Mastering the Data Explosion in the Earth and Environmental Sciences

Australian Academy of Science Elizabeth and Frederick White Conference

Dealing with unknown discontinuities in data and models

Kerry Gallagher

John Stephenson

Chris Holmes

Imperial College London

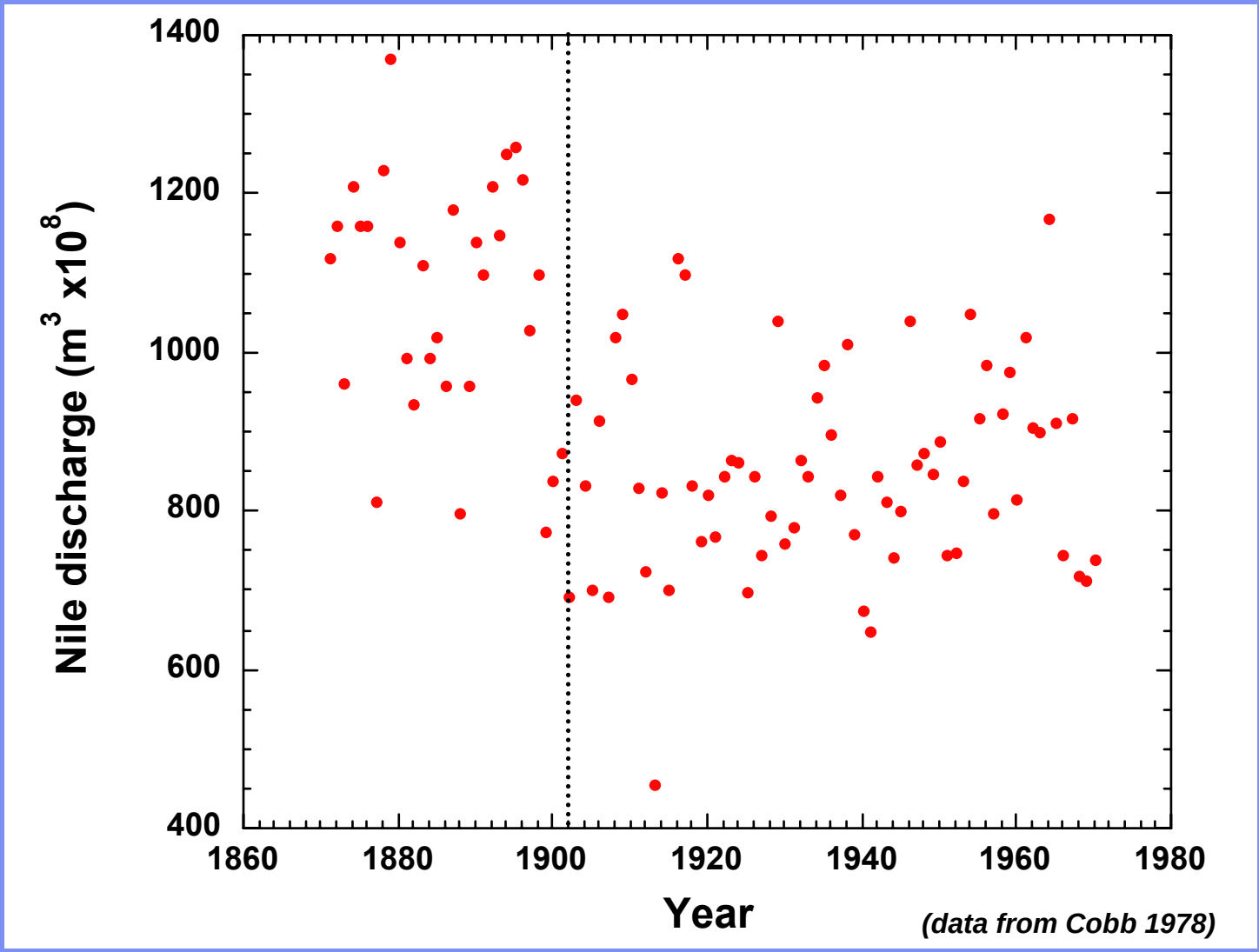
Discontinuities occur in both data and processes in the Earth and Environmental Sciences

