New Challenges and Bayes:

The world of computer models

Susie Bayarri (Universitat de València)

with many many collaborators ...

Bayes Lectures 2012

Edinburgh, August 29 - 30, 2012

Computer models

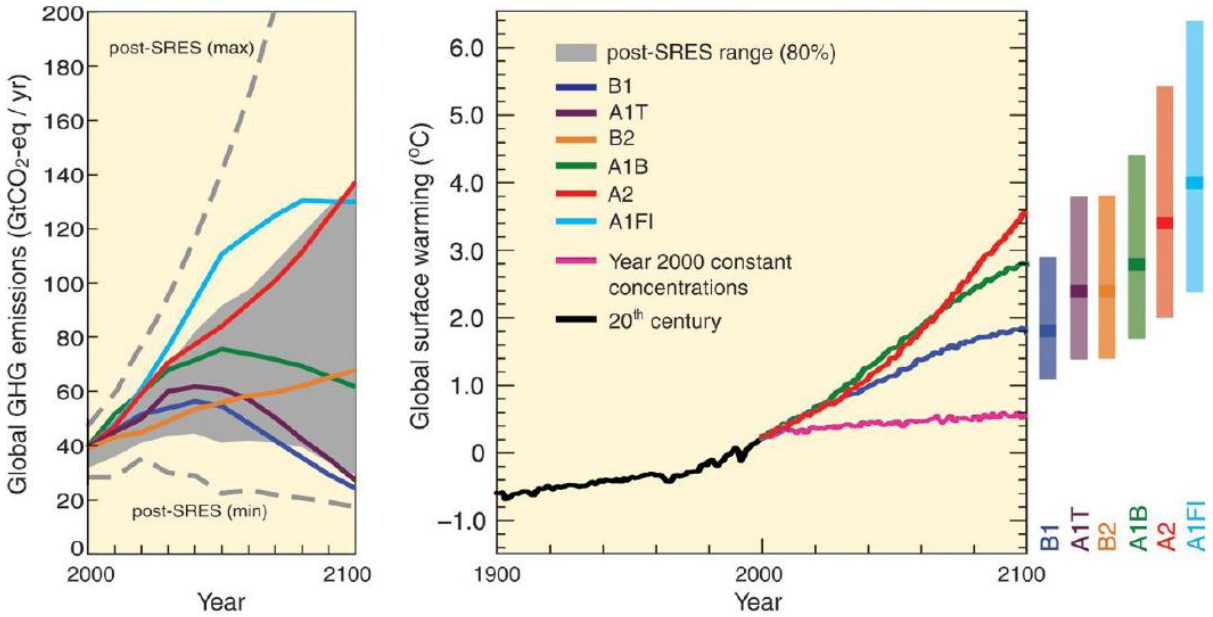
Simulators or computer models are intended as 'surrogates' (or at least good approximations) of reality

- Solutions of complex math/physics models (systems of ODE's, PDE's, ...) which try to mimic a real process
- For complex models, they have to be numerically solved (and hence the name *computer models*)
- CCM are (delayed) black boxes: feed specific values for the inputs $(\boldsymbol{x}, \boldsymbol{u})$, get back some output $y^M(\boldsymbol{x}, \boldsymbol{u})$, or computer model run at those inputs (predictions of the real process)
  - $\boldsymbol{x} \rightsquigarrow$ vector of controllable (known) inputs (speed of a crash)
  - $\boldsymbol{u} \rightsquigarrow$ calibration/tuning/random (unknown) parameters (friction/fudge factor/Richter index)

- Usually deterministic: if you run the computer model at the same inputs, you get the same outputs
- Complex computer models are expensive to run: a single run can take hours or even days (so we don't have many)
- outputs are complex; $y^M(\boldsymbol{x}, \boldsymbol{u})$ refers to a simple (here scalar) function of the outputs (the QoI)
- They are most common when physical data is scarce or non existent, but they are becoming ubiquitous (crashes, porous media, atmospheric ozone, health effects of pollutants, bomb deflagration, infections dynamics, engineering, weather, hurricanes, ecology models, high-energy physical collisions, social infections, traffic networks, ocean models, computer networks, social networks, engineering, medical devices, cell transport, PK/PD models ...)

Three views of computer modeling:

- It is the future of science, technology and society.
- It can be highly successful, but requires careful statistical validation.
- It is too difficult to be useful in most practical scenarios



Predictions from different climate models/scenarios (IPCC Assessment Report)

Uncertainties

There is usually considerable uncertainty in

- the "true" values of the $\boldsymbol{u}$ inputs
- the numerical implementation (approximation to the proposed math/physics model)
- the adequacy of the models to describe reality (the bias function)
- observation of reality (field data)
- the output of complex computer models at un-run inputs
- relation among 'similar' experiments and/or models (ensemble)
- ... and many others

Deal with them mostly with Bayesian Analysis

Why Bayes is so good in the analysis of computer models

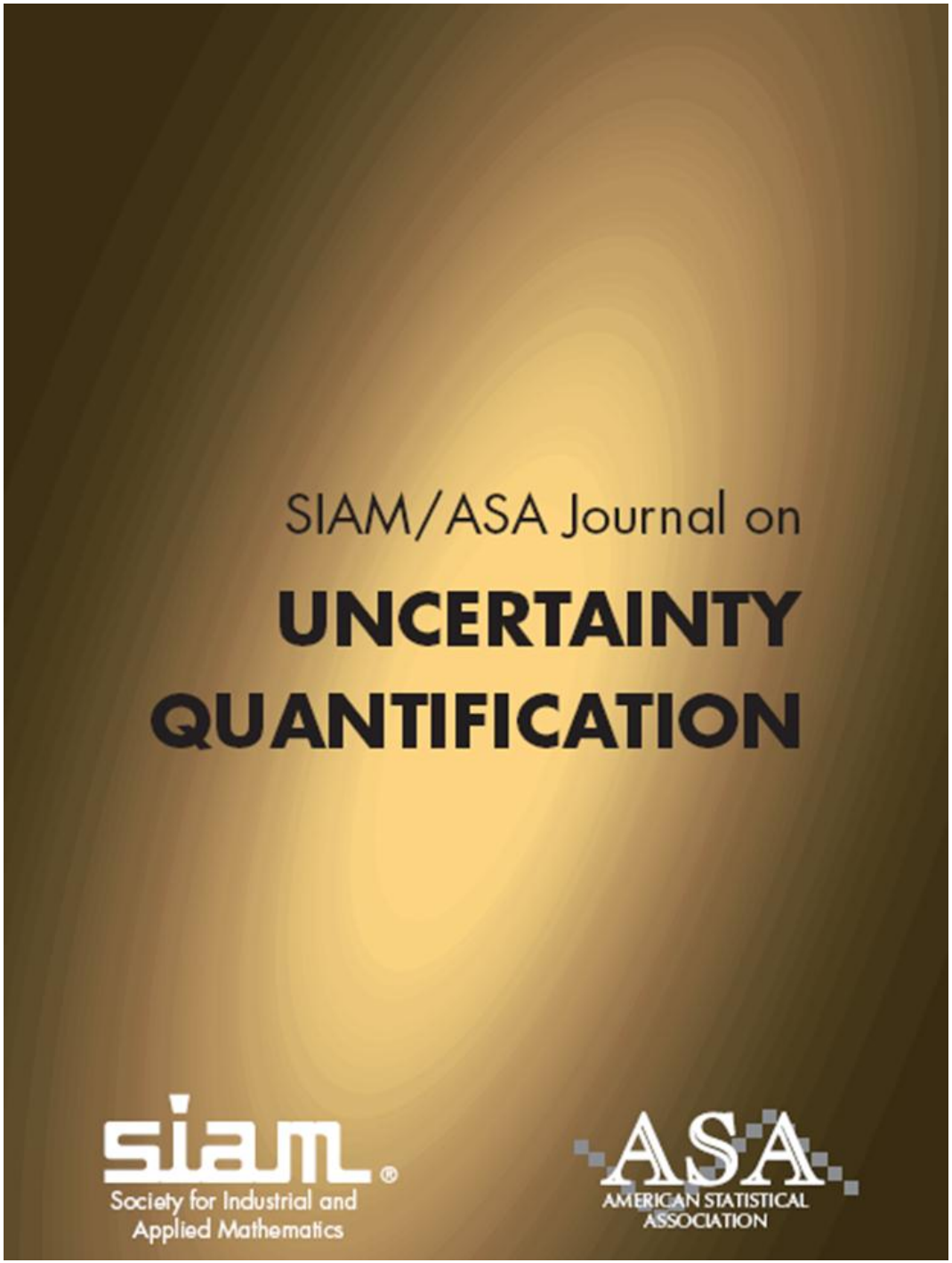
- Bayesian analysis can handle, quantify, update, combine and propagate all these uncertainties in the same analysis (probability rules)
- In Bayesian analysis all uncertainties are quantified through probability distributions $\rightsquigarrow$ no need to treat differently epistemic uncertainty from aleatoric uncertainty
- Bayesian analysis is about the only satisfactory way to go from P(data | inputs) to P(inputs | data) (methods not explicitly incorporating a prior distribution will not work)

- Applied math/Engineers/Physicists acknowledged uncertainty in their analysis (they called it *Uncertainty Quantification*); Statisticians acknowledged the role of deterministic models ... but very little interaction Math/Stat for a while
- Increased collaboration is changing the meaning (and the world) of Uncertainty Quantification

What is Uncertainty Quantification (UQ)?

Vague Definition: Uncertainty quantification is everything at the intersection of 'mathematical modeling of processes' with 'probability and statistics.'

- UQ has become a major area of science and statistics:
- Has put together SIAM (Society for Industrial and Applied Mathematics) and ASA (American Statistical Association)
  - Inaugural SIAM/ASA conference on UQ, on April 2-5, 2012 drew 417 participants
  - There is a new SIAM/ASA Journal on Uncertainty Quantification (www.siam.org/journals/juq.php)


SIAM/ASA Journal on
UNCERTAINTY
QUANTIFICATION
siam
Society for Industrial and
Applied Mathematics
ASA
AMERICAN STATISTICAL
ASSOCIATION

Four views on uncertainty quantification of computer models:

- Hard Core Modeler: "Don't bother me; I need every waking moment to work on the science/math/computation to improve the model."
- Hard Core Statistician: "Practical use of the model is irresponsible unless all sources of uncertainty have been properly accounted for."
- Soft Core Modeler:
  - "I'll talk to statisticians if it help's me to improve the model;"
  - "I'll consign some time and model runs to dealing with uncertainty."
- Soft Core Statistician: "What I want in terms of model, field data, and information about uncertainties is not possible, so I'll take what is available and try to do something."

Statistical areas of uncertainty quantification

- *Design:* Designing runs of the simulator
- *Emulation:* Approximating the simulator (surrogates, reduced order)
- *Sensitivity analysis and variance decomposition:* Determine sensitivity of simulator to inputs, often with goal of ignoring part of input space.
- *Diagnosis:* Detection of flaws in the simulator
- *Parameterization:* Incorporating probabilistic or statistical components
- *Inverse problems:* Determining tuning or calibration parameters of the simulator, or the initial states of the system
- *Output analysis:* Determining how stochastic inputs affect outputs of the simulator (uncertainty propagation)
- *Data assimilation:* Combining simulator runs with observational data for prediction (e.g., 'pseudo Kalman-filtering' as used in weather prediction)
- *Making decisions* using the computer model and UQ analysis
- *Validation:* Instead, estimating the discrepancy of the simulator from reality, and deciding if predictions are accurate enough for intended use.

