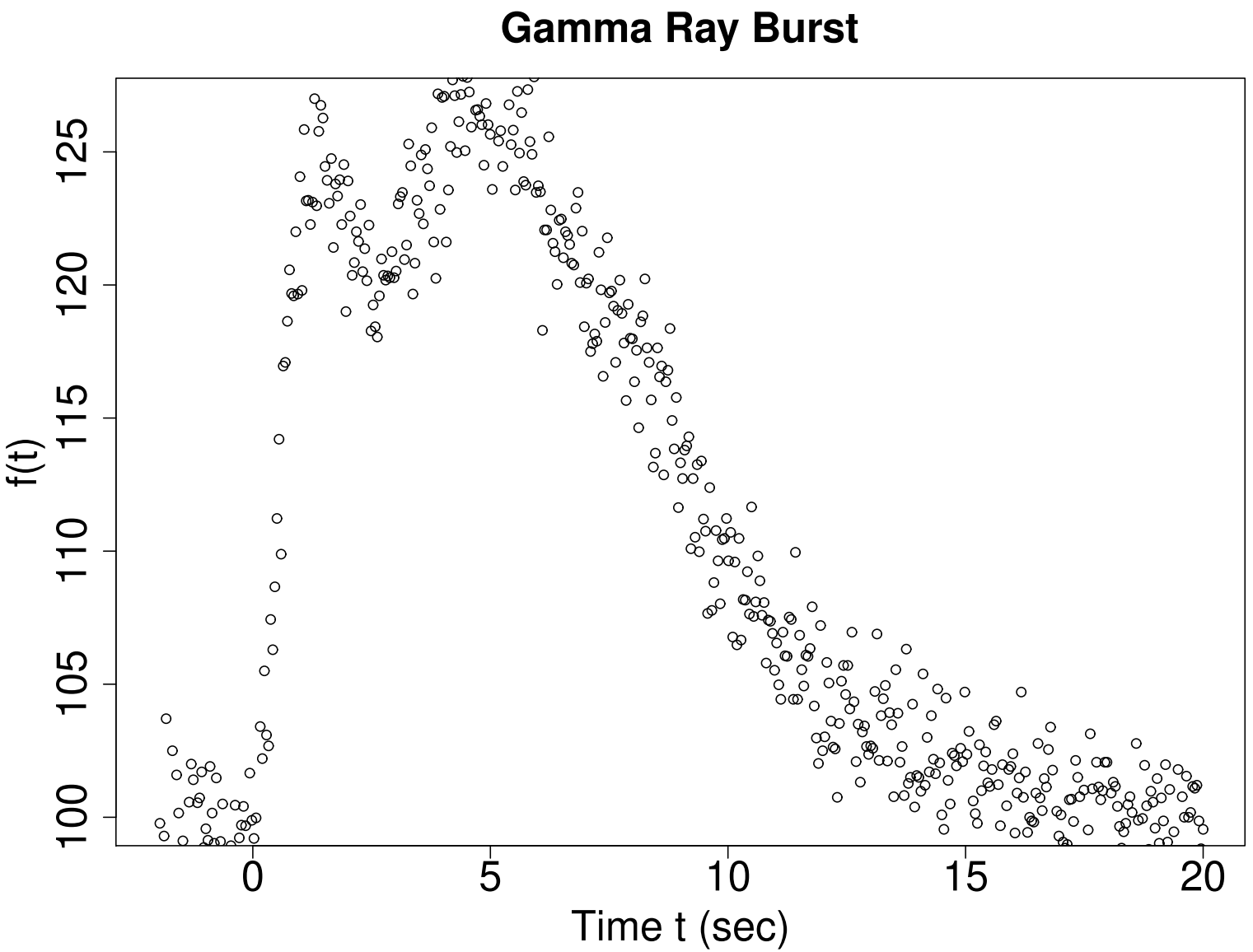
$$k_j(t \mid \theta) =$$



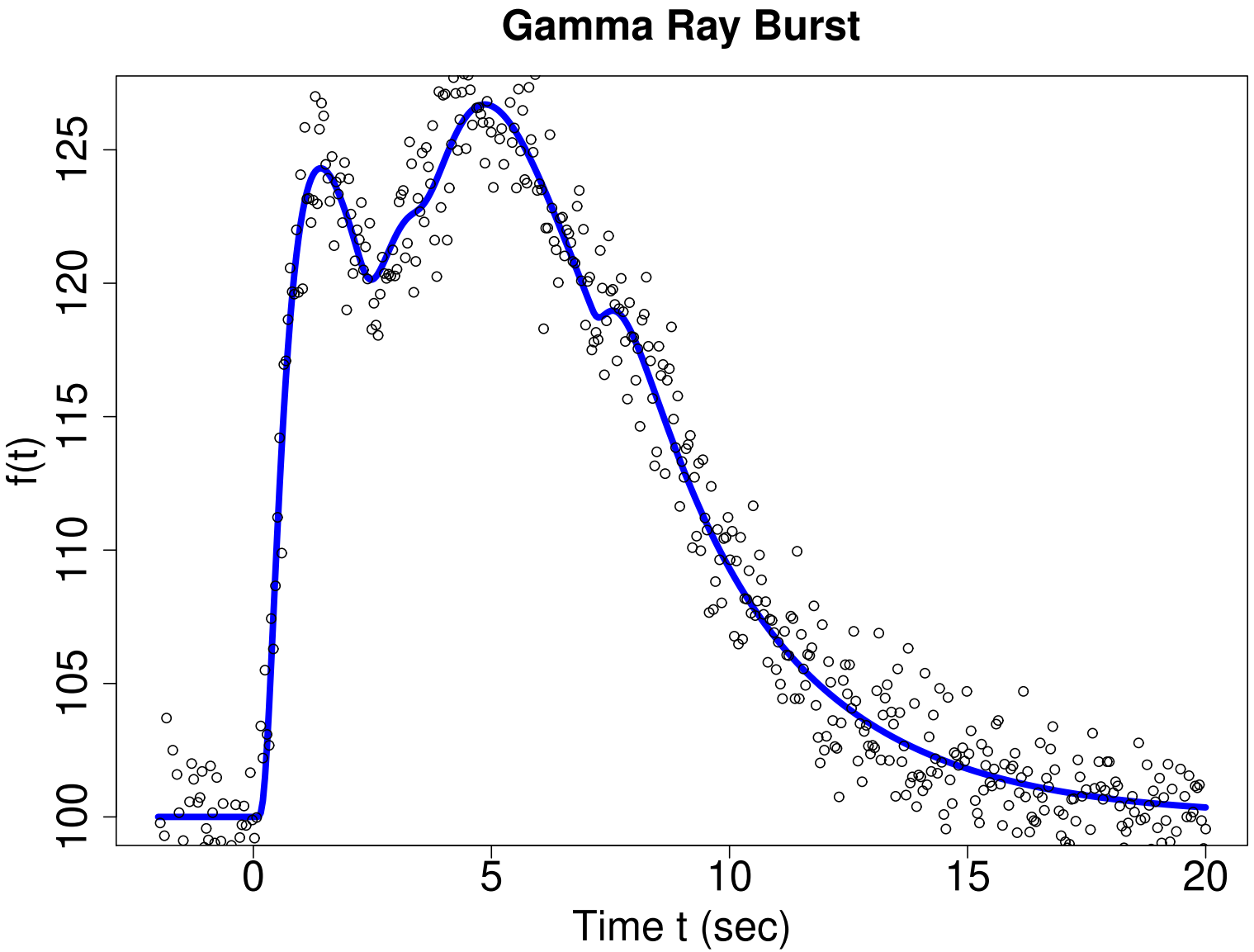
with uncertain polydimensional parameter