Spatial : faults, topography, lithology, phase, composition,…

Temporal : climate, seismicity, tectonics,…

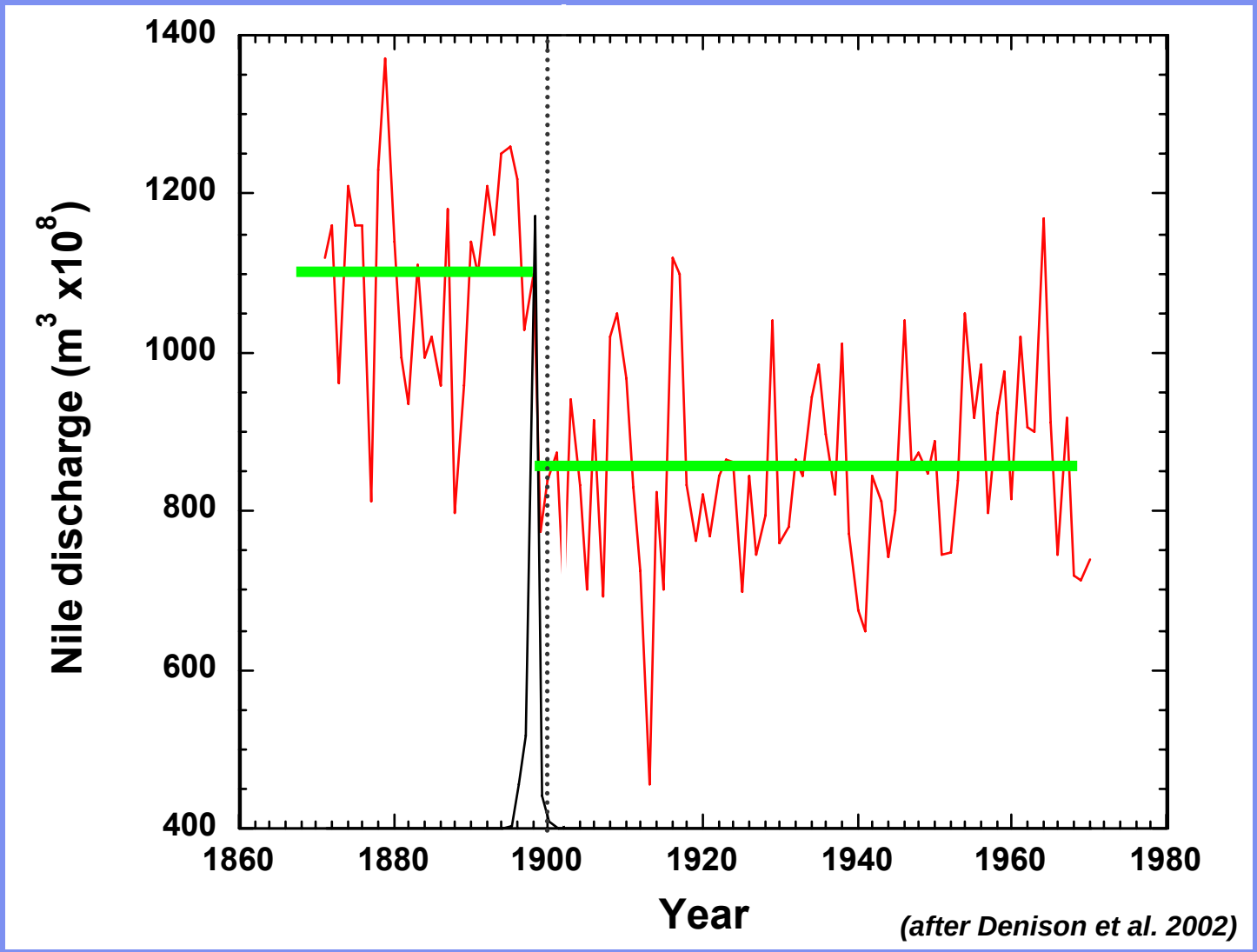
What is the appropriate question ?