Quantifying Risks of Geophysical Hazards*

with J.O. Berger, E. Calder, A. Patra, B. Pitman, E. Spiller, R. Wolpert

(Duke U, U Buffalo, Marquette U)

- A promising application of computer models (because it involves serious extrapolation)
- General methodology is applicable to many risks analyses
- It presents a novel use of an 'inverse problem', which is solved by a combination of computer models and Bayes

* *From a 2006-07 samsi Program on Development, Assessment and Utilization of Complex Computer Models. Partially supported with several NSF grants*

Assessing risks of extreme hazards

- Events producing mild damage occur periodically. In rare occasions, they are catastrophic (hurricanes, tsunamis, earthquakes, flooding, forest fires, pyroclastic flows, ...)
- Interest for now: Pr(at least a catastrophic event in the next T years) at certain locations
- Usual risk assessments are based on:
  - expert opinion $\rightsquigarrow$ but phenomena way too complex
  - statistical/probabilistic models $\rightsquigarrow$ but data way too scarce
  - computer implementations of math models describing the phenomena and extrapolating to unseeing situations $\rightsquigarrow$ but LOTS of uncertainties $\rightsquigarrow$ it needs statistics

What we do

Combine use of computer models and statistical models to assess the risk of a volcanic hazard. (Test-bed: pyroclastic flows of Soufrière Hills Volcano in the island of Montserrat.) We compute

Pr (a catastrophic inundation occurs in the next T years)

at specified locations, utilizing

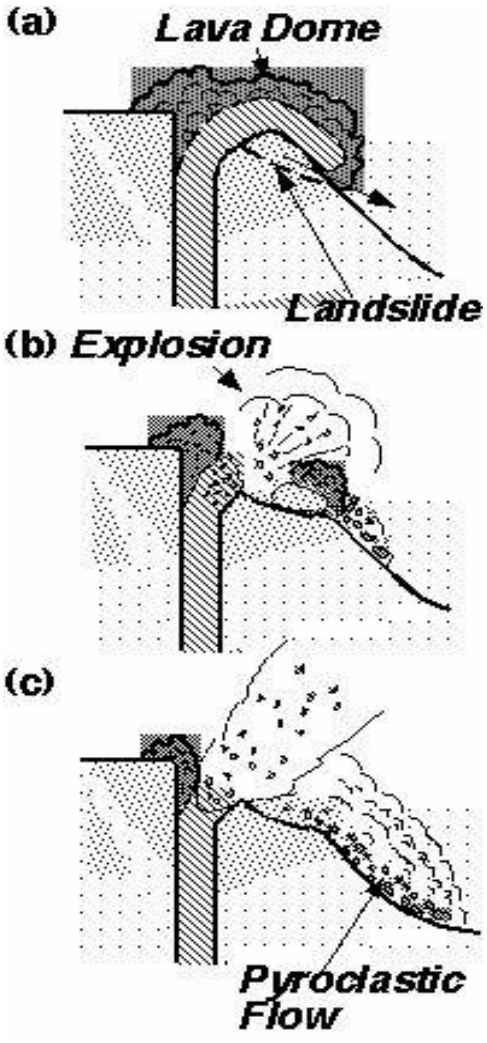
- computer implementations of mathematical models of flows to allow extrapolation to unseen situations;
- statistical models for needed stochastic inputs to the computer model, to calibrate unknown parameters of the computer model, and to account for uncertainties;
- a computational strategy for rare events, based on development of adaptive approximations to the computer model.

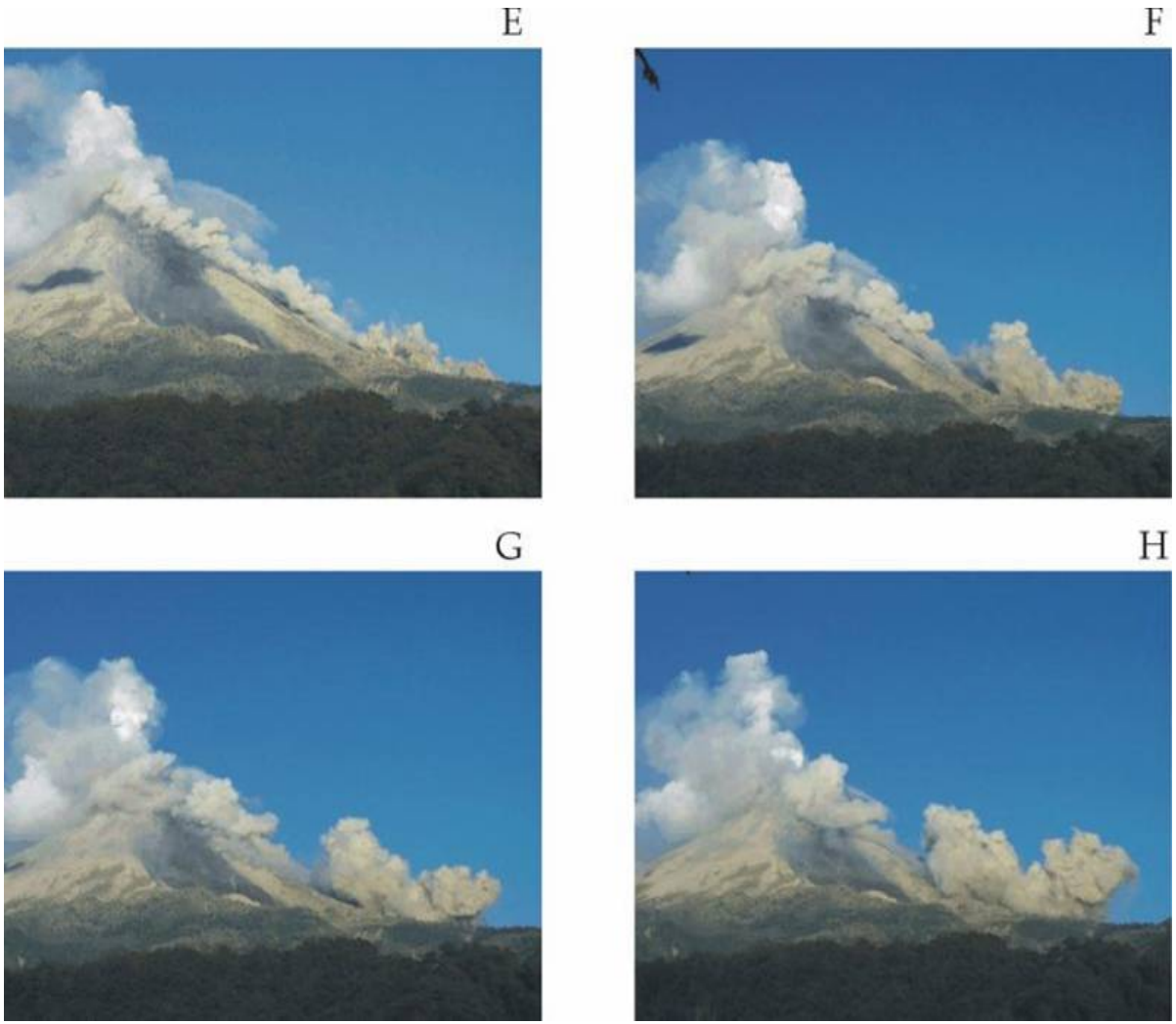
Although we present our methodology in the volcano hazard scenarios, it is quite general, applicable to assess risk of catastrophic events (probability of rare events) when:

- there is little or no data on the rare event
- a math model/simulator is available to describe the phenomena
- a statistical model is needed to feed the simulator, and there is data available (even if scarce and challenging) to fit such a model
- the simulator is too expensive to run to evaluate the probability of hazards by 'brute force'

Soufrière Hills Volcano

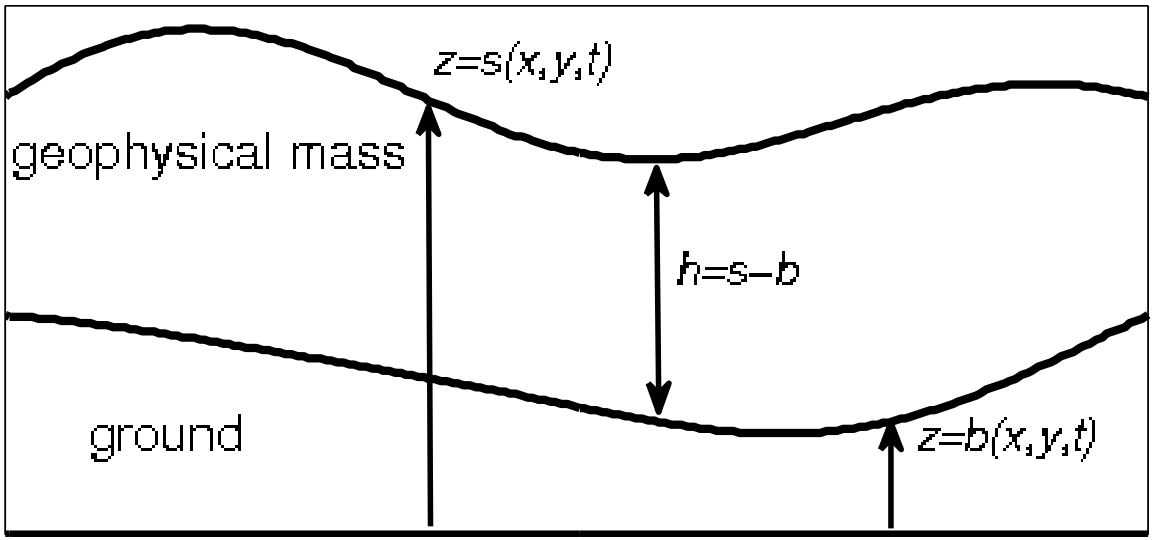
Pyroclastic flows





The Geophysical/Math Model

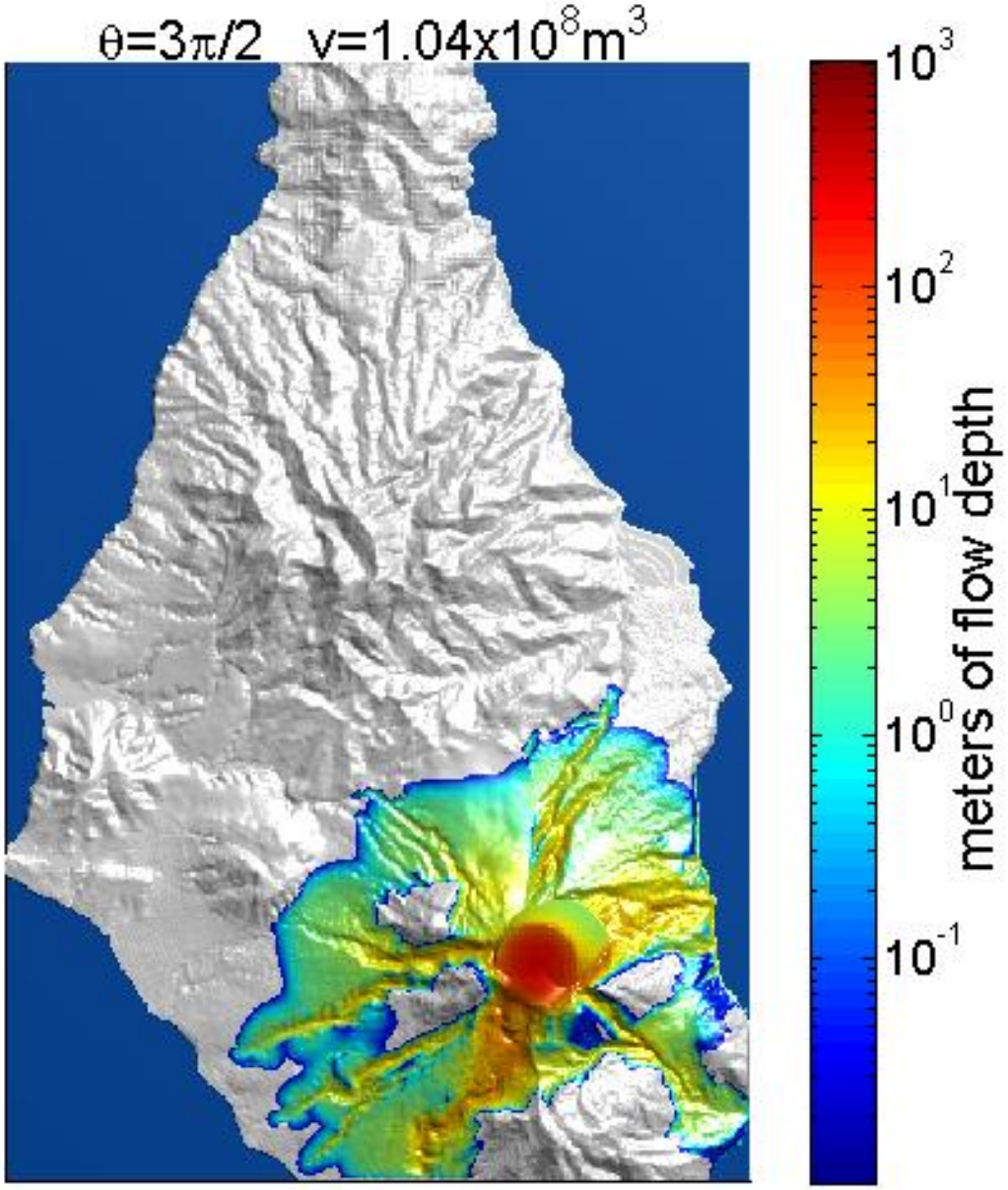
- Use 'thin layer' modeling $\rightsquigarrow$ system of PDE for the flow depth and the depth-averaged momenta.
- Important feature: Incorporates topographical data from a digital elevation map (DEM).