$$\theta = \big(B, J, \vec{A}, \vec{t_0}, \vec{\tau}_1, \vec{\tau}_2\big)$$

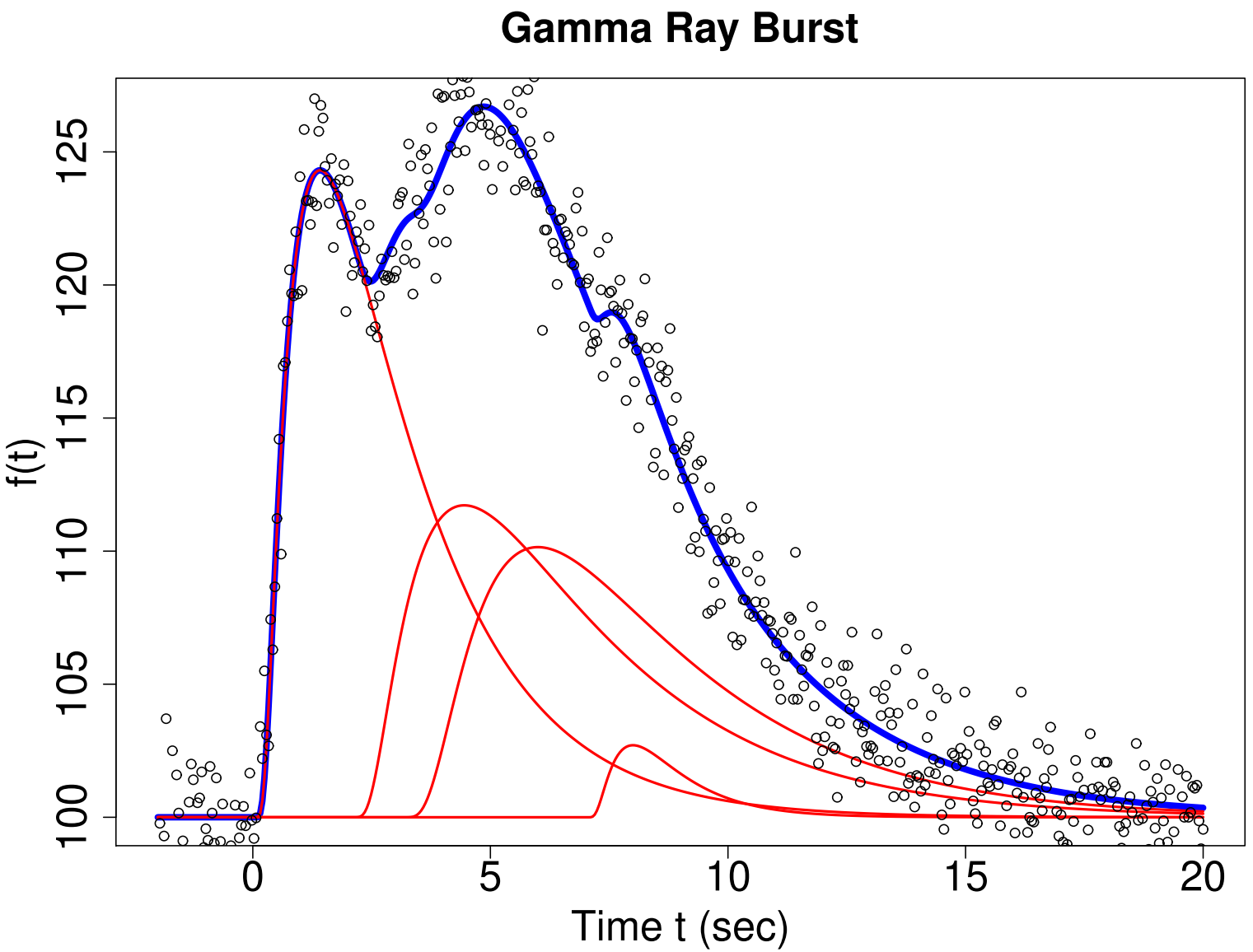
GRB Data (simulated BATSE Poisson bin counts):



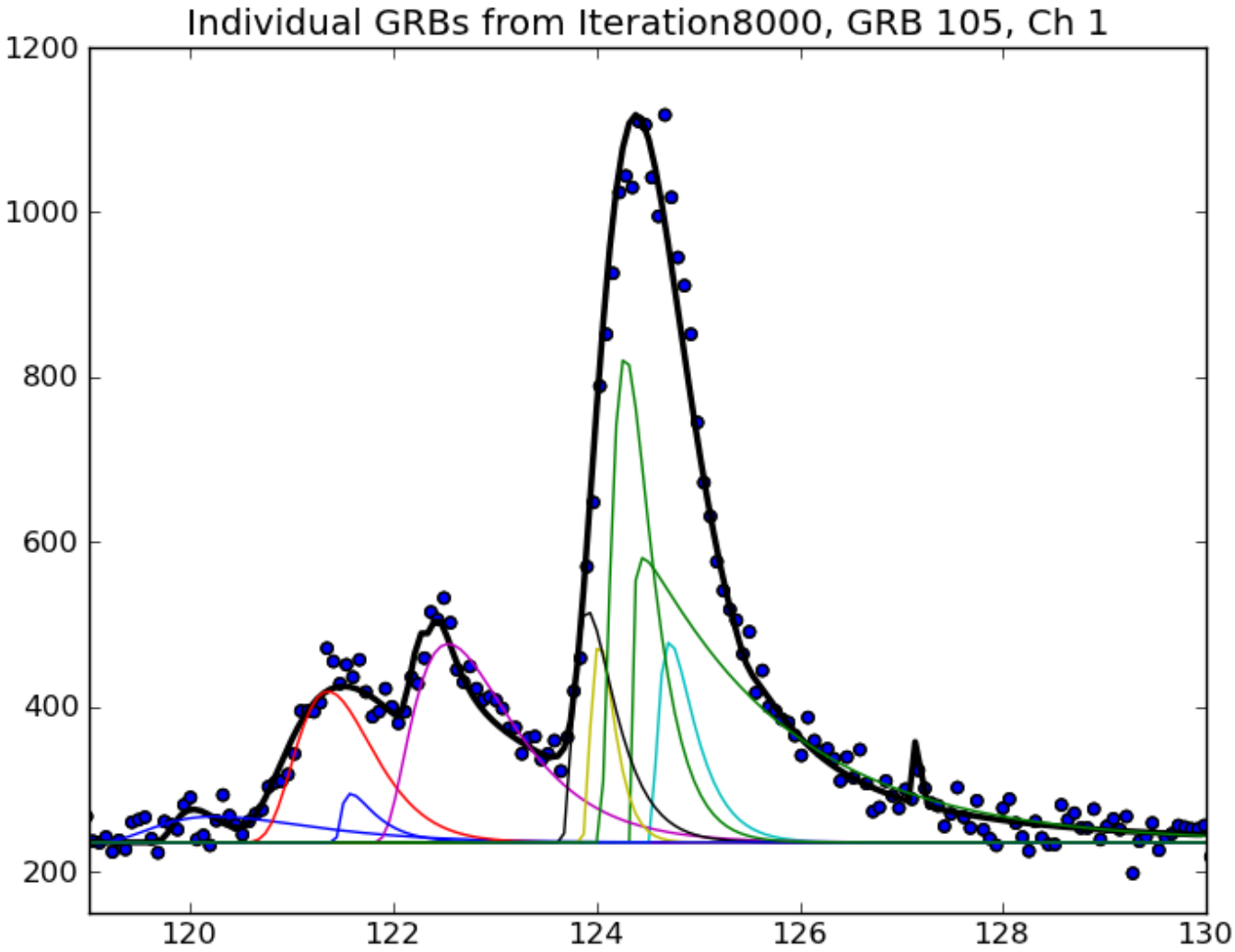
GRB Smoothed estimate of Poisson mean:



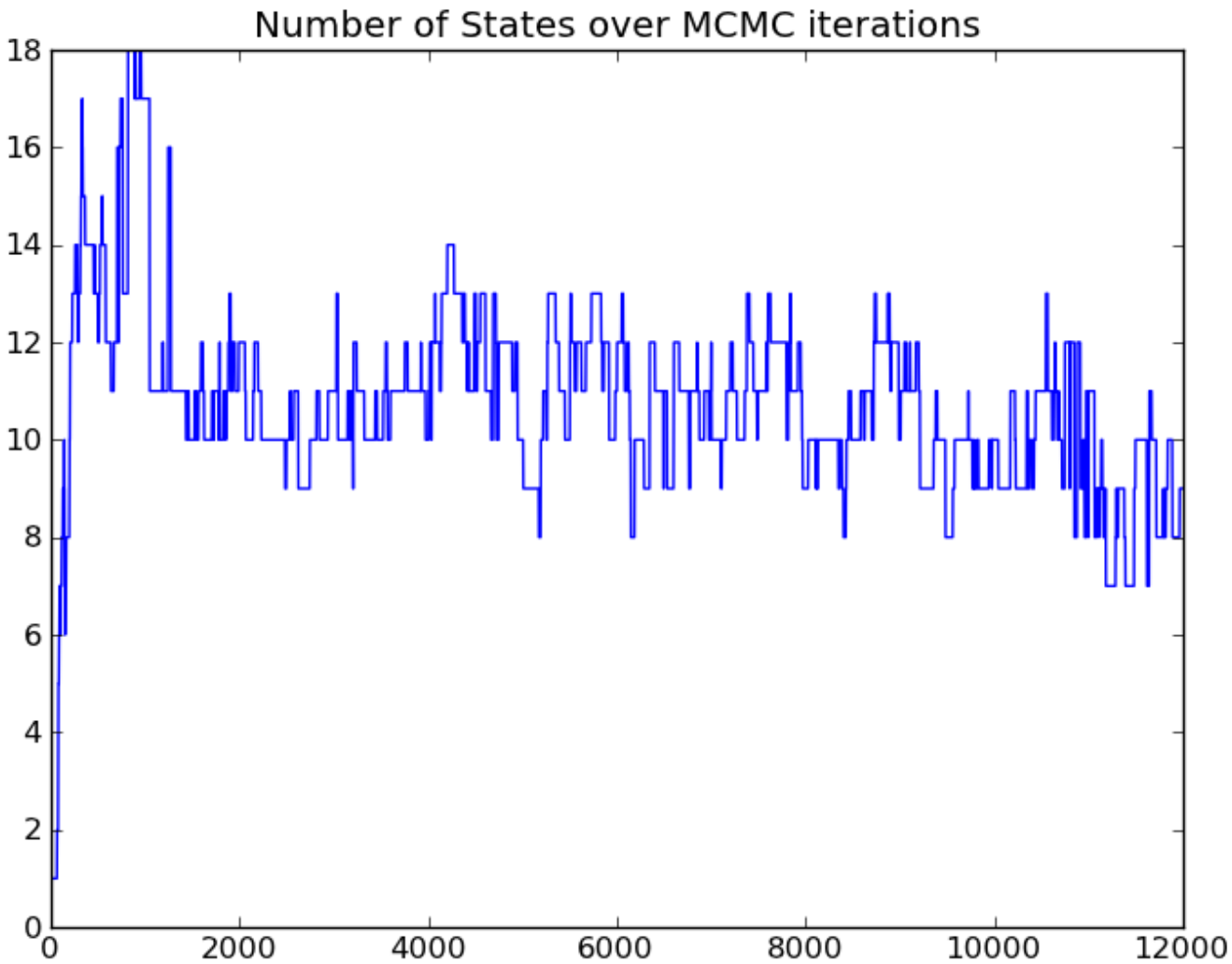
GRB Resolution of burst into pulses:



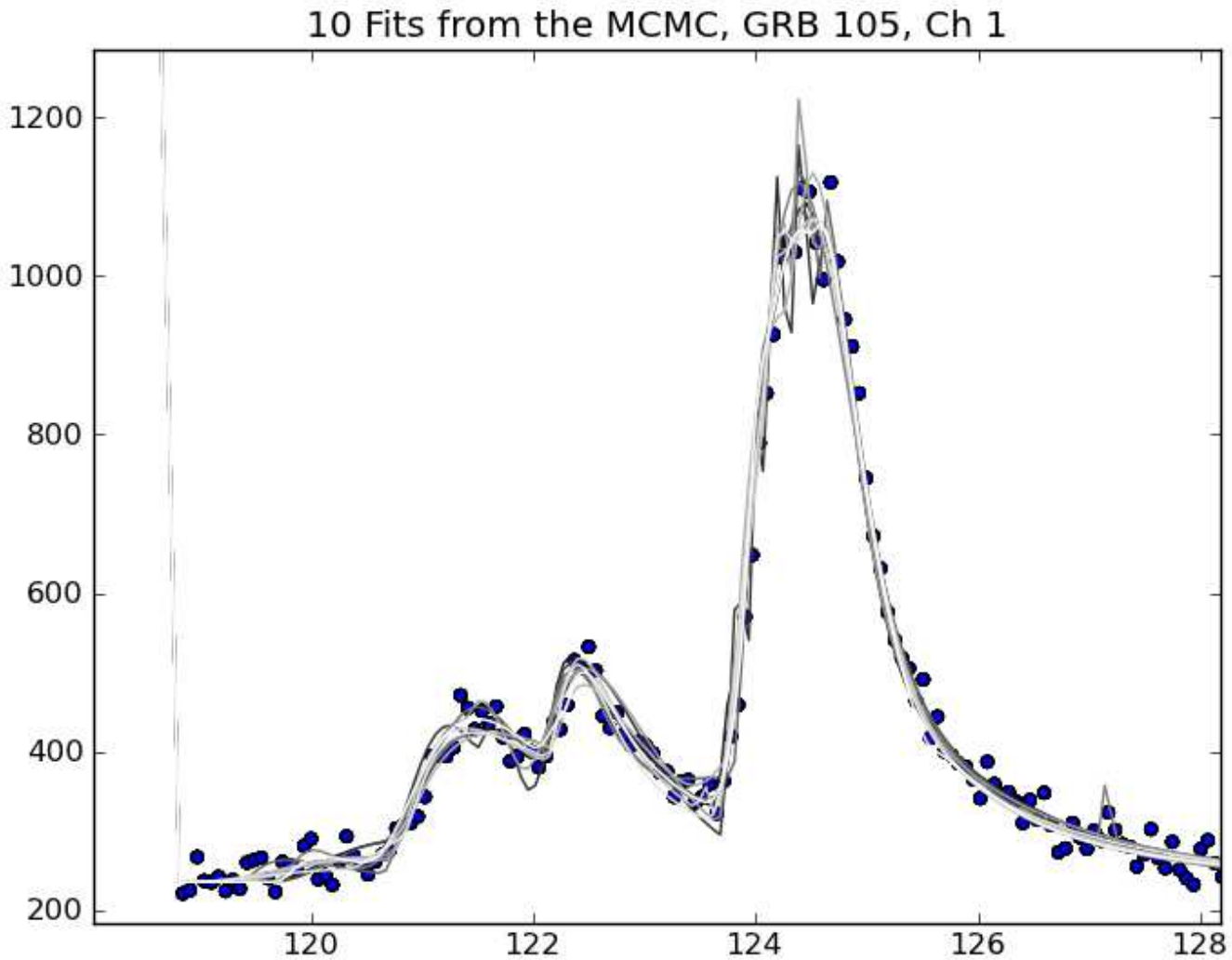
A real BATSE GRB and one possible resolution:



Uncertainty about pulse count J:



Resulting uncertainty about Light Curve:



The Frequentist approach:

Frequentist:

- Try fitting one pulse to light curve:

$$f(t \mid \theta_1) = B + A_1 k(t \mid \theta_1)$$

The Frequentist approach:

Frequentist:

- Try fitting one pulse to light curve:

$$f(t \mid \theta_1) = B + A_1 k(t \mid \theta_1)$$

- Like it? Quit and report $\theta = (B, J = 1, A_1, \theta_1)$.

The Frequentist approach:

Frequentist:

- Try fitting one pulse to light curve:

$$f(t \mid \theta_1) = B + A_1 k(t \mid \theta_1)$$

- Like it? Quit and report $\theta = (B, J = 1, A_1, \theta_1)$.
- No? Try two:

$$f(t \mid \theta_1, \theta_2) = B + \sum_{j=1}^{2} A_j k(t \mid \theta_j)$$

The Frequentist approach:

Frequentist:

- Try fitting one pulse to light curve:

$$f(t \mid \theta_1) = B + A_1 k(t \mid \theta_1)$$

- Like it? Quit and report $\theta = (B, J = 1, A_1, \theta_1)$.
- No? Try two:

$$f(t \mid \theta_1, \theta_2) = B + \sum_{j=1}^{2} A_j k(t \mid \theta_j)$$

- Like it? Quit and report $\theta = (B, J = 2, \vec{A}, \vec{\theta}_j)$.
- No? Try three... or four... until you do, and report θ.