What was the signficance of the opening of the Aswan Dam ?



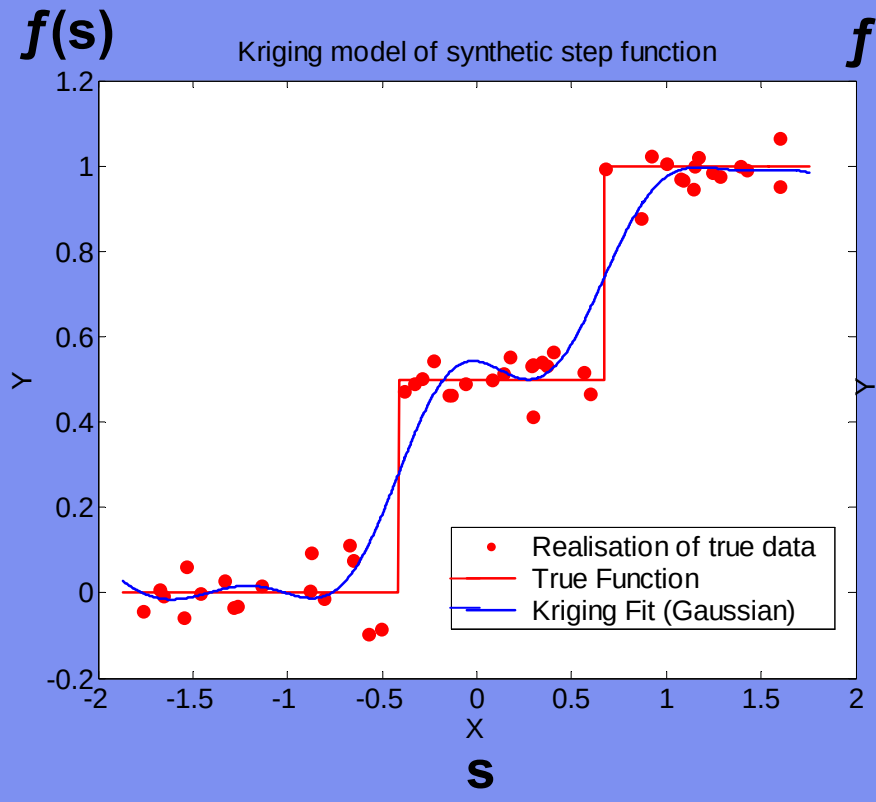
When was the change ?

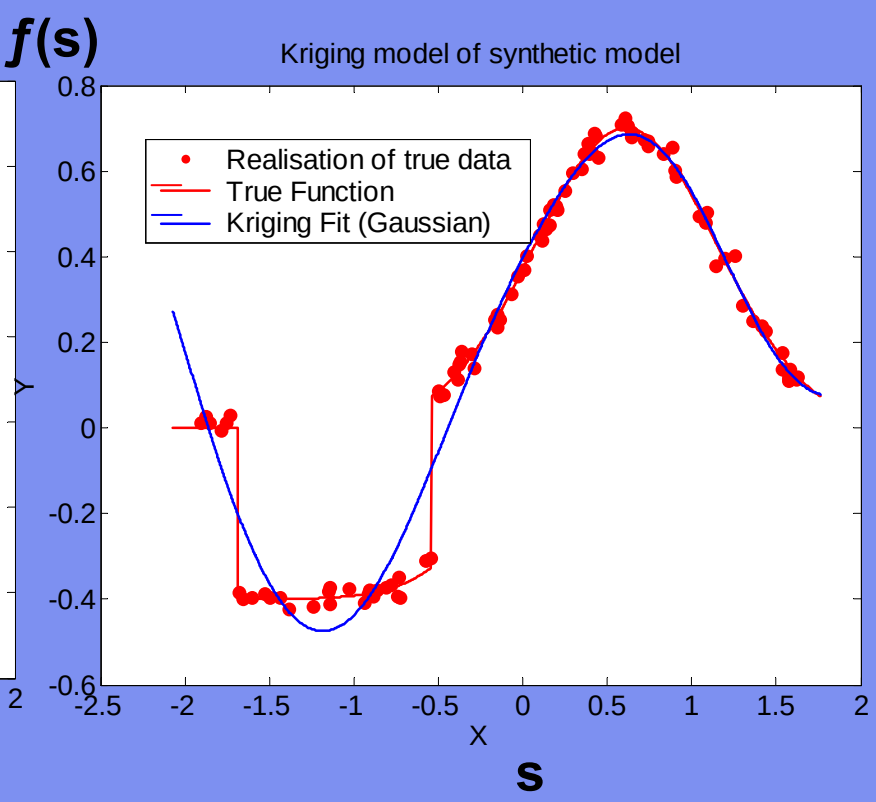
$$f(t) = \mu_1 I(t \le t_c) + \mu_2 I(t > t_c)$$