The Computer Model Implementation

TITAN2D (U Buffalo) computes solution to the math model

- Stochastic inputs whose randomness is the basis of the risk uncertainty (note: all inputs are denoted by $\boldsymbol{x}$ here; none controlable)
  - $x_1 =$ initial volume V (size of initial flow),
  - $x_2 =$ initial angle φ (direction of initial flow).
- Main deterministic inputs: internal friction, $x_3 =$ basal friction b (which is very uncertain), initial velocity (set to zero).
- Output: flow height and depth-averaged velocity at every grid point at every time step; we will focus on the maximum flow height at each grid point.
- Each run takes about 1 hour


θ=3π/2 v=1.04x10^8m^3
10^3
10^2
10^1
10^0
10^-1
meters of flow depth

a naïve risk assessment

- develop a distribution for the inputs $\boldsymbol{x} = (V, \varphi, b)$
  (needs to take into account frequency and severity)
- feed the simulator with simulations from this distribution
- estimate the probability of a "hit" by straight MC or importance sampling

This is unfeasible since

- simulator is very slow
- we are interested in the probabilities of very rare events

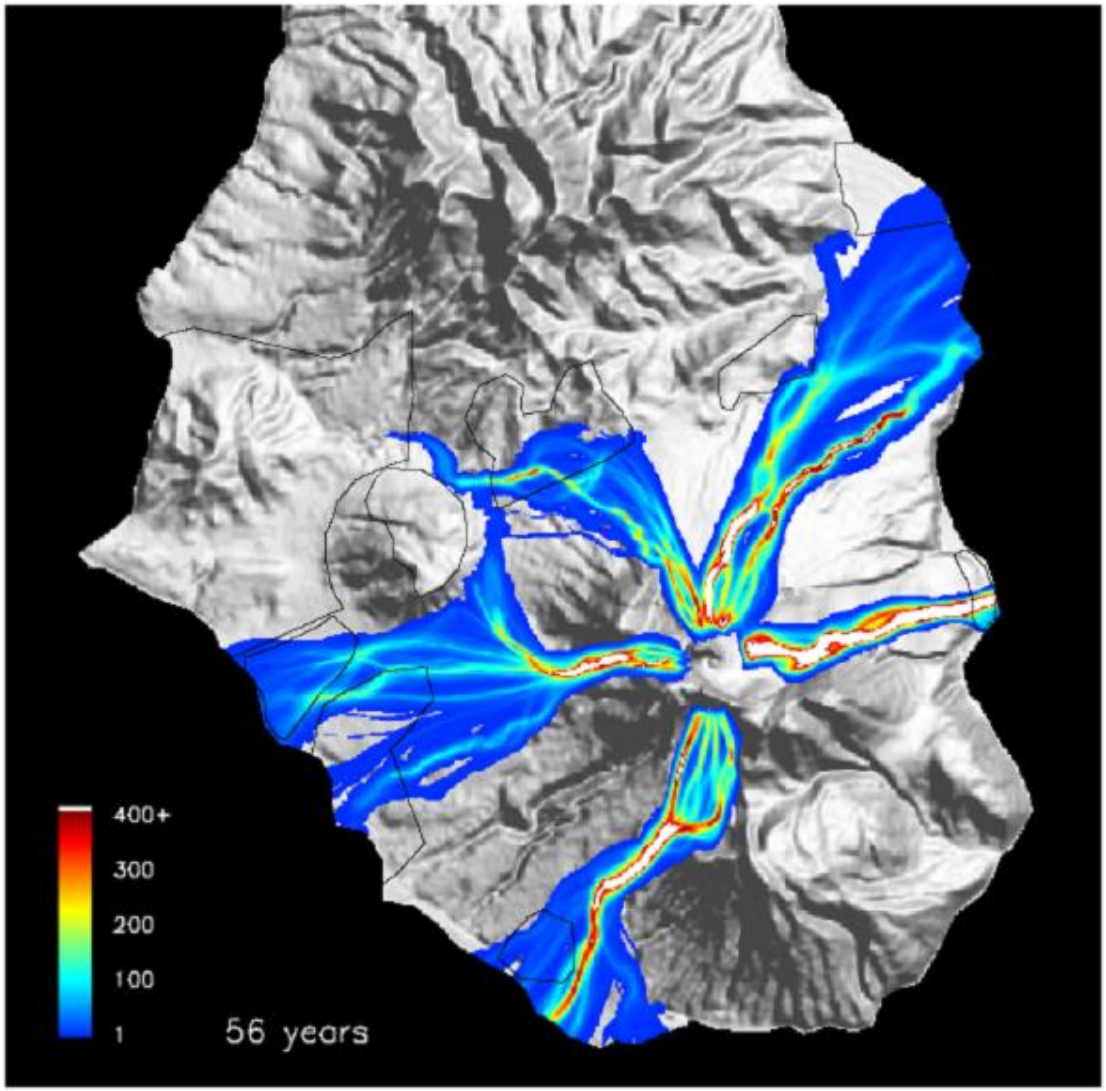


Fig. 4 PYROFLOW ensemble Monte Carlo simulation of pyroclastic flow and surge hazard from Soufriere Hills Volcano. The ensemble represents an equivalent of 56 years of typical 1996-98 activity and the colour scale shows how many times a given spot is overrun by flows and surges in that period. Four sources of flows are simulated: Gage's, Farrell's, Tar River Valley and Galway's.

Risk Assessment Part I: Defining a Catastrophe

- Let $y^M(\boldsymbol{x})$ be the computer model prediction, for input $\boldsymbol{x} \in \mathcal{X}$, of the characteristic that defines catastrophic events.

  SHV: $\boldsymbol{x} = (V, \varphi, b) \in \mathcal{X} = (0, \infty) \times [0, 2\pi) \times (0, \infty)$ and $y^M(\boldsymbol{x})$ = maximum height of the pyroclastic flow in downtown Plymouth (or airport) for an eruption of characteristics $\boldsymbol{x}$.

- Catastrophe occurs if $y^M(\boldsymbol{x}) \in \mathcal{Y}_C$.

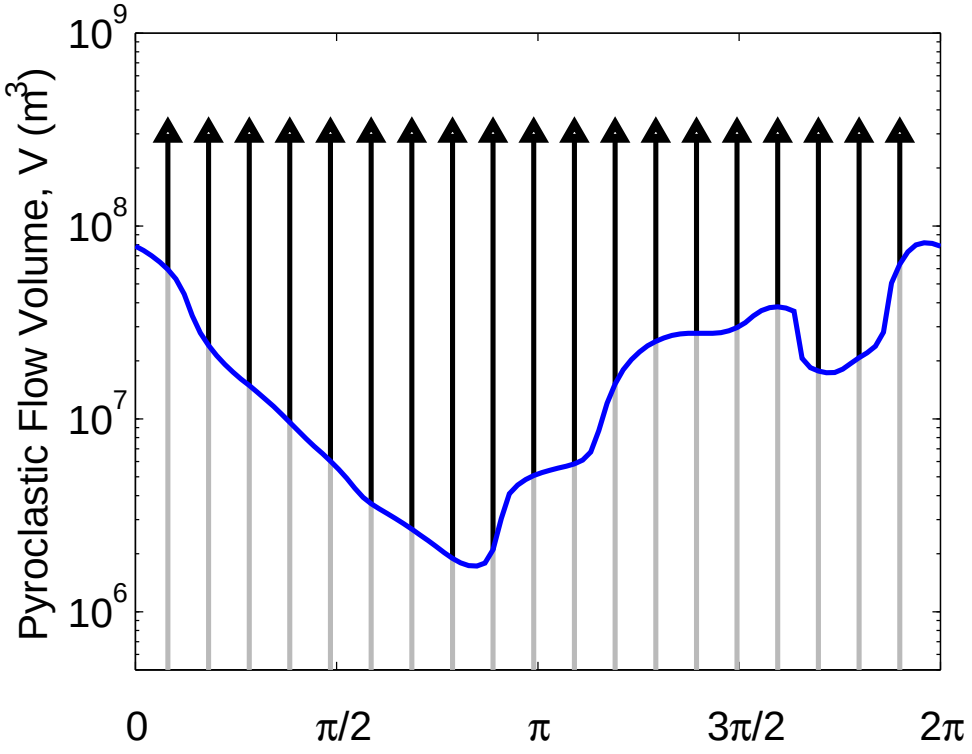
  SHV: Catastrophe if $\boldsymbol{x}$ is such that $y^M(\boldsymbol{x}) \geq 1\text{m}$.

- what kind of inputs produce an inundation of at least 1m in the location of interest? determine 'catastrophic region' $\mathcal{X}_C$ in the input space:

$$\mathcal{X}_C = \{\boldsymbol{x} \in \mathcal{X} \ : \ y^M(\boldsymbol{x}) \in \mathcal{Y}_C\}$$

SHV: $\mathcal{X}_C = \{(V, \varphi, b) \in \mathcal{X} \ : \ V > \Psi(\varphi, b)\}$, where the *critical contour*, separating catastrophic and benign events, is

$$\Psi = \Psi(\varphi, b) = \{\text{value of } V \text{ such that } y^M(V, \varphi, b) = 1\text{m}\}$$



Developing an emulator to help determine Ψ

- It is obviously impossible to determine the critical boundary Ψ with simulator's runs. We use instead an emulator/surrogate
- Emulators are very fast, statistical "approximations" (surrogates) to some of the outputs $y^M(\boldsymbol{x})$ of (slow) computer models.
- More specifically, an emulator
  - is fitted based on a set of runs of the computer model at specified "design points" $\boldsymbol{x} \in \mathcal{D}$;
  - is a statistical predictor of $y^M(\boldsymbol{x})$ for untried $\boldsymbol{x}$, which provides an estimate of the error incurred in the prediction;
  - exactly equals the computer model runs at the design points and interpolates at other values of $\boldsymbol{x}$.

localized emulator

- Here we construct a "localized" emulator, namely only around the Critical Region $\mathcal{X}_C$ (for us, critical contour Ψ)
- Localized emulators work better, and are easier to fit
- Initial Computer Model runs
  - Begin with a Latin hypercube statistical (space filling) design to select N design points in a large region $\mathcal{X} = [10^5 \mathrm{m}^3, 10^{9.5} \mathrm{m}^3] \times [0, 2\pi] \times [5, 25]$.
  - Run the computer model at these preliminary points.
  - For the purpose of fitting an emulator to find $\mathcal{X}_C$, keep only design points $\mathcal{D}$ in a region 'close' to $\mathcal{X}_C$:

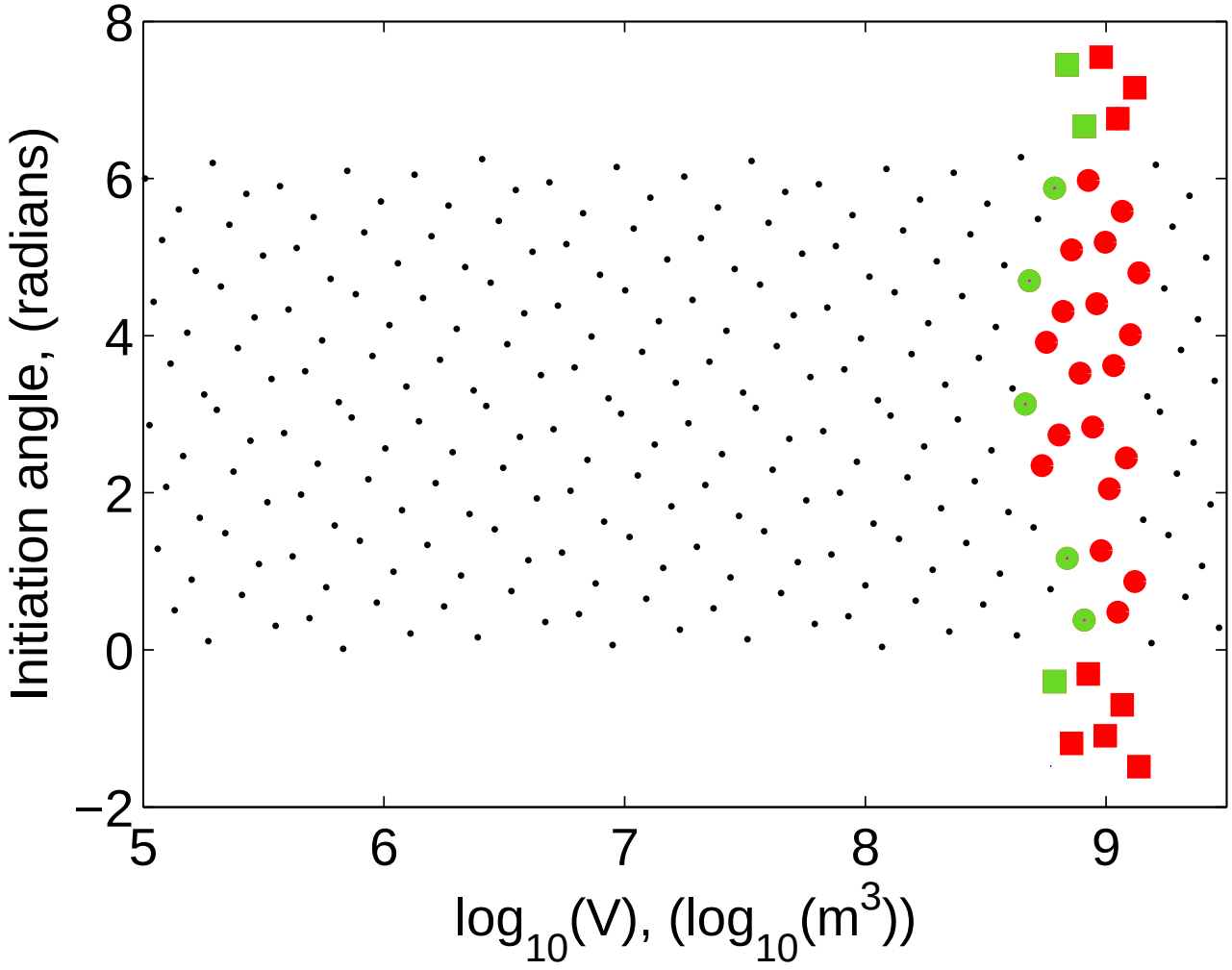


Figure 1: Red dots: $y^M(V, \varphi, 15) > 0$ at Plymouth. Green dots: $y^M(V, \varphi, 15) = 0$.

Remember the goal: find the contour where $y^M(V, \varphi, b) = 1$.

Gaussian process emulators in the $\mathcal{X}_C$ region

- Since we are interested in regions where the flow is small (1m), we fit an emulator to $\tilde{y}^M(\boldsymbol{x}) = \log(y^M(\boldsymbol{x}) + 1)$. Let $\widetilde{\boldsymbol{y}}$ be the transformed vector of computer model runs $y^M(\boldsymbol{x})$ for $\boldsymbol{x} \in \mathcal{D}$.
- In any given analysis, $\tilde{y}^M(\boldsymbol{x})$ is effectively an unknown function (known only for $\boldsymbol{x} \in \mathcal{D}$) $\rightsquigarrow$ assess a prior; standard in the field is use of GaSP, a Gaussian process
- Interpret a GaSP as saying Multivariate Normal. Indeed, the unknown function $\tilde{y}^M(\dot{}) \sim GaSP$ if the joint distribution of any finite set of L realizations $(\tilde{y}^M(\boldsymbol{x}_1), \ldots, \tilde{y}^M(\boldsymbol{x}_L))$ is a multivariate normal with mean a variance as specified by the mean function and the covariance function

Our GaSP choices:

We followed basically (but not entirely) the standard choices in the area. Specifically our GaSP has:

- mean $\beta + m\,V$ (a constant mean is often used)
  note: we expect monotonicity in V, but not φ (b later)
- variance σ_z^2;
- a product exponential correlation structure, i.e., for any two $\boldsymbol{x}_i = (V_i, \varphi_i, b_i), \boldsymbol{x}_j = (V_j, \varphi_j, b_j) \in \mathcal{D}$, the correlation matrix is $\boldsymbol{R} = [R_{ij}]$:

$$R_{ij} = \exp(-\theta_V |V_i - V_j|^{\alpha_V}) \exp(-\theta_\varphi |\varphi_i - \varphi_j|^{\alpha_\varphi}) \exp(-\theta_b |b_i - b_j|^{\alpha_b}),$$

where $\theta's$ ($\alpha's$) are range (smoothness) parameters.

- this is the usual correlation function, but here initiation angle is periodic ($\varphi = 1$ is the same as $\varphi = 361$)
- there are several ad-hoc solution. We formally modify the factor corresponding to φ so that it is periodic (and defines a positive definite correlation function):

$$\sum_{k=-\infty}^{\infty} \exp\big(-\beta_\varphi|\varphi_i - \varphi_j + 2\pi k|^{\alpha_\varphi}\big)/c(\beta_\varphi, \alpha_\varphi)$$

with normalizing constant $c(\beta_b, \alpha_b) = 1 + 2\sum_{k=1}^{\infty} \exp(-\beta_b|2\pi k|^{\alpha_b})$.
In most cases, a three term approximation is sufficient

- These assessments produce the likelihood function of the unknown $\boldsymbol{\theta} = (\theta_V, \theta_\varphi, \theta_b, \alpha_V, \alpha_\varphi, \alpha_b, \sigma_z^2, \beta, m))$ for the given data $\tilde{\boldsymbol{y}}$ (the computer model runs at inputs in $\mathcal{D}$) given by

$$p(\tilde{\boldsymbol{y}} \mid \boldsymbol{\theta}) = \frac{1}{2\pi\sigma_z^2|\mathbf{R}|^{\mathbf{1/2}}} \exp[-\frac{1}{2\sigma_z^2}(\boldsymbol{y} - \beta\mathbf{1} - m\mathbf{V})^T\mathbf{R}^{-1}(\boldsymbol{y} - \beta\mathbf{1} - m\mathbf{V})]$$

Traditional (pre-Bayesians) emulator

- Compute MLE $\widehat{\boldsymbol{\theta}}$
- EB posterior $p(\tilde{y}^M(\cdot) \mid \widehat{\boldsymbol{\theta}}, \tilde{\boldsymbol{y}})$ is another GaSP with mean $\hat{y}^*(\boldsymbol{x}^*)$ and variance $\hat{s}^2(\boldsymbol{x}^*)$ at any input $\boldsymbol{x}^* = (V^*, \varphi^*)$, given by usual kriging expressions

$$\hat{y}^*(\boldsymbol{x}^*) = \hat{\beta} + \widehat{m} V^* + \widehat{\mathbf{r}}' \widehat{\mathbf{R}}^{-1} (\mathbf{y}^g - \hat{\beta}\mathbf{1} - \widehat{m}\mathbf{V}),$$

$$\hat{s}^2(\boldsymbol{x}^*) = \hat{\sigma}_z^2 (1 - \widehat{\mathbf{r}}' \widehat{\mathbf{R}}^{-1} \widehat{\mathbf{r}} + \frac{(1 - \mathbf{1}' \widehat{\mathbf{R}}^{-1} \widehat{\mathbf{r}})^2}{\mathbf{1}' \widehat{\mathbf{R}}^{-1} \mathbf{1}}),$$

  where $\mathbf{r} = (R(\boldsymbol{x}^*, \boldsymbol{x}_1), \ldots, R(\boldsymbol{x}^*, \boldsymbol{x}_N))'$ with $\boldsymbol{x}_i \in \mathcal{D}$.

- This (posterior) GaSP was the Plug-in emulator used for years
- We improve it in several ways with similar computational effort.

Handling the unknown hyperparameters $\boldsymbol{\theta} = (\theta_V, \theta_\varphi, \theta_b, \alpha_V, \alpha_\varphi, \alpha_b, \sigma_z^2, \beta, m)$

- Deal with the crucial parameters (σ_z^2, β, m) via a fully Bayesian analysis (here an extension of Kriging) using objective priors: $\pi(\beta) \propto 1$, $\pi(m) \propto 1$, and $\pi(\sigma_z^2) \propto 1/\sigma_z^2$;
- Use a "good estimate" (more later) $\hat{\boldsymbol{\xi}}$ of the correlation (nuissance) parameters: $\boldsymbol{\xi} = (\theta_V, \theta_\varphi, \theta_b, \alpha_V, \alpha_\varphi, \alpha_b)$ then $\boldsymbol{R}(\hat{\boldsymbol{\xi}})$ is completely specified (big simplification).
  - A fully Bayesian analysis, accounting for uncertainty in $\hat{\boldsymbol{\xi}}$, is certainly doable, but it is difficult and rarely affects the final answer significantly because of confounding of variables.

The estimate of $\boldsymbol{\xi} = (\theta_V, \theta_\varphi, \theta_b, \alpha_V, \alpha_\varphi, \alpha_b)$

- MLE fitting of $\boldsymbol{\xi}$ has enormous problems: we have given up on it.

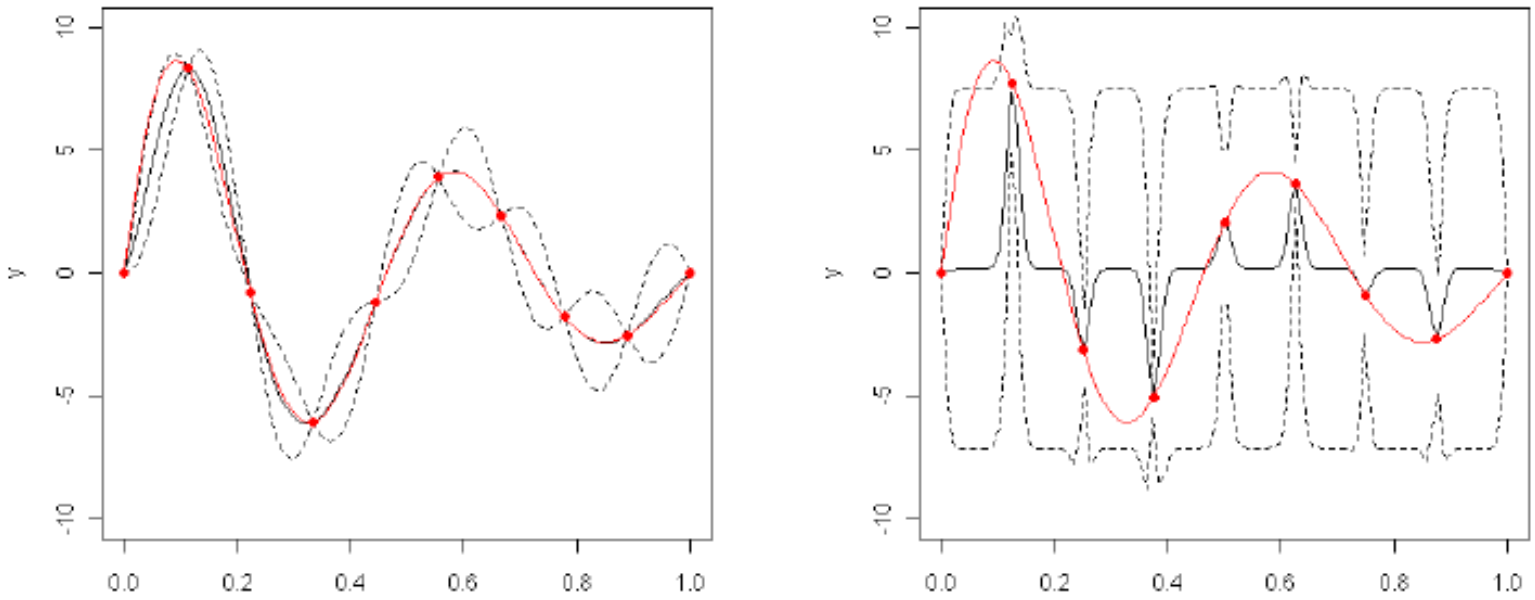


Figure 2: GASP fit to a damped sine wave for m=10 (left) and m=9 (right) points.