The Bayesian Approach

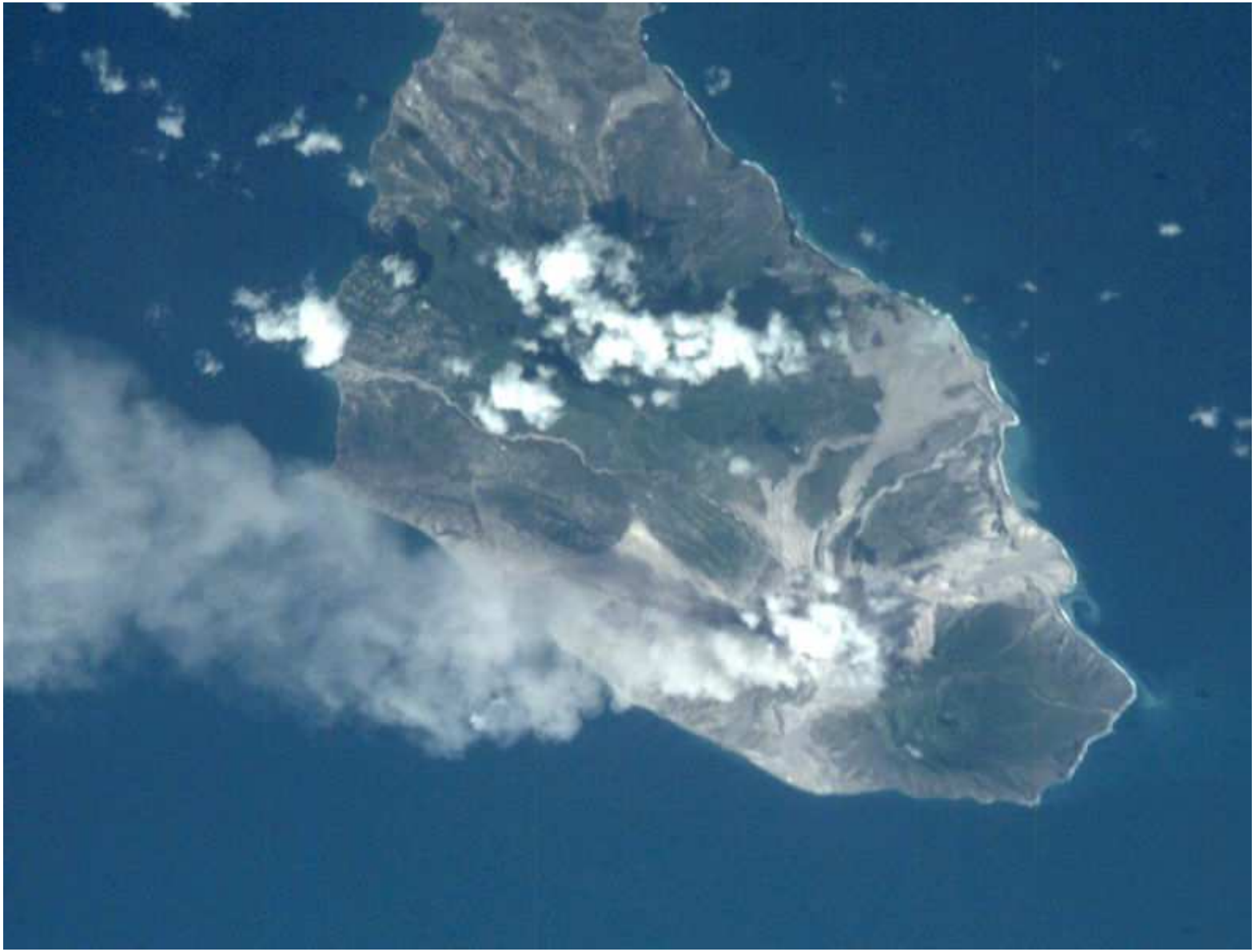
- Choose joint prior on $\theta = (B, J, \vec{A}, \vec{\theta_j})$.
  For J and the amplitudes $\{A_j\}$ with $A_j \geq \varepsilon$ for some threshold $\varepsilon > 0$,
  we[1] use Lévy processes built on Gamma random fields
  (so J has Poisson dist'n, and $\{(A_j, \theta_j)\}_{1 \leq j \leq J}$ are iid, given J.)

- Use Reversible Jump (varying J) Metropolis/Hastings MCMC to sample $\{\theta^{(t)}\}_{t \in \mathbb{N}}$ from posterior distribution.

- Report marginal distributions (or means & variances) of any feature of interest— like $\{J\}$ or total number of photons or max amplitude or duration at half-max-height or ...

[1] Joint work with Tom Loredo, Jon Hakkila, and Duke Stats PhD **Mary Beth Broadbent**

Example 3: Volcanic Hazard on Montserrat



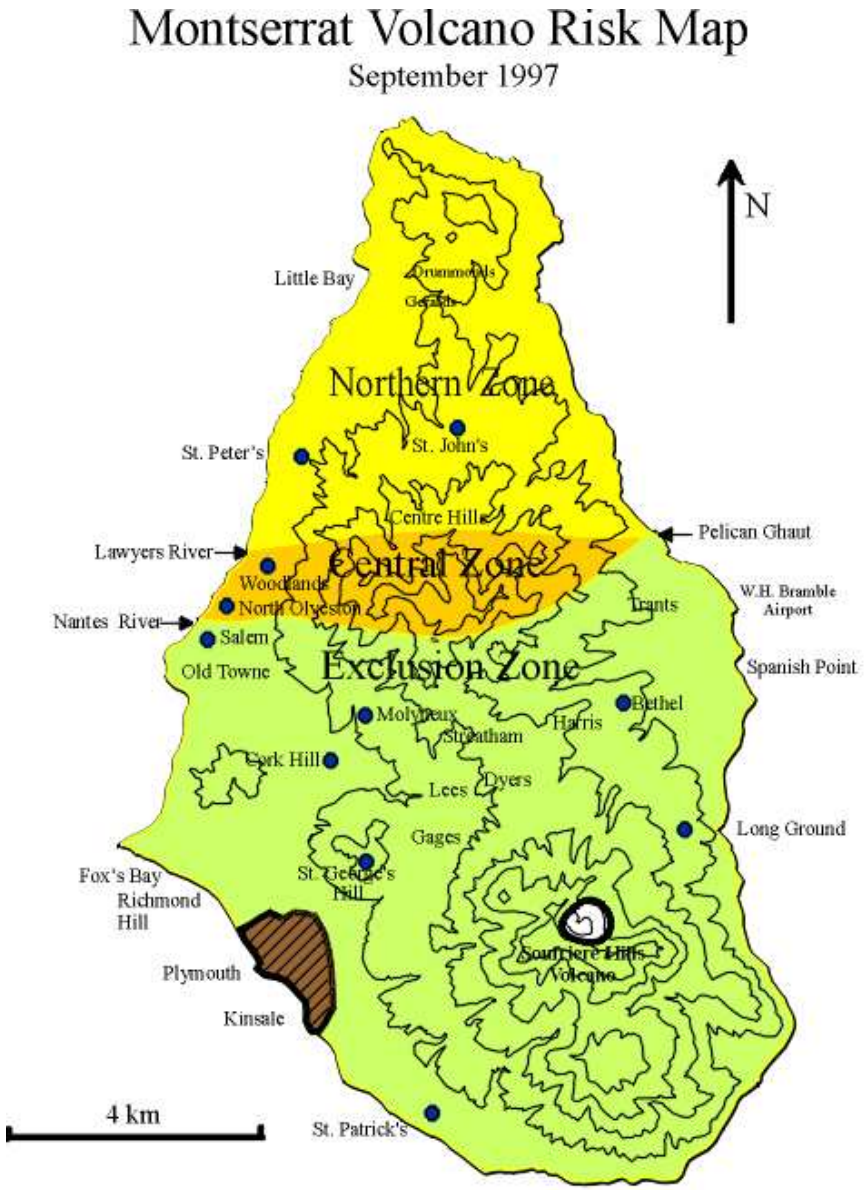
Plymouth, Montserrat, 1994

Plymouth, Montserrat, 1998

Montserrat Risk Map on Current MVO Website

rtn

Montserrat Risk Map on 1997 MVO Website



rtn

Our goal:

Scientifically-based quantitative risk maps: **Contour Maps** giving probability of inundation in specified time period, based on:

- **Topography** captured in high-precision Digital Elevation Model (DEM);
- **Rheology:** Viscosity, mineral variety, gas and water content, grain size distribution, etc.
- **Historical** data on volumes, directions, and frequencies of Pyroclastic Density Currents (PDCs), reflecting
- **Current Conditions:** Seismic activity, Degassing, Dome shape, etc.

Our goal:

Scientifically-based quantitative risk maps: **Contour Maps** giving probability of inundation in specified time period, based on:

- **Topography** captured in high-precision Digital Elevation Model (DEM);
- **Rheology:** Viscosity, mineral variety, gas and water content, grain size distribution, etc.
- **Historical** data on volumes, directions, and frequencies of Pyroclastic Density Currents (PDCs), reflecting
- **Current Conditions:** Seismic activity, Degassing, Dome shape, etc.

We have data on each of these;

Our goal:

Scientifically-based quantitative risk maps: **Contour Maps** giving probability of inundation in specified time period, based on:

- **Topography** captured in high-precision Digital Elevation Model (DEM);
- **Rheology:** Viscosity, mineral variety, gas and water content, grain size distribution, etc.
- **Historical** data on volumes, directions, and frequencies of Pyroclastic Density Currents (PDCs), reflecting
- **Current Conditions:** Seismic activity, Degassing, Dome shape, etc.