Data interpolation and prediction with discontinuities

Standard methods may be too smooth

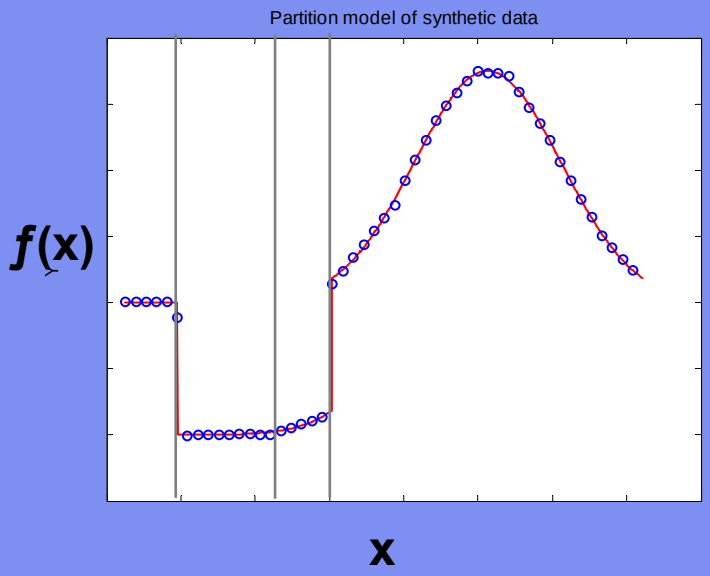




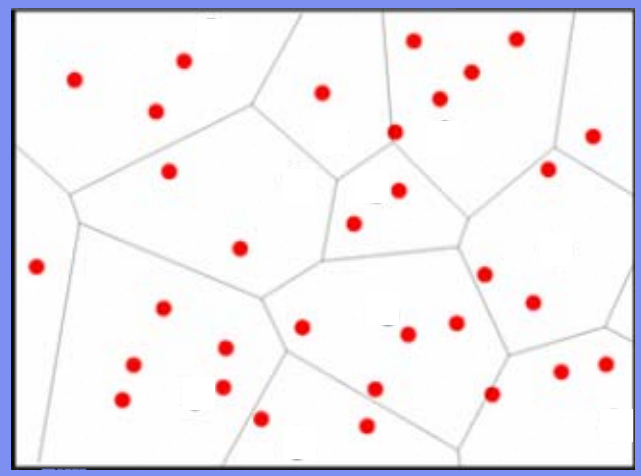
Need a method that can deal with an unknown number of discontinuities in unknown locations