- A big improvement is finding the marginal MLE of $\boldsymbol{\xi}$ from the marginal likelihood for $\boldsymbol{\xi}$, found by integrating out over the objective prior $\pi(\beta, m, \sigma_z^2) = 1/\sigma_z^2$; the expression is

$$L(\boldsymbol{\xi}) \propto |\boldsymbol{R}(\boldsymbol{\xi})|^{-\frac{1}{2}} |\mathbf{X}'\boldsymbol{R}(\boldsymbol{\xi})^{-1}\mathbf{X}|^{-\frac{1}{2}} (S^2(\boldsymbol{\xi}))^{-(\frac{n-q}{2})} \; ;$$

  - $\mathbf{X} = (\mathbf{1}, \boldsymbol{V})$ is the design matrix for the linear parameters, $\boldsymbol{\mu} = (\beta, m)$ (having dimension $q = 2$), $\mathbf{1}$ is the column vector of ones, and $\boldsymbol{V}$ is the vector of volumes in the data set
  - $S^2(\boldsymbol{\xi}) = (\widetilde{\boldsymbol{y}} - \mathbf{X}\hat{\boldsymbol{\mu}})'\boldsymbol{R}(\boldsymbol{\xi})^{-1}(\widetilde{\boldsymbol{y}} - \mathbf{X}\hat{\boldsymbol{\mu}})$
  - $\hat{\boldsymbol{\mu}} = (\mathbf{X}'\boldsymbol{R}(\boldsymbol{\xi})^{-1}\mathbf{X})^{-1}\boldsymbol{R}(\boldsymbol{\xi})^{-1}\widetilde{\boldsymbol{y}}$

- An even bigger improvement arises by finding the posterior mode from $L(\boldsymbol{\xi})\pi^R(\boldsymbol{\xi})$, where $\pi^R(\boldsymbol{\xi})$ is the reference prior for $\boldsymbol{\xi}$ (Paulo, 2005 AOS). Note that it is computationally expensive to work with the reference posterior in an MCMC, but using it for a single maximization to determine the posterior mode is easy.
- The reference prior for $\boldsymbol{\xi}$ is $\pi^R(\boldsymbol{\xi}) \propto |I^\star(\boldsymbol{\xi})|^{1/2}$, where

$$I^\star(\boldsymbol{\xi}) = \begin{pmatrix} (n-q) & \mathrm{tr}\mathbf{W}_1 & \mathrm{tr}\mathbf{W}_2 & \cdots & \mathrm{tr}\mathbf{W}_p \\ & \mathrm{tr}\mathbf{W}_1^2 & \mathrm{tr}\mathbf{W}_1\mathbf{W}_2 & \cdots & \mathrm{tr}\mathbf{W}_1\mathbf{W}_p \\ & & \ddots & \cdots & \vdots \\ & & & & \mathrm{tr}\mathbf{W}_p^2 \end{pmatrix}$$

$$\mathbf{W}_k = \left(\frac{\partial \boldsymbol{\Sigma}}{\partial \xi_k}\right) \boldsymbol{R}(\boldsymbol{\xi})^{-1}[\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\boldsymbol{R}(\boldsymbol{\xi})^{\mathbf{X}})^{-1}\mathbf{X}'\boldsymbol{R}(\boldsymbol{\xi})^{-1}],$$

$q = 2$ being the dimension of $\boldsymbol{\mu}$ and $p = 6$ the dimension of $\boldsymbol{\xi}$.

The posterior predictive distribution at input x^*, conditional on $\widetilde{y}$ and $\hat{\xi}$, yields the final emulator (in transformed space) at x^*

$$\tilde{y}^M(\boldsymbol{x}^*) \mid \widetilde{\boldsymbol{y}},\ \hat{\boldsymbol{\xi}} \ \sim\ t(\, y^*(\boldsymbol{x}^*),\, s^2(\boldsymbol{x}^*),\, N-2)\,,$$

the t-distribution with $N-2$ degrees of freedom and parameters

$$y^*(\boldsymbol{x}^*) = \boldsymbol{r}^T\boldsymbol{R}^{-1}\widetilde{\boldsymbol{y}} + \frac{\mathbf{1}^T\boldsymbol{R}^{-1}\widetilde{\boldsymbol{y}}}{\mathbf{1}^T\boldsymbol{R}^{-1}\mathbf{1}}(1-\boldsymbol{r}^T\boldsymbol{R}^{-1}\mathbf{1}) + \frac{\widetilde{\boldsymbol{V}}^T\boldsymbol{R}^{-1}\widetilde{\boldsymbol{y}}}{\widetilde{\boldsymbol{V}}^T\boldsymbol{R}^{-1}\widetilde{\boldsymbol{V}}}(\widetilde{V}^* - \boldsymbol{r}^T\boldsymbol{R}^{-1}\widetilde{\boldsymbol{V}})$$

$$s^2(\boldsymbol{x}^*) = \left[(1-\boldsymbol{r}^T\boldsymbol{R}^{-1}\boldsymbol{r}) + \frac{(1-\boldsymbol{r}^T\boldsymbol{R}^{-1}\mathbf{1})^2}{(\mathbf{1}^T\boldsymbol{R}^{-1}\mathbf{1})} + \frac{(\widetilde{V}^* - \boldsymbol{r}^T\boldsymbol{R}^{-1}\widetilde{\boldsymbol{V}})^2}{(\widetilde{\boldsymbol{V}}^T\boldsymbol{R}^{-1}\widetilde{\boldsymbol{V}})}\right]$$

$$\times\ \frac{1}{N-2}\left[(\widetilde{\boldsymbol{y}})^T\boldsymbol{R}^{-1}\widetilde{\boldsymbol{y}} - \frac{(\mathbf{1}^T\boldsymbol{R}^{-1}\widetilde{\boldsymbol{y}})^2}{\mathbf{1}^T\boldsymbol{R}^{-1}\mathbf{1}} - \frac{(\widetilde{\boldsymbol{V}}^T\boldsymbol{R}^{-1}\widetilde{\boldsymbol{y}})^2}{\widetilde{\boldsymbol{V}}^T\boldsymbol{R}^{-1}\widetilde{\boldsymbol{V}}}\right],$$

where $\widetilde{V}_i = V_i - V_R$, $\widetilde{V}^* = V^* - V_R$, $V_R = \mathbf{1}^T\boldsymbol{R}^{-1}\boldsymbol{V}/\mathbf{1}^T\boldsymbol{R}^{-1}\mathbf{1}$, and $\boldsymbol{r}^T = (R(\boldsymbol{x}^*, \boldsymbol{x}_1), \ldots, R(\boldsymbol{x}^*, \boldsymbol{x}_N))$.

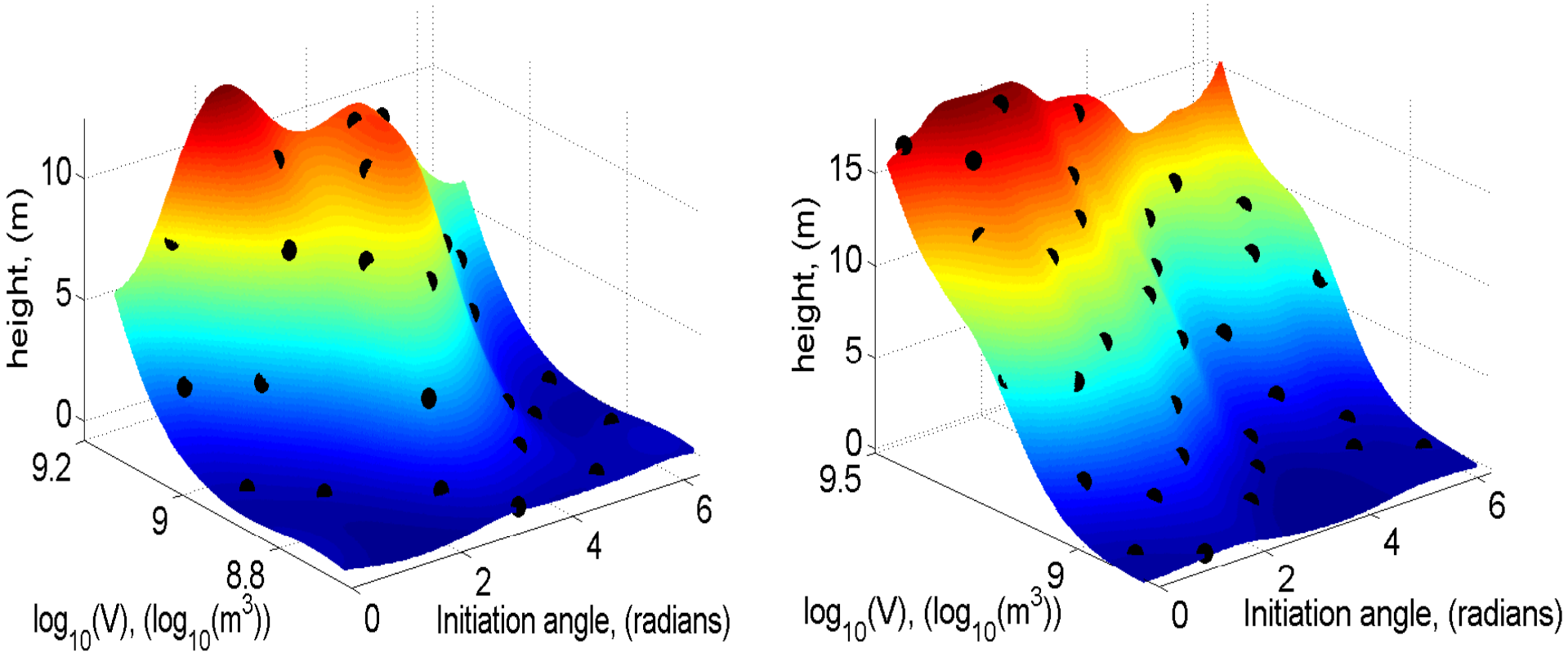
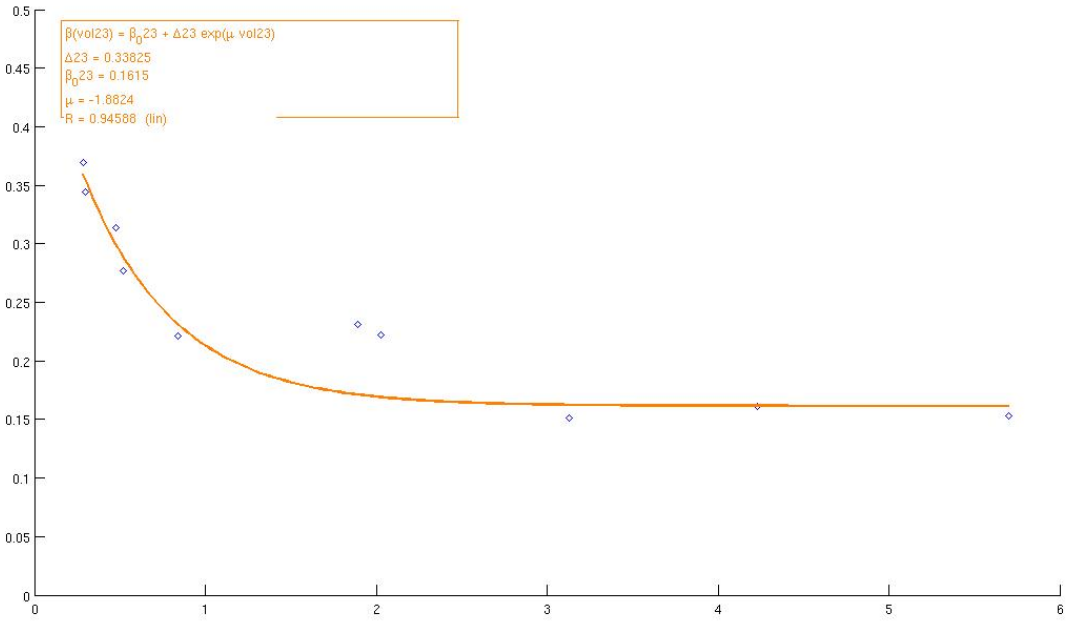


Figure 3: Median of the emulator, transformed back to the original space. Left: Plymouth, Right: Bramble Airport. Black points: max-height simulation outputs at design points.