We have data on each of these;

Now we have to **Fuse** them!

But....

We don't fuse them all at once (Remember why?)

But....

We don't fuse them all at once (Remember why?)

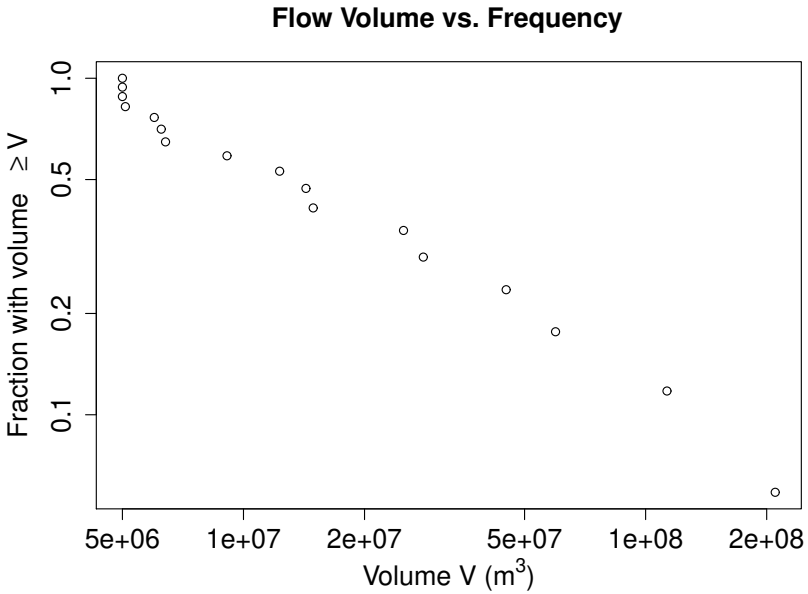
We'll put them together one-by-one, like

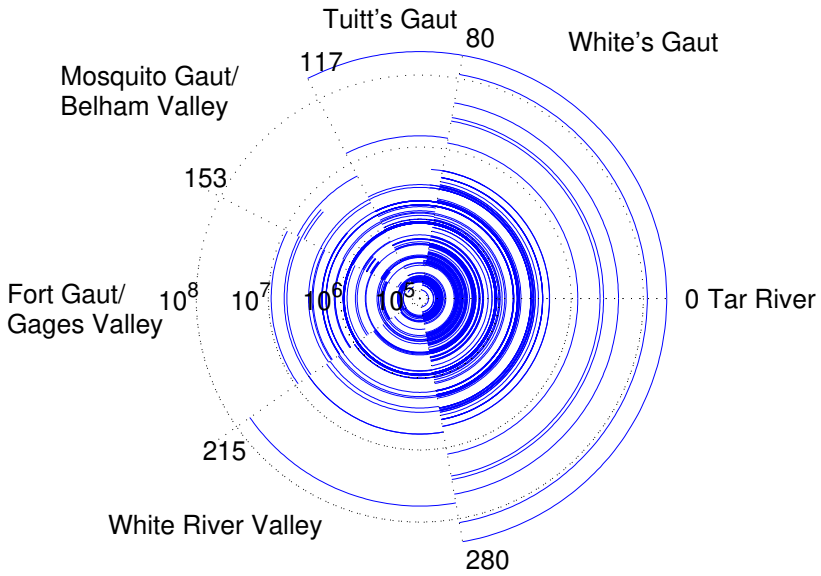
Topography:

- The **TITAN2D** Flow Simulator solves the hyperbolic balance equations describing granular flow down a
  - Specified terrain, starting with a
  - Specified volume V of material, of
  - Specified basal friction angle, with
  - Specified initial direction.
- We can get the terrain from LIDAR and satellite sources, with 2–10m precision.
- What about inputs V, ϕ, and θ? And what about Rates?

... & History:

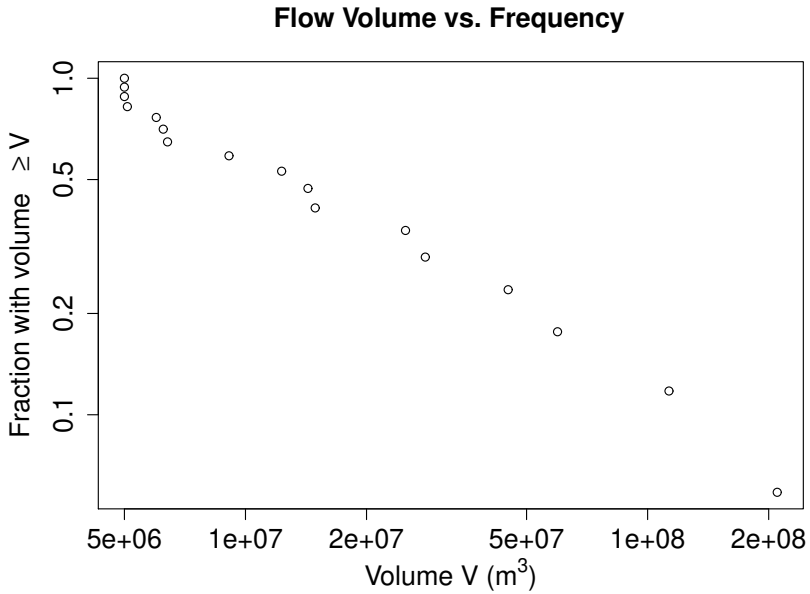
The Historical Record suggests (to us) a **Pareto/vonMises/Poisson** stochastic model for Volumes and Directions and Frequencies:

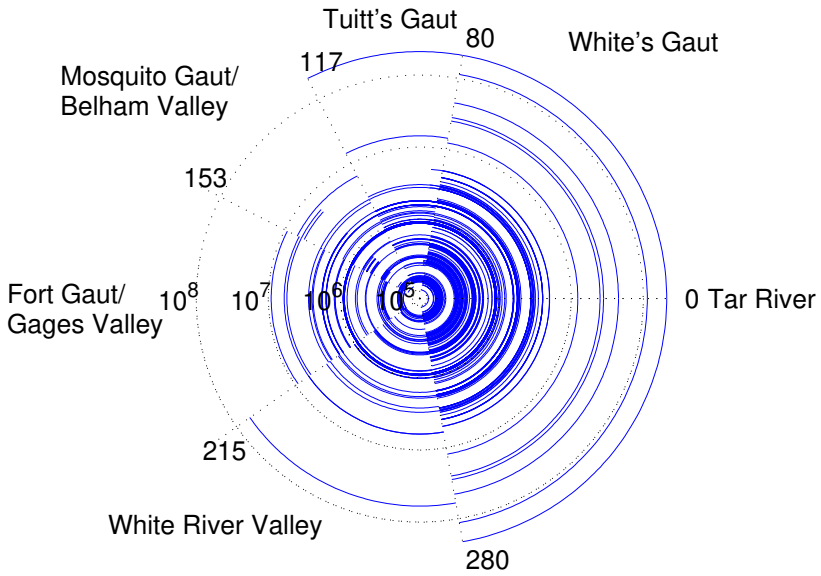




... & History:

The Historical Record suggests (to us) a **Pareto/vonMises/Poisson** stochastic model for Volumes and Directions and Frequencies:





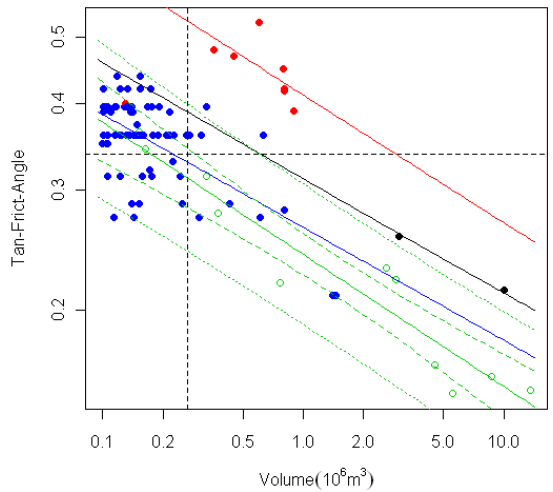
Flow Simulator + **Stochastic Model** ⇒ **Hazard Forecasts**

... & Rheology:

It understates runout for **Large Volume Flows** V, which have been observed to be more mobile than expected— i.e. exhibit lower **Basal Friction** ϕ. Empirically $\tan\phi \propto V^{-\alpha}$, or