Partition Modelling

1D

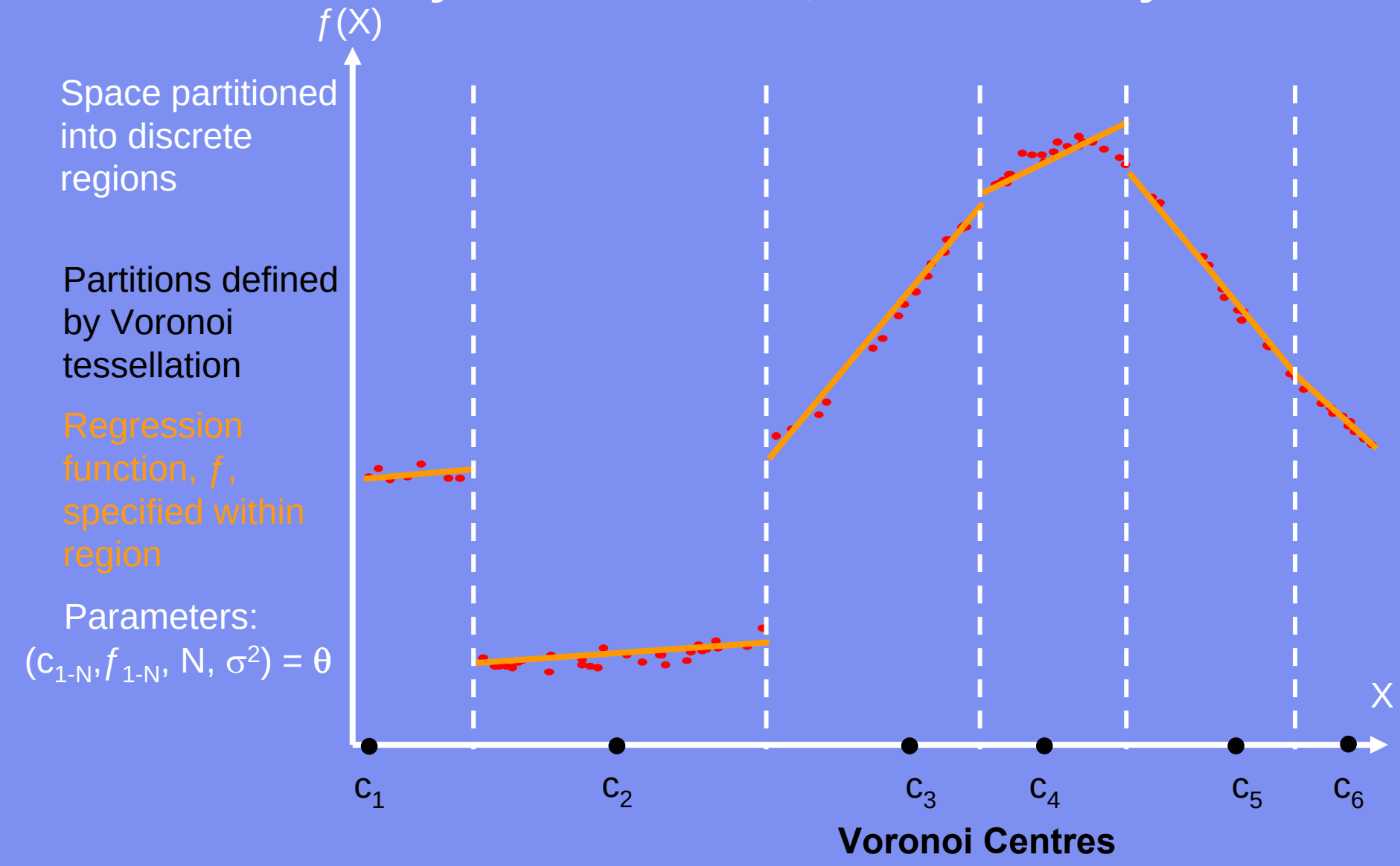


2D



Formulating a Partition Model

How many discontinuities, where are they ?



Generating Partition Models

Prediction

$$p(y \mid D) = \int_{\Theta} p(y \mid \theta, D)\, p(\theta \mid D)\, d\theta$$

Posterior distribution

Monte Carlo integration

y = value to be predicted
D = observed data
θ = model parameters

$$p(y \mid D) \approx \frac{1}{N} \sum_{i=1}^{N} p(y \mid \theta_i, D)\, p(\theta_i \mid D)$$

Bayes' Theorem

$$p(\theta \mid D) \propto p(D \mid \theta)\, p(\theta)$$

Posterior Likelihood Prior

Use Markov chain Monte Carlo (MCMC) to sample the posterior distribution, *p(θ|D)*

Sampling with (transdimensional) MCMC

Initialise θ

Iterate

- **Propose new** θ'
- **Calculate likelihood with new** θ'
- **Accept new** θ' **or retain current** θ