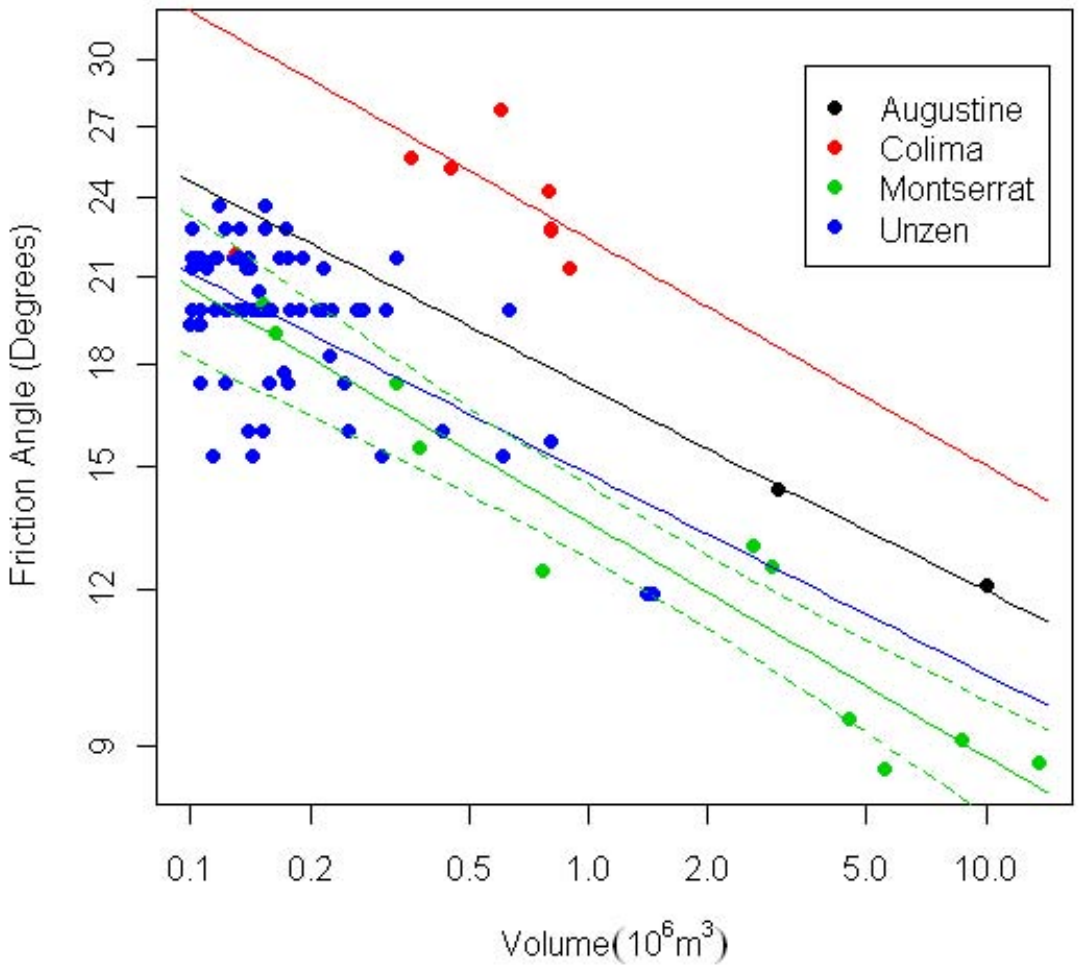
One more problem: b is highly dependent on V:



- for each such function, replace b in the emulator by this fucntion, thus the emulator becomes only a function of (V, φ)
- but $b(V)$ is very uncertain, and each possible function produces a different critical contour (inserting b in the T2D was not possible)
- we perform a Hierarchical Bayes Analysis to "borrow information" from similar volcanos, and account for uncertainty

Volume/Basal-Friction Relationship

Data from Sarah Ogburn; hierarchical Bayes analysis by Danilo Lopes

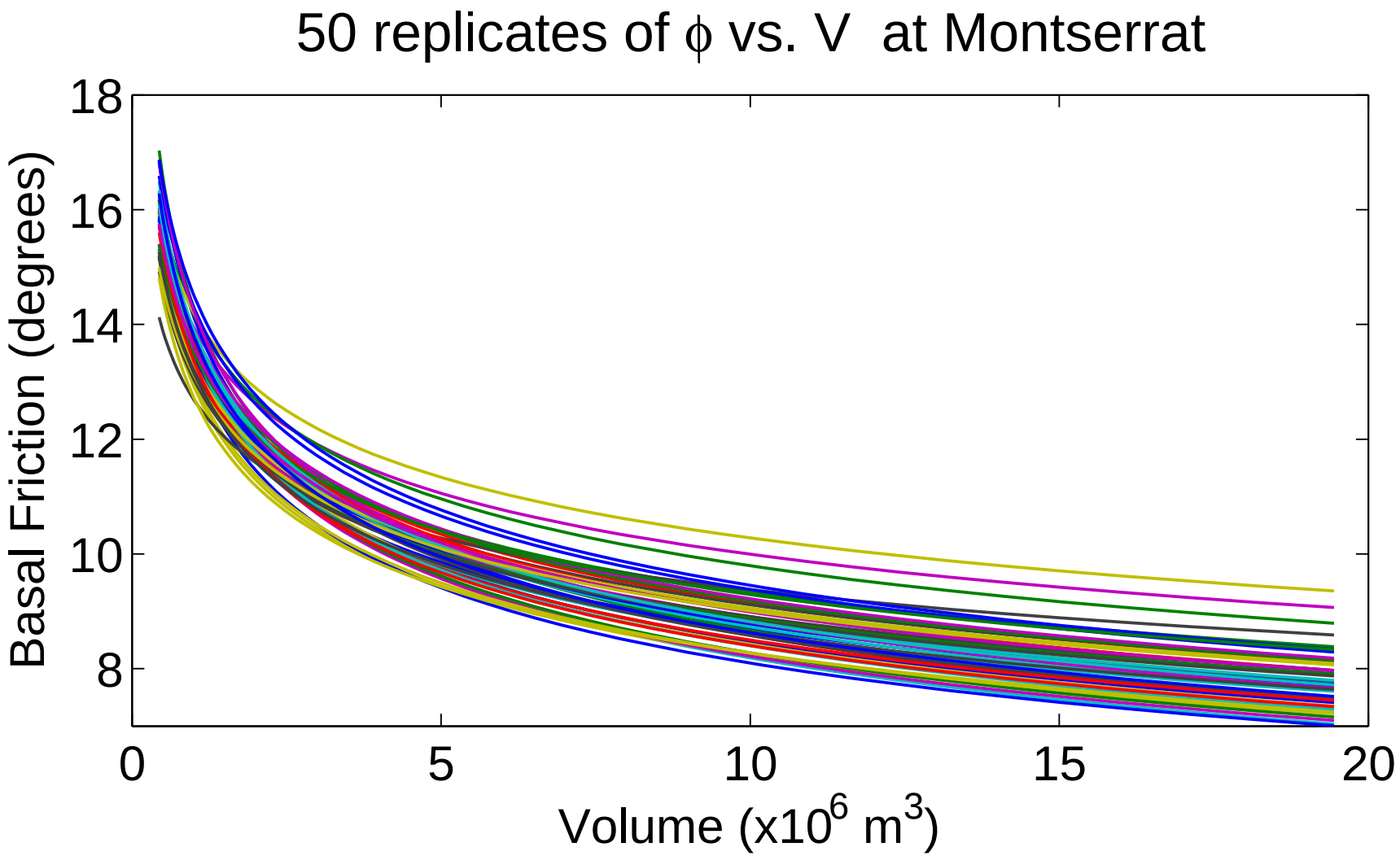


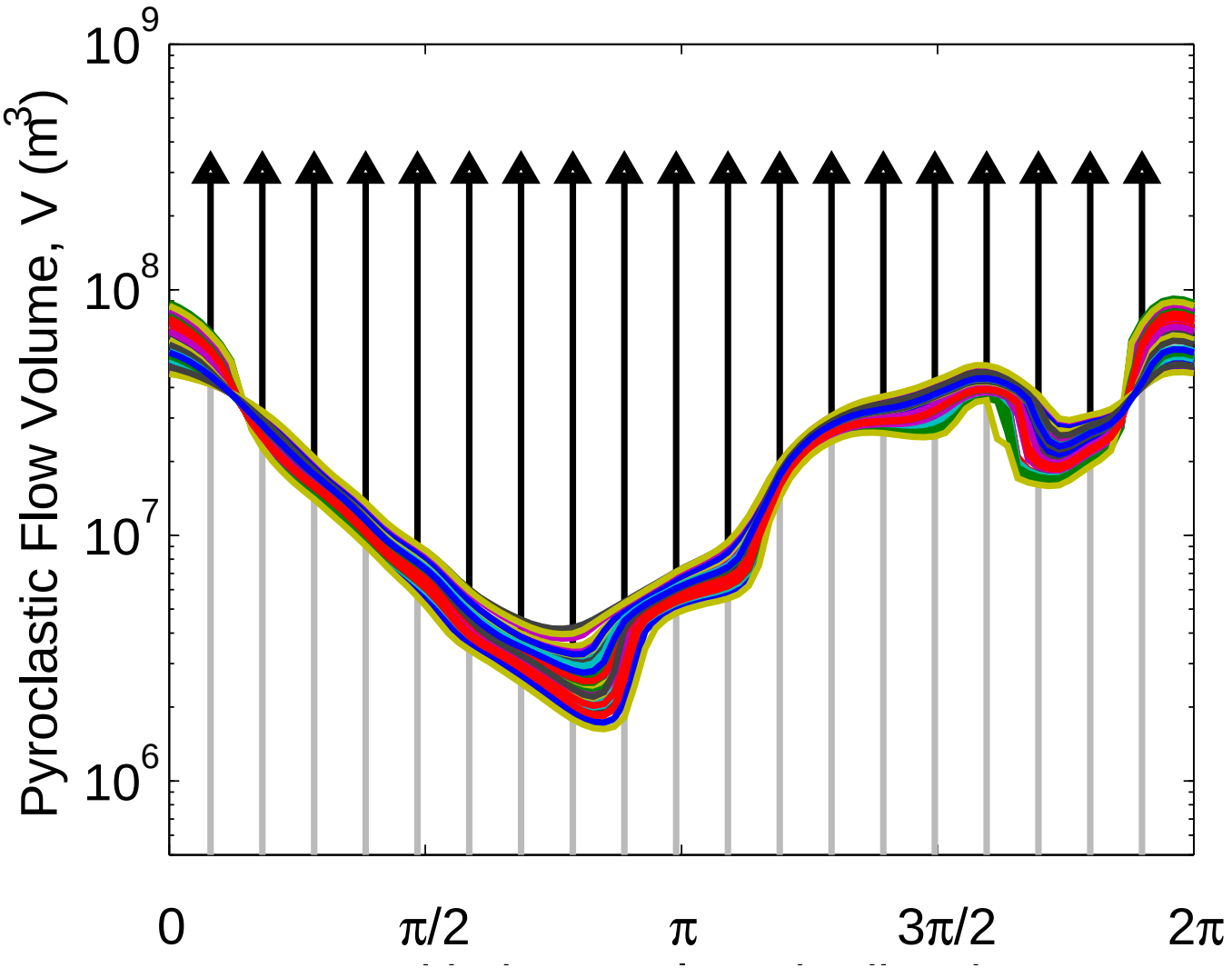
Hierarchical Modeling:

- For the j^{th} PF at i^{th} volcano, model the relationship of basal friction angle ϕ_{ij} to volume V_{ij} by

$$\log \tan \phi_{ij} = a_i + b_i \log V_{ij} \, .$$

- Assume that the a_i and b_i from different volcanoes arise from normal distributions $N(\mu_a, \sigma_a^2)$ and $N(\mu_b, \sigma_b^2)$.
- Assign prior distributions to $\mu_a, \sigma_a^2, \mu_b, \sigma_b^2$ and find the posterior distribution.
- Generate a sample of (a, b) for Montserrat, and utilize the resulting volume/basal friction curves when required in the emulator to determine the critical contours $\Psi(\varphi \mid a, b)$.


50 replicates of ϕ vs. V at Montserrat
Basal Friction (degrees)
18
16
14
12
10
8
0
5
10
15
20
Volume (x10^6 m^3)


Pyroclastic Flow Volume, V (m3)
10^9
10^8
10^7
10^6
0
π/2
π
3π/2
2π

Risk Assessment Part II: Probability of Catastrophe

Recall what we did in Part I for quantifying risk of a hazard:

- Define catastrophic event in term of the outputs $y^M(\boldsymbol{x}) \in \mathcal{Y}_C$.
- Determine the 'catastrophic region' $\mathcal{X}_C$ in the input space:

$$\mathcal{X}_C = \{\boldsymbol{x} \in \mathcal{X} \ : \ y^M(\boldsymbol{x}) \in \mathcal{Y}_C\}$$

What we have to do now is to

- Assess a suitable distribution for the inputs (fitted with data)
- Use this distribution to compute the probability of at least a catastrophic event ($\boldsymbol{x} \in \mathcal{X}_C$) in t years.

SHV: assessing the risk of a catastrophic inundation

- Once the critical boundary Ψ is found, we need to determine the distribution of the stochastic variables (V, φ) to compute:

$$\begin{aligned} &\Pr(\text{at least one } (V,\varphi) \in \mathcal{X}_C \text{ in the next } t \text{ years }) \\ &= \Pr(\text{at least one } V \geq \Psi(\varphi) \text{ in next } t \text{ years}) \end{aligned}$$

  (Note that this computation is solely a probability and statistics computation: no more computer model runs are needed.)

- Volcanologists viewed the assumption of uniformity of φ as being reasonable for the larger flows of interest.