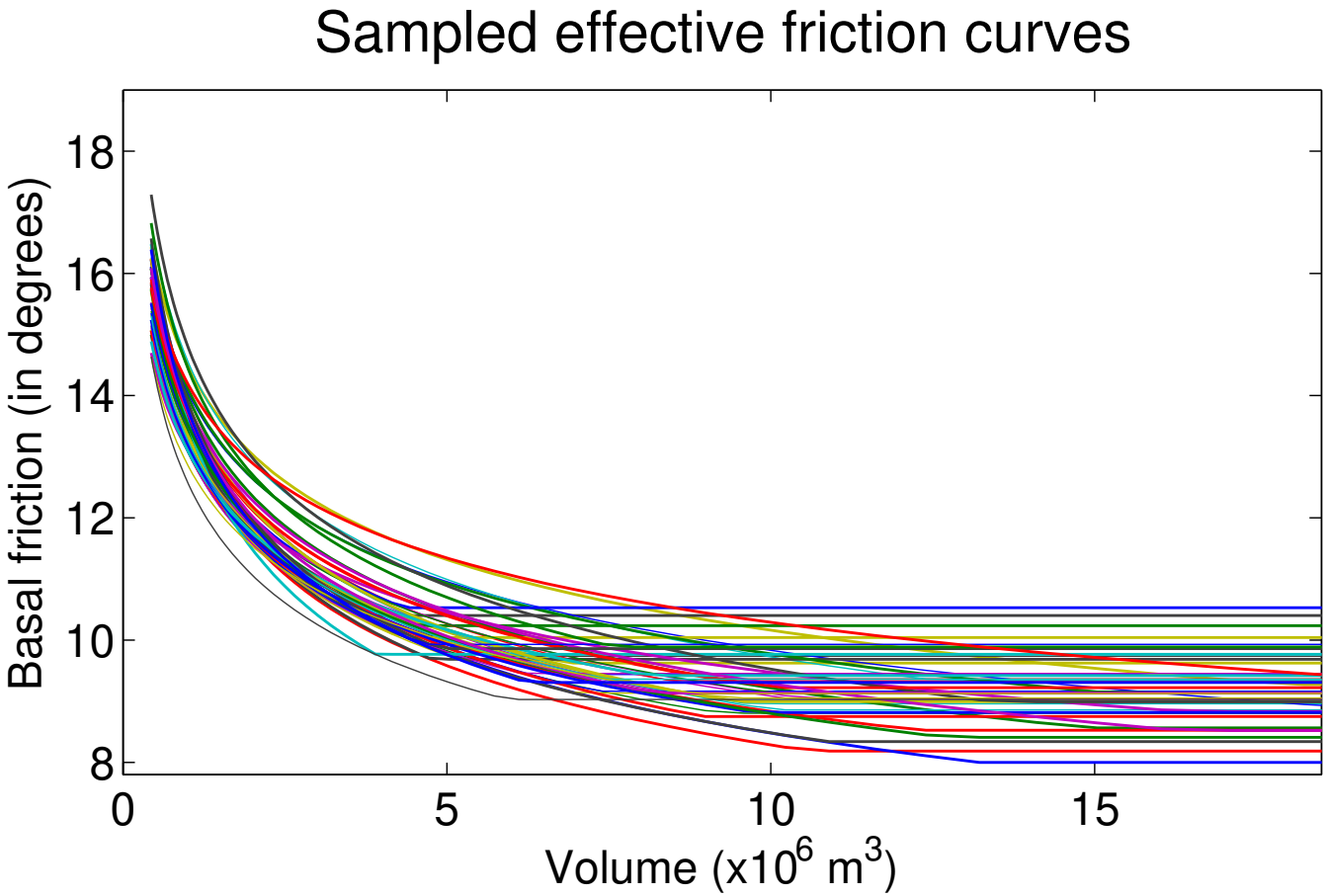
$$\log\tan\phi_j = \beta - \alpha \log V_j$$

Here are data from four similar volcanoes, showing similar slope α:



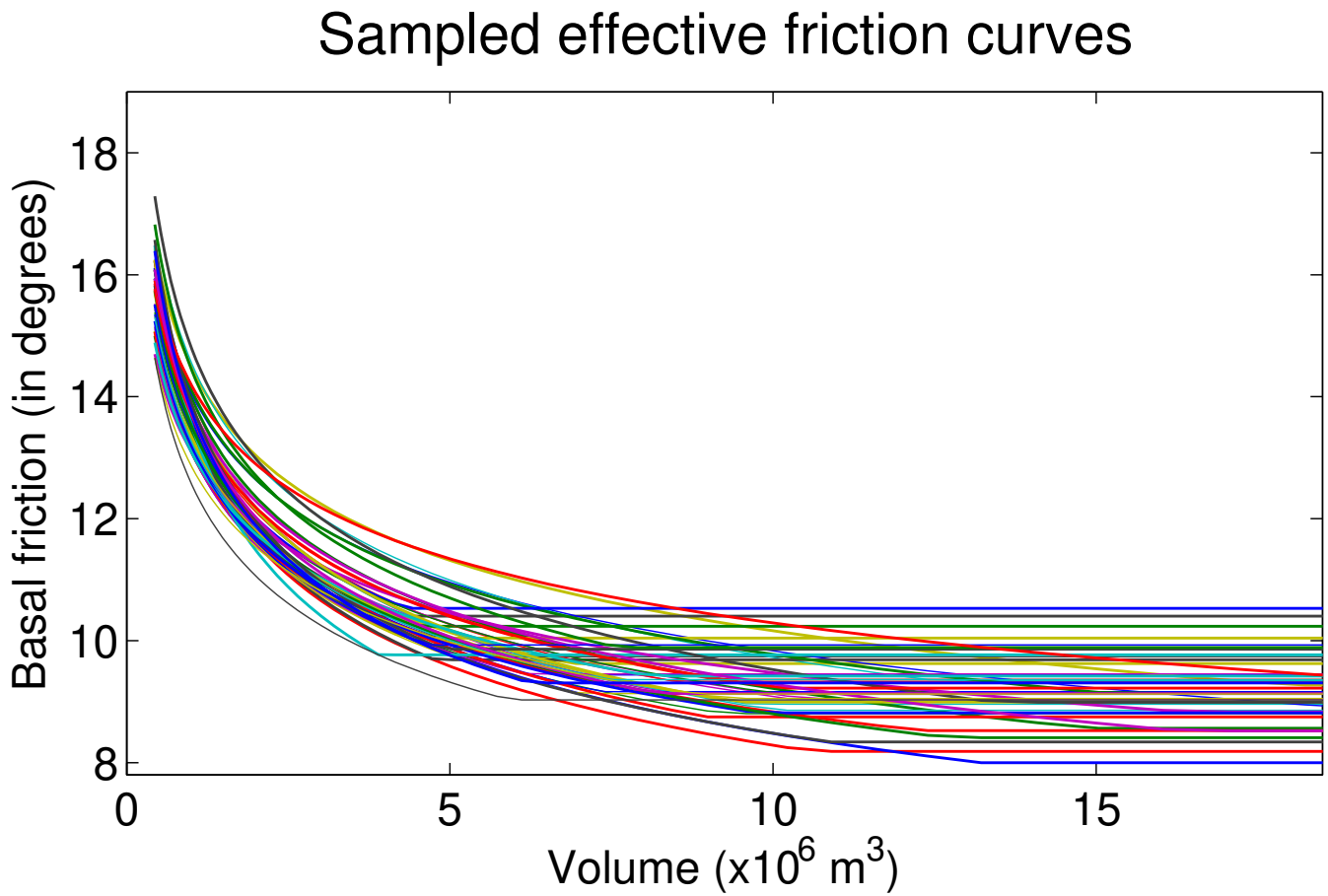
... & Rheology (cont):

Here are posterior samples of the ϕ-vs.-V relation:



... & Rheology (cont):

Here are posterior samples of the ϕ-vs.-V relation:

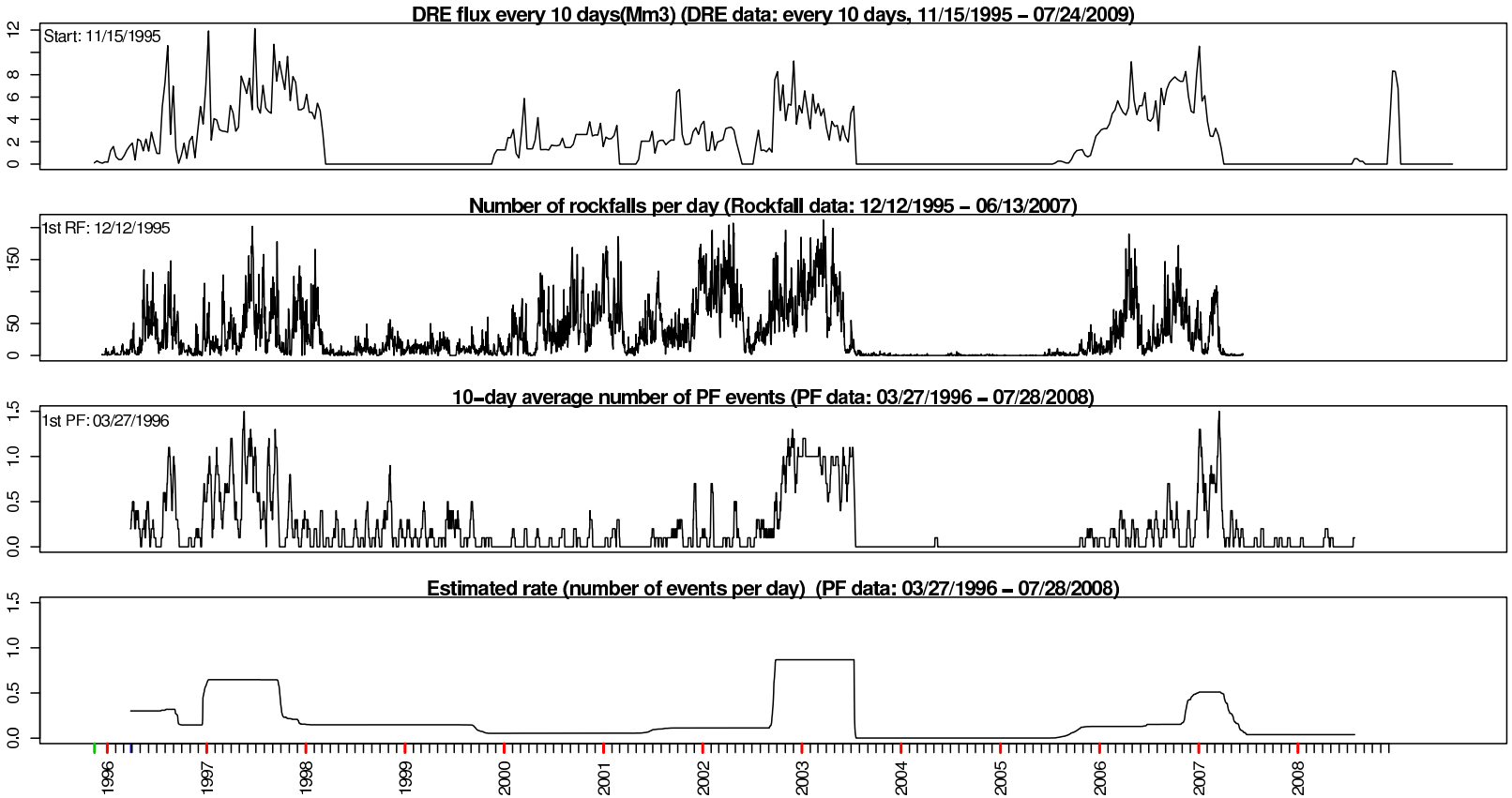


Time to Fuse Again!

Flow Sim + **Stoch Mod** + **Hier Friction Model** $\Rightarrow$ **Better Hazard Forecasts**

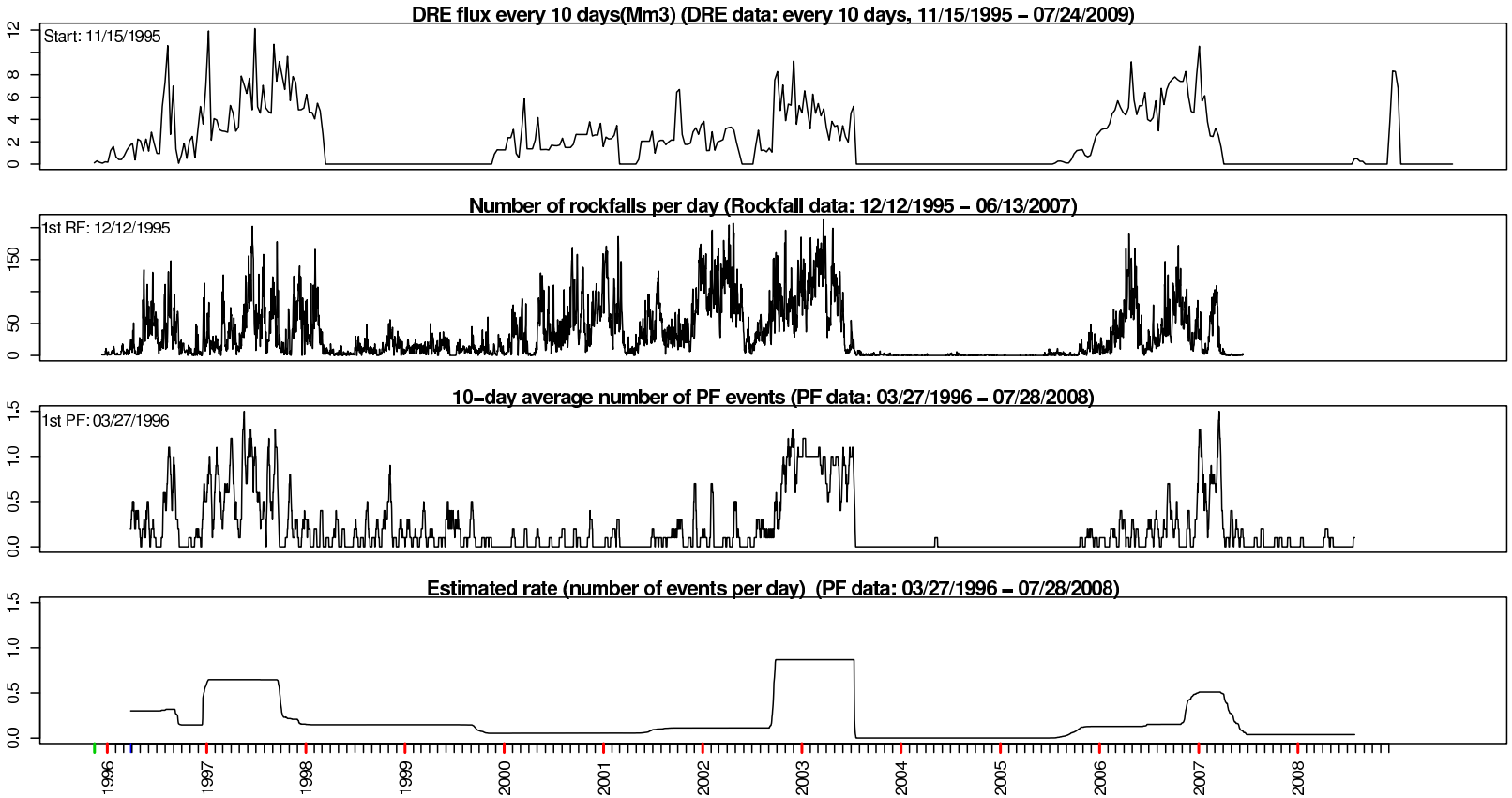
... & Current Conditions:

We're just beginning to Fuse data from Seismic, Dome shape, Degassing, and such. Here's an indication that **Rockfall Data** (from work of Duke PhD **Jianyu Wang**) may be useful precursors:



... & Current Conditions:

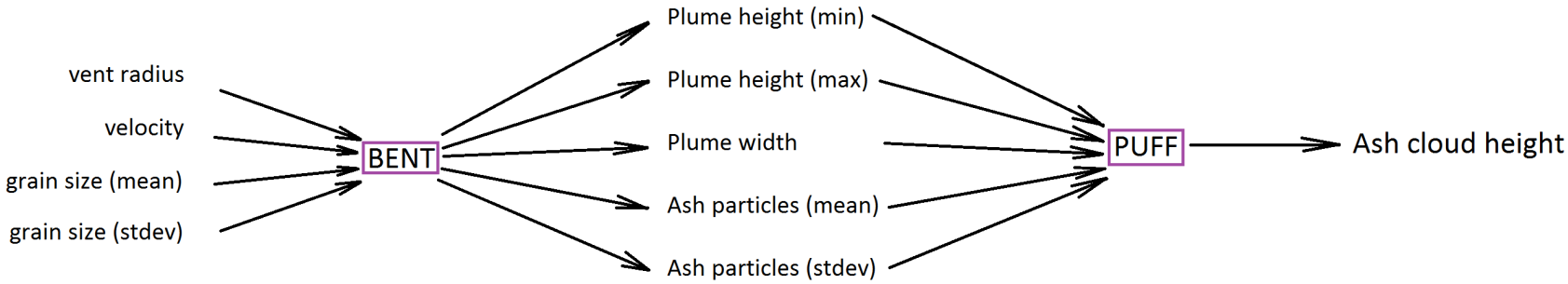
We're just beginning to Fuse data from Seismic, Dome shape, Degassing, and such. Here's an indication that **Rockfall Data** (from work of Duke PhD **Jianyu Wang**) may be useful precursors:



And Yet Again! **Flow Simulator** + **Stoch Model** +
Hier Friction Mod + **Precursors** ⇒ **Even Better Hazard Forecasts**

... & Consequences:

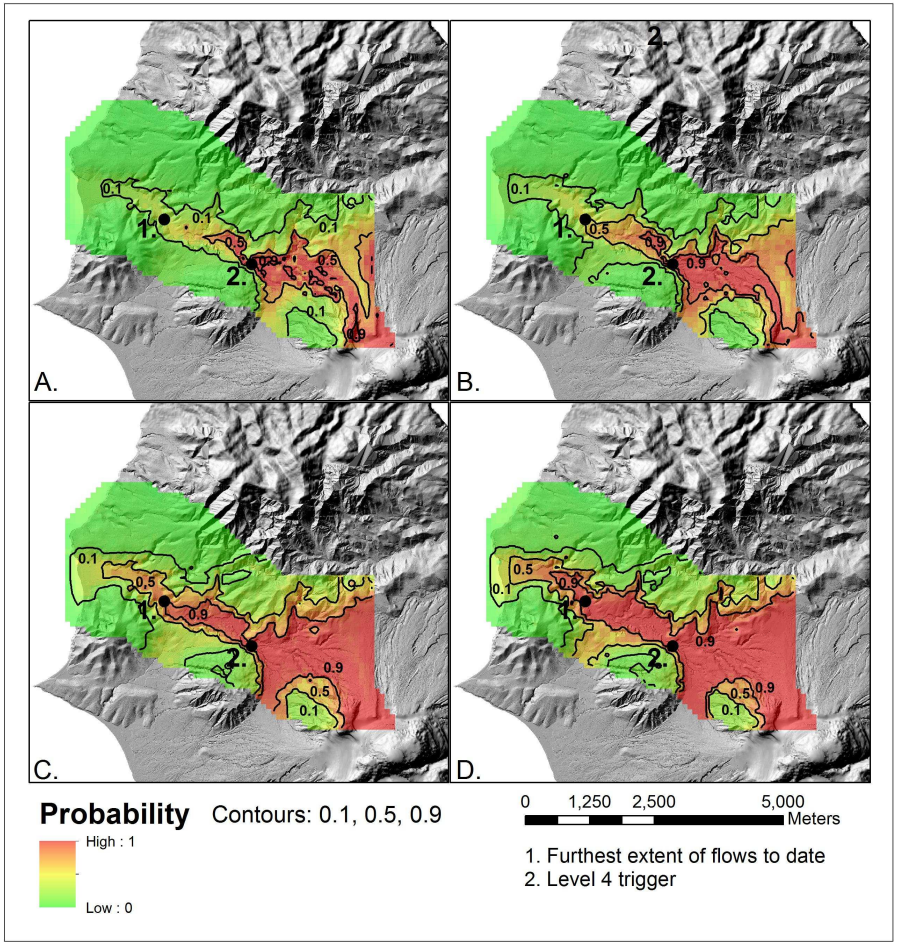
We've been forecasting **Hazard**, the *probability* of an adverse event in specified place and time. Often interest centers on **Risk**, the *expected cost* of such events reflecting both probability and frailty.
One approach, following recent Duke Ph.D. work of **Ksenia Kyzyurova**, is to fuse models of Hazard and of Consequence by **fusing their emulators**.
She studied coupling of a volcanic ash cloud generation model **Bent** with an atmospheric dispersion model **Puff** through their emulators:



and found this approach is far more efficient than building a single emulator for the compound model construct.

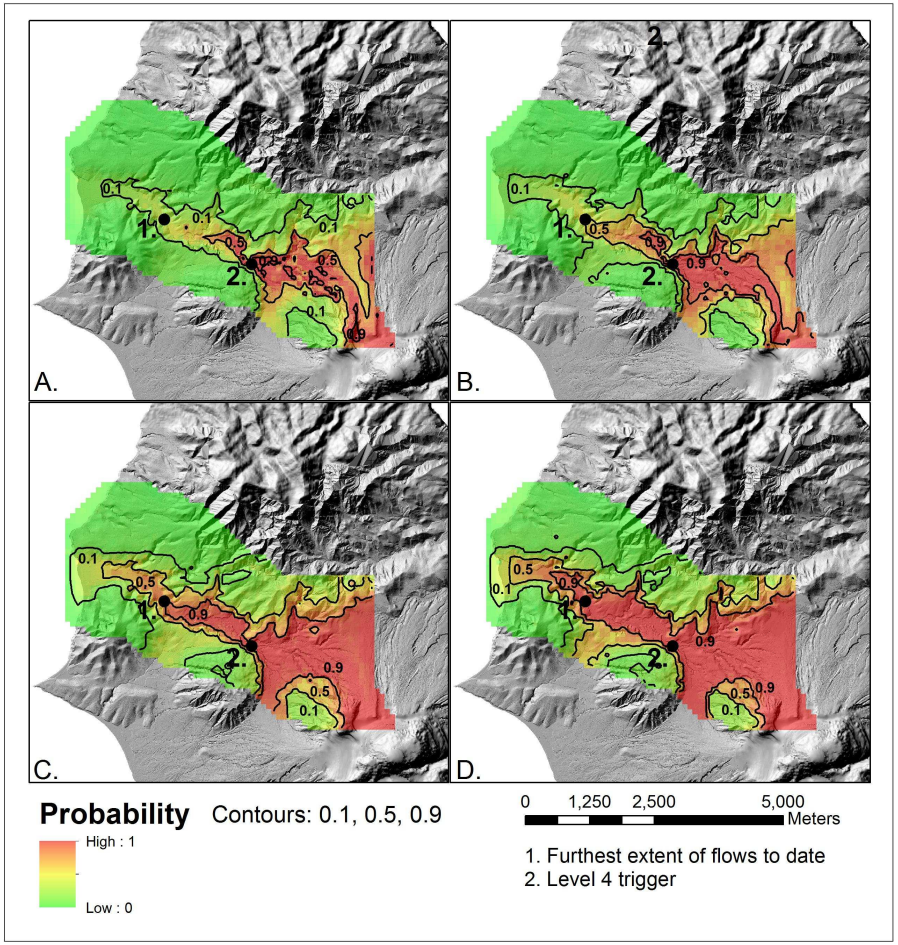
Back to Hazard Maps— where we are so far:

Here's one for the **Belham Valley** on Montserrat:



Back to Hazard Maps— where we are so far:

Here's one for the **Belham Valley** on Montserrat:



Actually here are *four* hazard contour maps, for different possible Current Conditions. Compare these to the (maps in current use)

Thanks!

More details (references, this talk in .pdf, related work) are available on request from

rlw@duke.edu.

Thanks to Jim Berger, SAMSI, NASA, and the NSF!

Thanks!

More details (references, this talk in .pdf, related work) are available on request from

rlw@duke.edu.

Thanks to Jim Berger, SAMSI, NASA, and the NSF!
And to my colleagues in this team effort!

Along with many current and former Ph.D. students...

Recorded at AIR Studios, Montserrat:

Over 500 albums were recorded at Sir George Martin's **Associated Independent Recording** (AIR) Studio in Montserrat between 1980 and 1995, by artists including:

Gerry Rafferty, Little River Band, Sheena Easton, Duran Duran, The Police, America, Rush, Black Sabbath, Elton John, Rolling Stones, Luther Vandross, Dire Straits, Lou Reed, Eric Clapton, The Fixx, Earth Wind & Fire, Keith Richards, Sting, Stray Cats, M1ke & The Mechanics, Indochine, Gregory Hines, Boy George, Status Quo, Roger Daltrey, Climax Blues Band, Art Garfunkle, Amy Grant, Jimmy Webb, UFO, OMD, James Taylor, Paul McCartney, Stray Cats, Art of Noise, Jimmy Buffet.

Buffet's song *Volcano* (refrain: *Where ya gonna go when the volcano blows?*) was written in 1979 on and about the Soufrière Hills Volcano on the island of Montserrat. The first SHV eruption in centuries came on July 18, 1995.

AIR studio, in Montserrat's "Exclusion Zone", has been abandoned for 25 years and lies in ruins.