Acceptance criterion

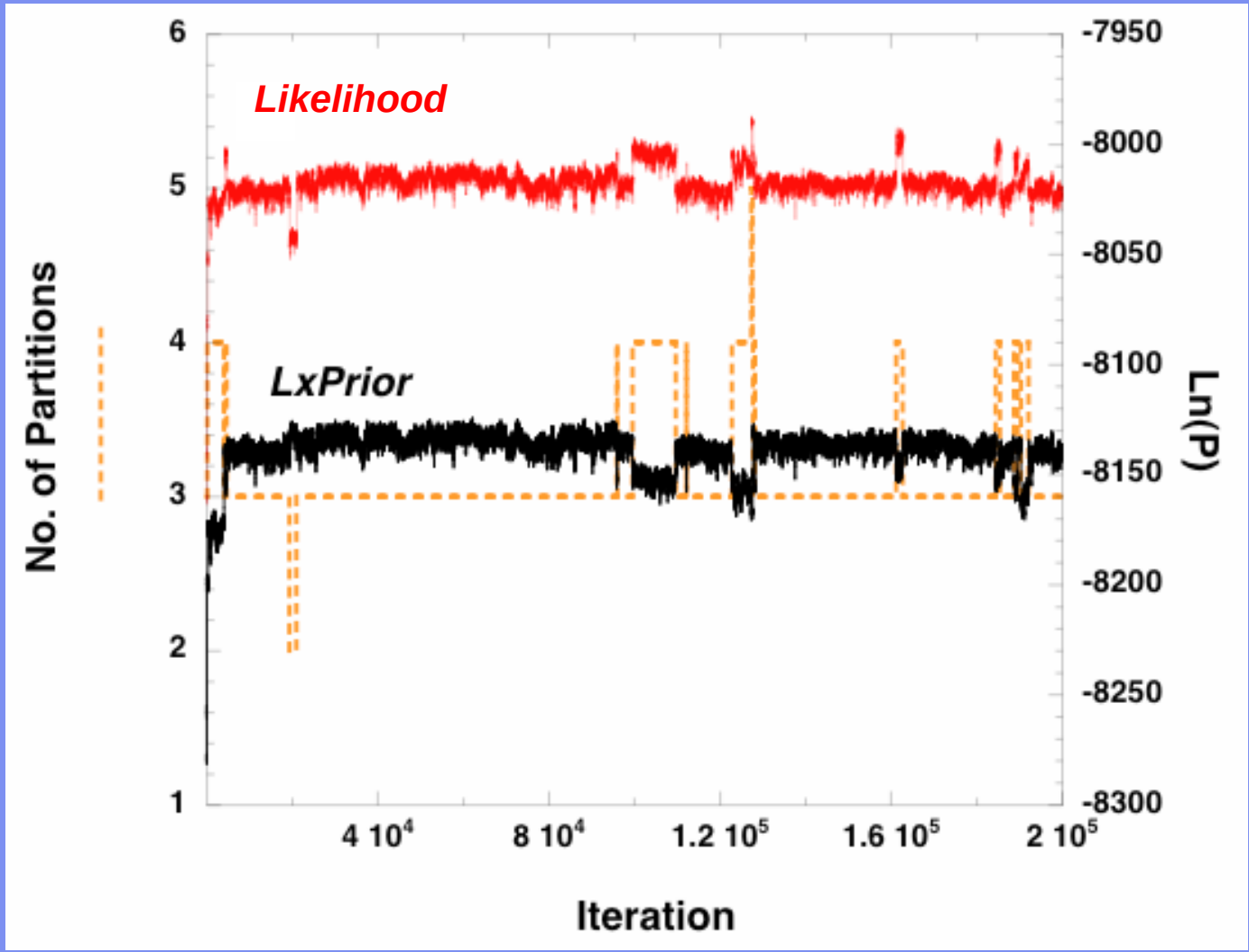
$$\alpha(\theta,\theta') = \min\left\{1, \frac{p(\theta')p(D|\theta')\,p(\theta/\theta')}{p(\theta)p(D|\theta)\,p(\theta'/\theta)}\,R\,|J|\right\}$$