- Volcanologists *originally* viewed φ to be independent of V (for larger flows), so we need only the distribution for the $V's$ and τ's (volumes and times of PF events).

Modeling the Input Distributions

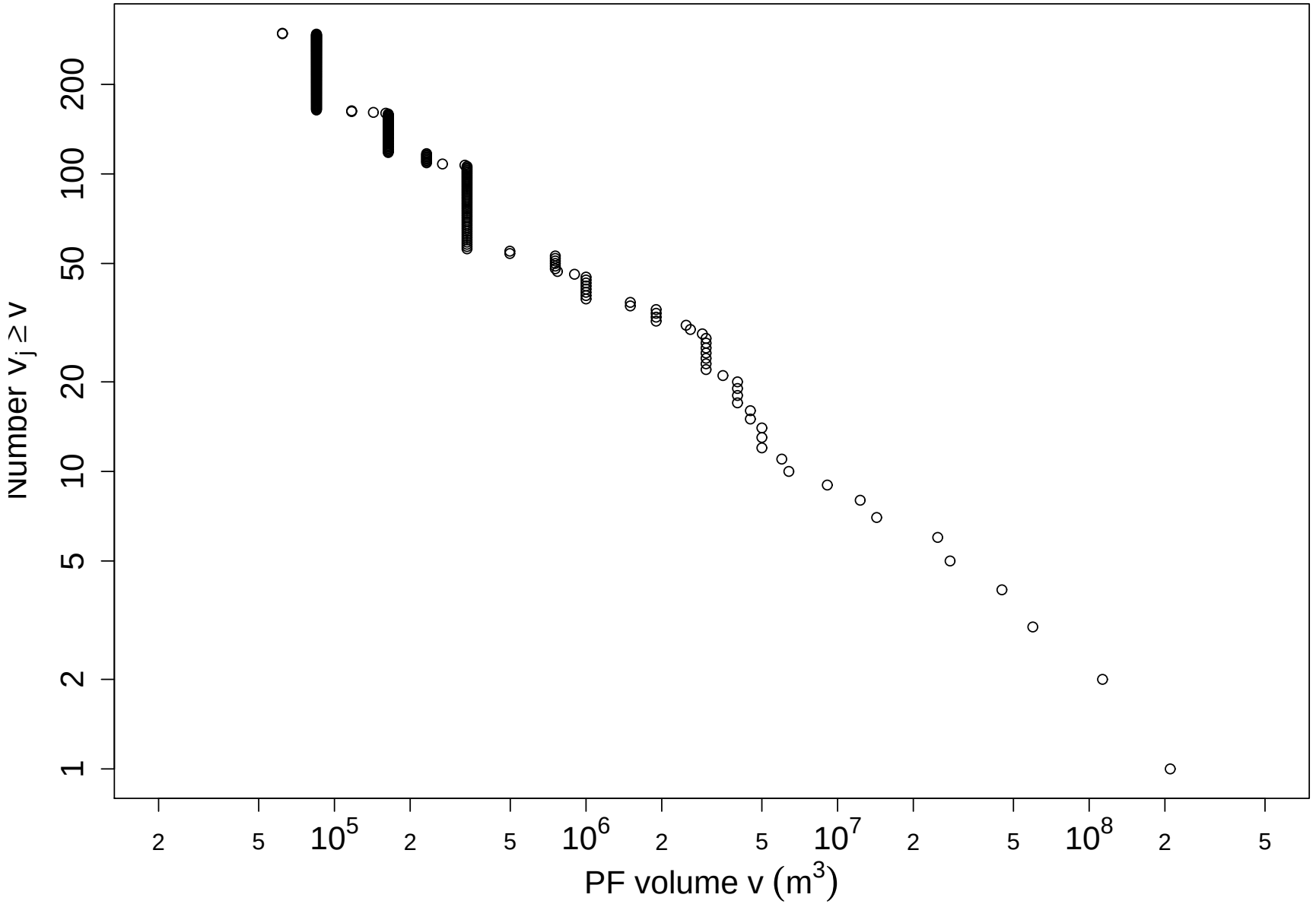
Initial assumptions:

- φ has a uniform distribution (for the larger flows of interest).
- φ is independent of V (for larger flows).
- PFs are stationary and independent over disjoint time intervals (so times of occurrence follow a stationary Poisson process) and

  Number of flows $V \geq \epsilon$ in $(0, t)$ is $Poisson(\lambda\, \epsilon^{-\alpha} t)$,

  with unknown rate parameter λ. (Here $\epsilon = 10^4$.)
- Arrival times and flow volumes are independent.

Then we only need to characterize the distribution of flow volumes


Number $v_j \geq v$
1
2
5
10
20
50
100
200
2
5
10^5
2
5
10^6
2
5
10^7
2
5
10^8
2
5
PF volume v (m^3)

This looks almost linear on a log-log scale

- Linear log-log plots of Magnitude vs. Frequency point to a Pareto distribution

$$\log \Pr(V \geq v \mid V \geq \epsilon) \approx -\alpha \log(v) + c, \qquad v \geq \epsilon$$

- in which case the probability density of V, for volumes greater than ϵ, would be Pareto $Pa(\alpha, \epsilon)$ with density

$$f(v \mid \alpha) = \frac{\alpha\, \epsilon^{\alpha}}{v^{\alpha+1}}\,, \qquad v \geq \epsilon\,,$$

  with unknown parameter α.

- Data suggest $\alpha \approx 0.64$ and hence enormous tails, infinite mean and variance, and significant chance of seeing in the future volumes larger than any we have seen in the past

Probability of a catastrophic event

It follows that, for any fixed $t > 0$, the number of catastrophic PF's (those with $V_i > \Psi(\varphi_i)$) in t years is Poisson with (conditional) mean

$$\begin{aligned} E(\#\text{ catastrophic PFs in } t \text{ yrs } \mid \alpha, \lambda) &= \int_0^{2\pi} \int_{\Psi(\varphi)}^{\infty} [\lambda \epsilon^{-\alpha} t] \frac{f(v \mid \alpha)}{2\pi} dv d\varphi \\ &= \frac{t\,\lambda}{2\pi} \int_0^{2\pi} \Psi(\varphi)^{-\alpha}\, d\varphi\,, \end{aligned}$$

$$\Pr(\text{At least one CPF in } t \text{ yrs} \mid \alpha, \lambda) = 1 - \exp\left[-\frac{t\,\lambda}{2\pi} \int_0^{2\pi} \Psi(\varphi)^{-\alpha}\, d\varphi\right],$$

$$\begin{aligned} P(t) &\equiv \Pr(\text{At least one CPF in } t \text{ yrs} \mid \text{data}) \\ &= 1 - \iint \exp\left[-\frac{t\,\lambda}{2\pi} \int_0^{2\pi} \Psi(\varphi)^{-\alpha}\, d\varphi\right] \pi(\alpha, \lambda \mid \text{data})\; d\alpha\; d\lambda\,, \end{aligned}$$

where $\pi(\alpha, \lambda \mid \text{data})$ is the posterior distribution of (α, λ).

Posterior distribution of (α, λ)

For a given ϵ and period $[0, t]$, the sufficient statistics are $J = \#$ of PF's on $[0, t]$, and $S = \sum \log(V_j)$, the log-product of their volumes. The likelihood function is: $L(\alpha, \lambda) \propto (\lambda\,\alpha)^J \exp[-\lambda t\,\epsilon^{-\alpha} - \alpha\,S]$.

Objective Priors:

- Jeffreys prior is $\pi_J(\alpha, \lambda) \propto |I(\alpha, \lambda)|^{1/2} \propto \alpha^{-1}\epsilon^{-\alpha}$
- Reference priors
  - α of interest gives $\pi_{R1}(\alpha, \lambda) \propto \lambda^{-1/2}\alpha^{-1}\epsilon^{-\alpha/2}$
  - λ of interest gives $\pi_{R2}(\alpha, \lambda) \propto \lambda^{-1/2}[\alpha^{-2} + (\log\epsilon)^2]^{1/2}\epsilon^{-\alpha/2}$ which also is Jeffrey's independent prior

Posterior: $\pi(\alpha, \lambda \mid \text{data}) \propto L(\alpha, \lambda)\,\pi(\alpha, \lambda)$.

Computing the probabilities of catastrophe

To compute Pr(at least one catastrophic event in t years | data) for a range of t, an importance sampling estimate is

$$P(t) \cong 1 - \frac{\sum_i \exp\left[-\frac{t\,\lambda_i\,\widehat{\Psi}(\alpha_i)}{2\pi}\right]\,\frac{\pi^*(\alpha_i,\lambda_i)}{f_I(\alpha_i,\lambda_i)}}{\sum_i \frac{\pi^*(\alpha_i,\lambda_i)}{f_I(\alpha_i,\lambda_i)}},$$

- where $\widehat{\Psi}(\alpha)$ is an MC estimate of $\int_0^{2\pi} \Psi(\varphi)^{-\alpha}\,d\varphi$ based on draws $\varphi_i \sim Un(0, 2\pi)$;
- $\pi^*(\alpha, \lambda)$ is the un-normalized posterior;
- (α_i, λ_i) are drawn from the importance sampling density $f_I(\alpha, \lambda) = t_2(\alpha, \lambda \mid \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}, 3)$, with d.f. 3, mean $\widehat{\boldsymbol{\mu}}^t = (\hat{\alpha}, \hat{\lambda})$, and scale $\widehat{\boldsymbol{\Sigma}} =$ inverse of observed Fisher information matrix.

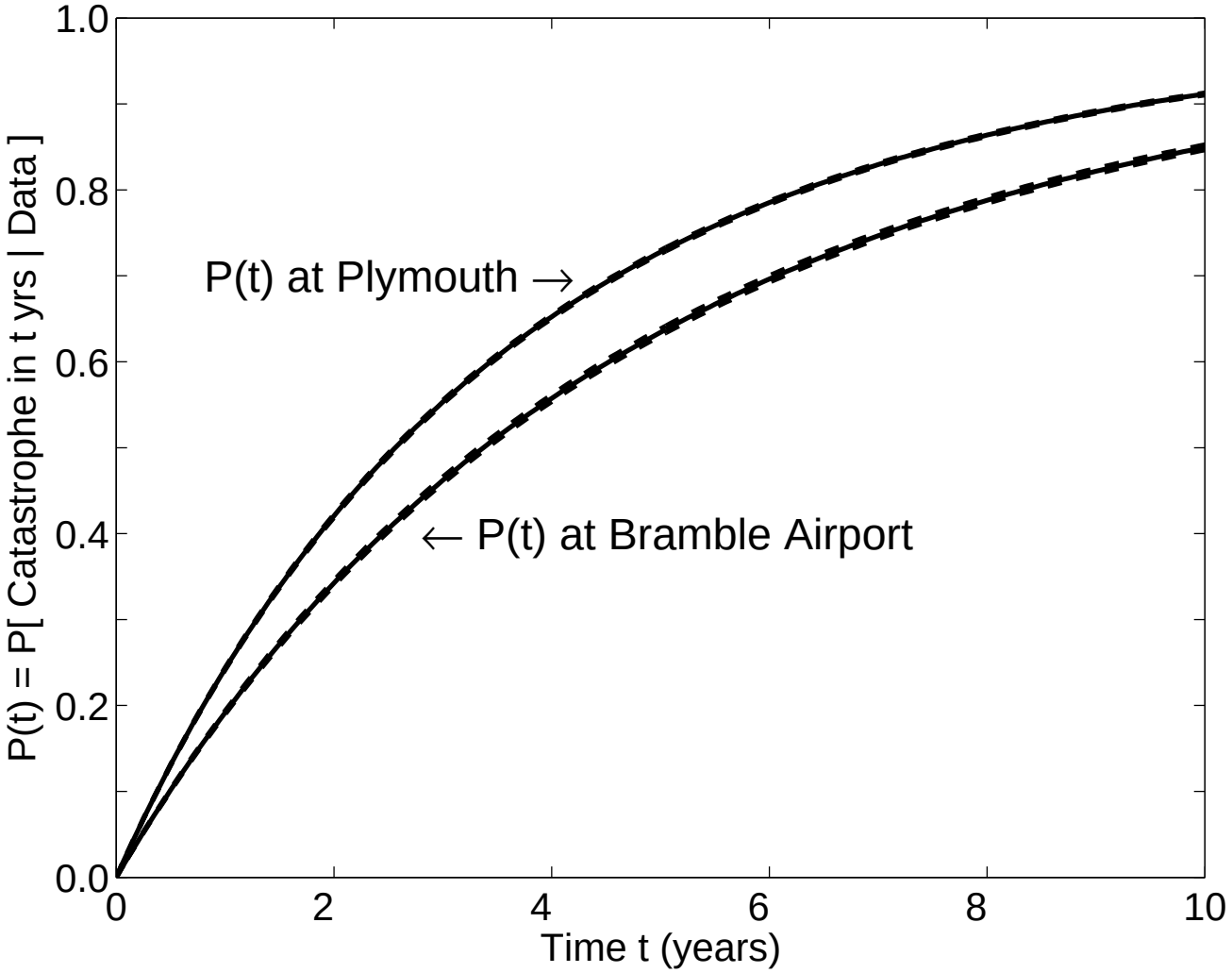


Figure 4: $P(t)$ at Plymouth (higher curves) and Airport (lower curves). Solid (dashed) $\rightsquigarrow$ computed with the upper (lower) 75% confidence bands. Different reference priors lead to overlapping curves.

Back to the Assumptions

- φ has a uniform distribution (for the larger flows of interest).
- φ is independent of V (for larger flows).
- PFs are stationary and independent over disjoint time intervals so

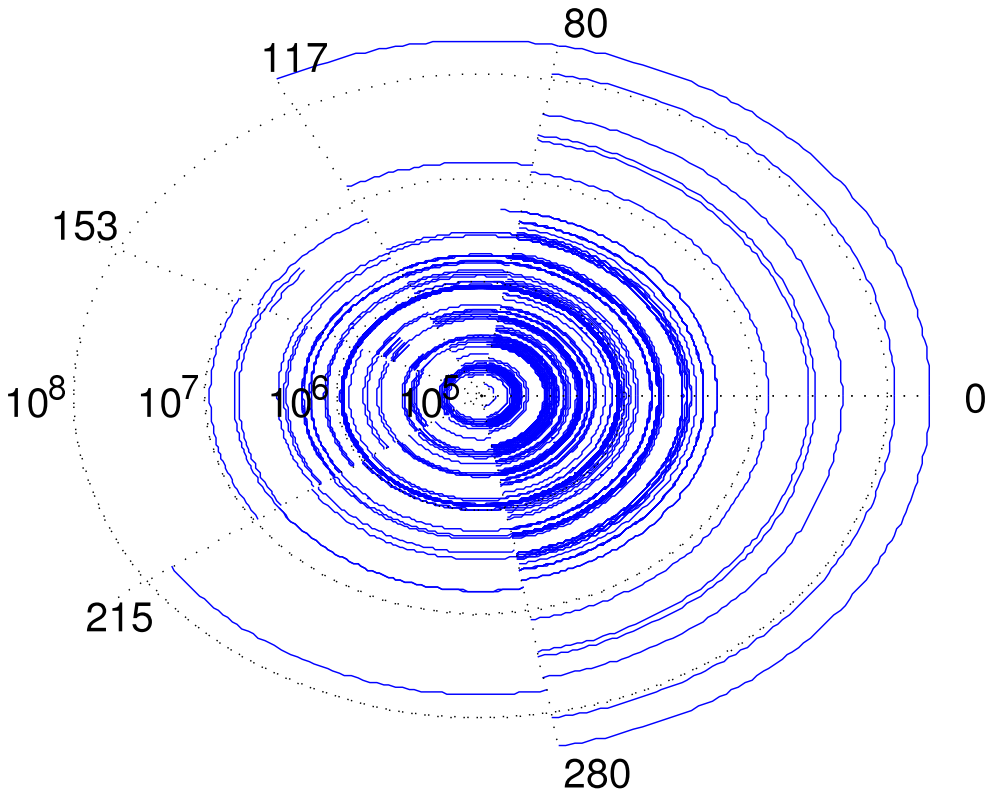
  Number of flows $V \geq \epsilon$ in $(0, t)$ is $Poisson(\lambda\, \epsilon^{-\alpha} t)$,

  with unknown rate parameter λ. (Here $\epsilon = 10^4$.)
- Arrival times and flow volumes are independent.
- Flow volumes are Pareto.

None are correct.

PF Initiation Angles

The data on angles is quite vague— we only know which of 7 or 8 *valleys* were reached by a given PF, from which we can infer a *sector* but not a specific angle φ:



Data seems to indicate non-uniformity in φ; also, the assumption of indepence of φ on volume V seems suspect

We need a *joint* density function for V and φ, and are using

$$V, \varphi \sim \alpha\, \varepsilon^{\alpha}\, V^{-\alpha-1}\, \pi_{\kappa}(\varphi) \qquad V > \varepsilon$$

where $\pi_{\kappa}(\varphi)$ is the von Mises distribution with pdf

$$f(\varphi \mid \kappa, \mu) = \frac{e^{\kappa \cos(\varphi - \mu)}}{2\pi I_0(\kappa)},$$

centered at $\varphi \approx \mu$ close to zero (East) with concentration κ that might depend on V if the data support that.

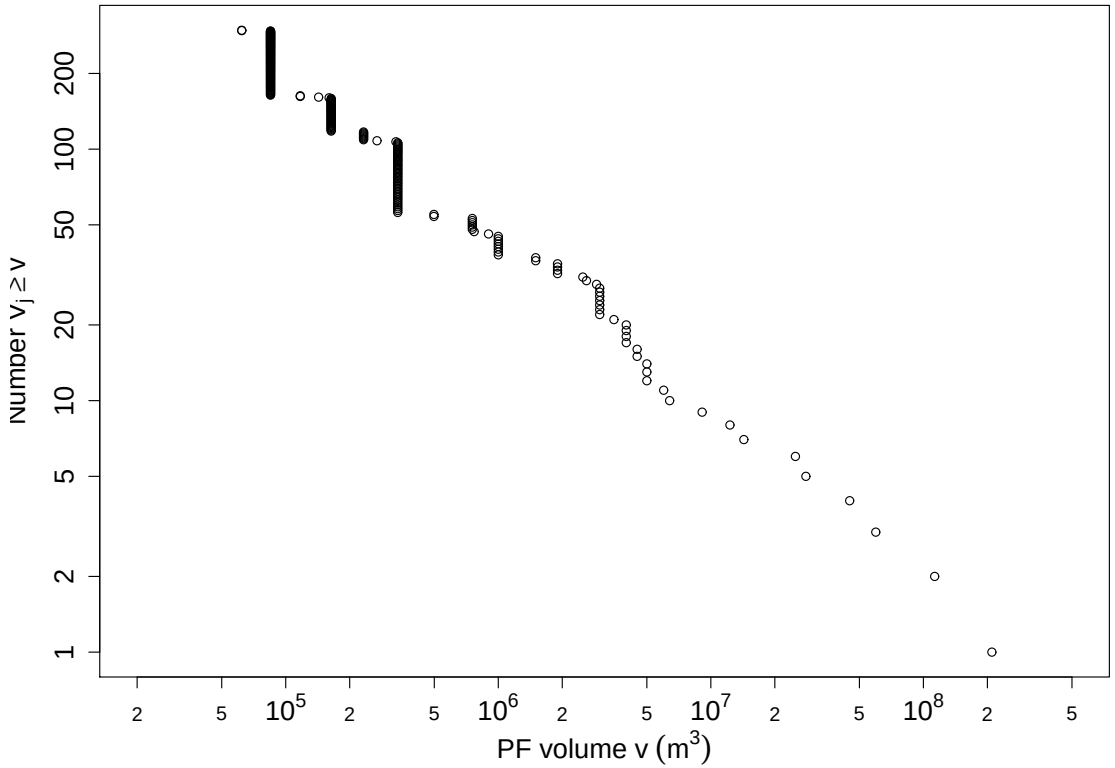
Stationarity and Independence of PFs

Wrong, but not necessarily a problem for long time periods.

For shorter time periods, nonstationary models are being developed (Jianyu Wang) where the Poisson rate parameter λ_t varies over time via a change point model, and can be zero over periods of time.

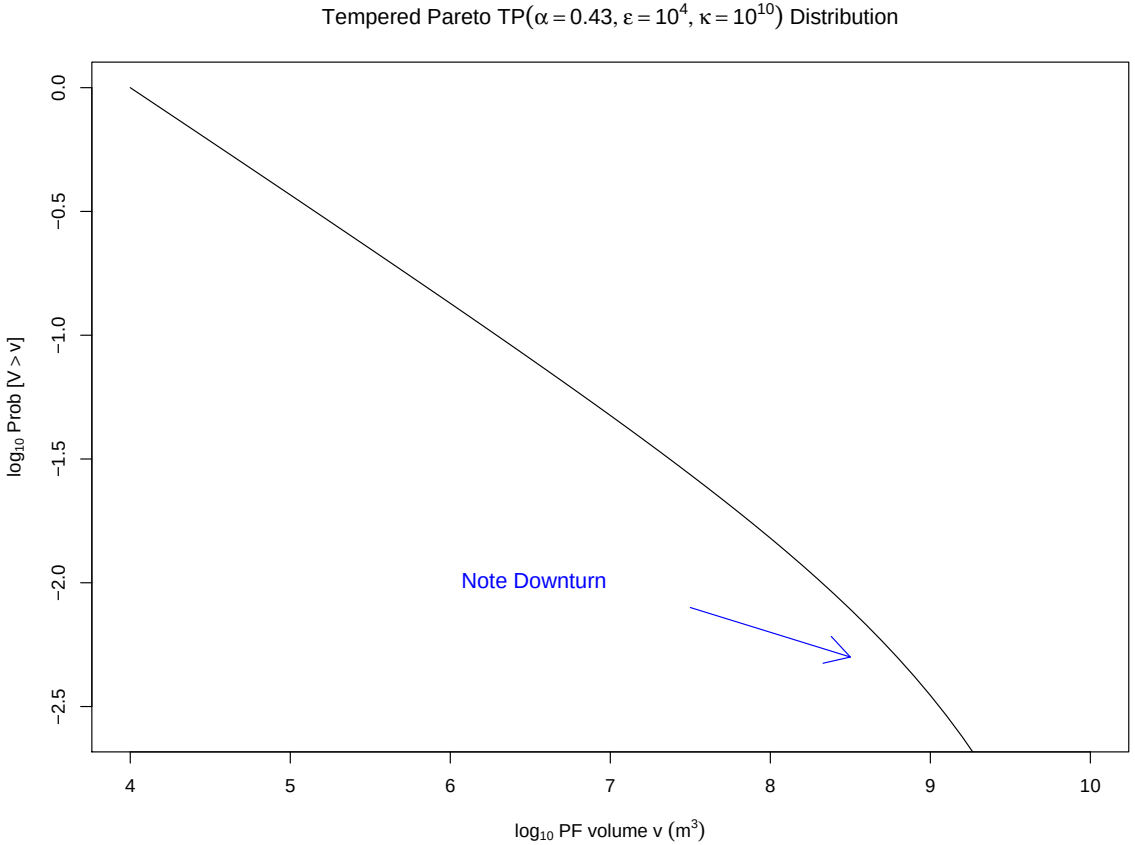
Assumption of Pareto Flows

Flows can't be arbitrarily large, but for SHV the Pareto tails imply 0.1% chances of a PF exceeding 10^{12} in a century (earth volume $\approx 10^{19}$). Clearly need *tempering* (or truncation) of the tail. Data also seems to suggest some tempering



A Tempered Pareto model is

$$P[V > v] = (v/\epsilon)^{-\alpha}\, e^{-\beta(v-\epsilon)}, \qquad v > \epsilon$$



Discussion

We have argued that:

- Risk assessment of catastrophic events (in the absence of lots of extreme data) requires
  - Mathematical computer modeling to extrapolate beyond the range of the data.
  - Statistical modeling of available (possibly not extreme) data to determine input distributions and perform calibration, while attempting to account for all uncertainties.
  - Statistical development of emulators of the computer model to determine critical event contours.
- Major sources of uncertainty can be combined and incorporated with a Bayesian analysis.

To be highlighted:

- we *only* fit an emulator around $\mathcal{X}_C$. This:
  - Improves the fitting (emulators work best when localized)
  - Crucially simplifies the otherwise "large data sets" GASP computations
- Computing the probability of interest (a catastrophic event in t years) only requires
  - The input distribution, which is derived independently of the computer model
  - The (distribution of) the critical region, which does not need the input distribution and is determined based on the computer model runs

- Therefore
  - Several improvements and changes in the input distribution (tempering the tails, taking into account dependence on volume and angles,...) can be entertained WITHOUT requiring new runs of the computer model.
  - This is not the case in the 'brute force' MC approach to risk analysis; any change in the distribution for the inputs, or simply getting more data, requires new model runs.

Numerous uncertainties are present in each and every step of this Risk Assessment; they are dealt with Bayesian analysis (mostly)

THANKS!!