Prior · Likelihood · Model · Proposal

Jump proposal · Jacobian

Distribution of accepted models $\theta \sim p(\theta|D)$

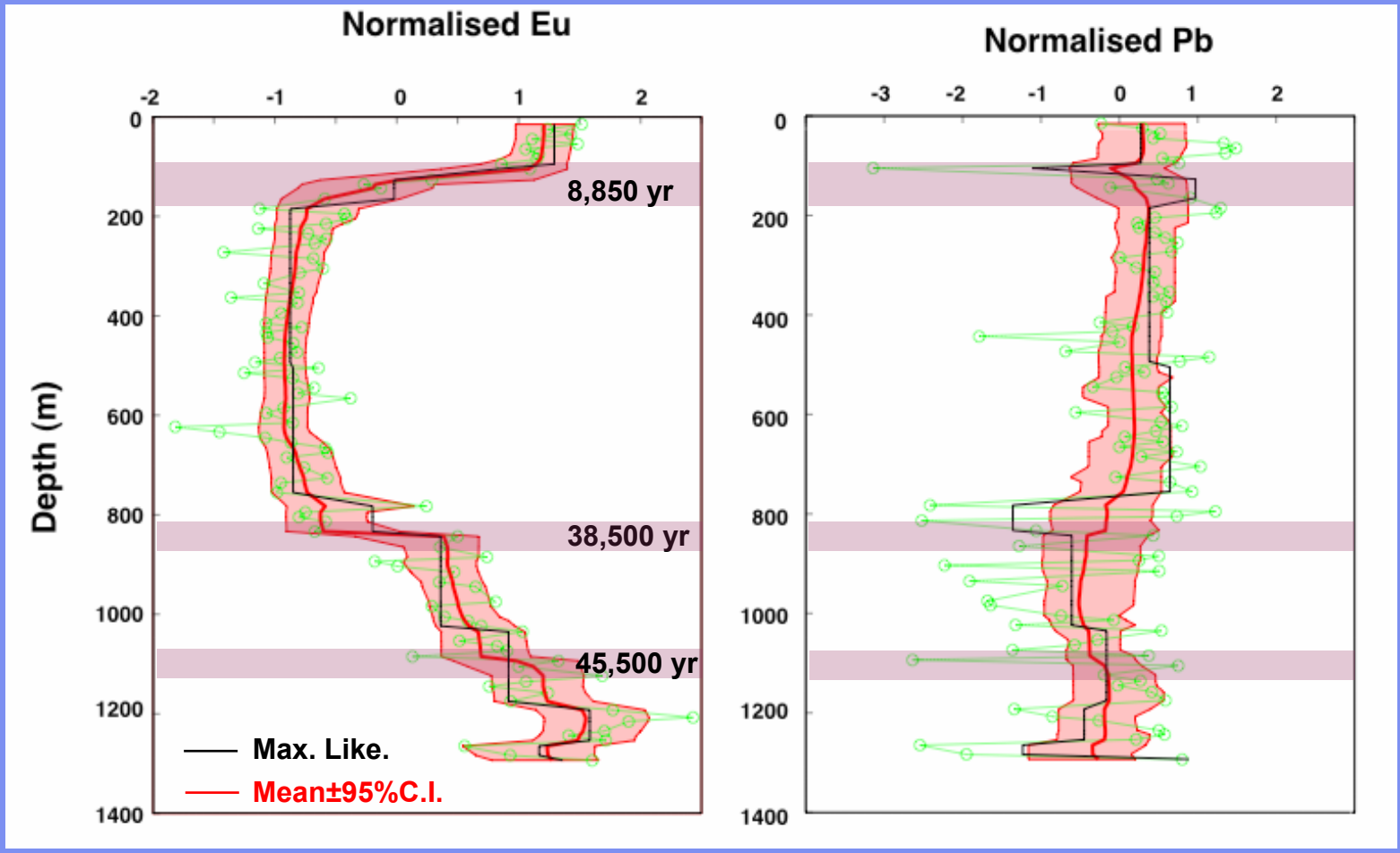
Sampling Partition Models
natural parsimony



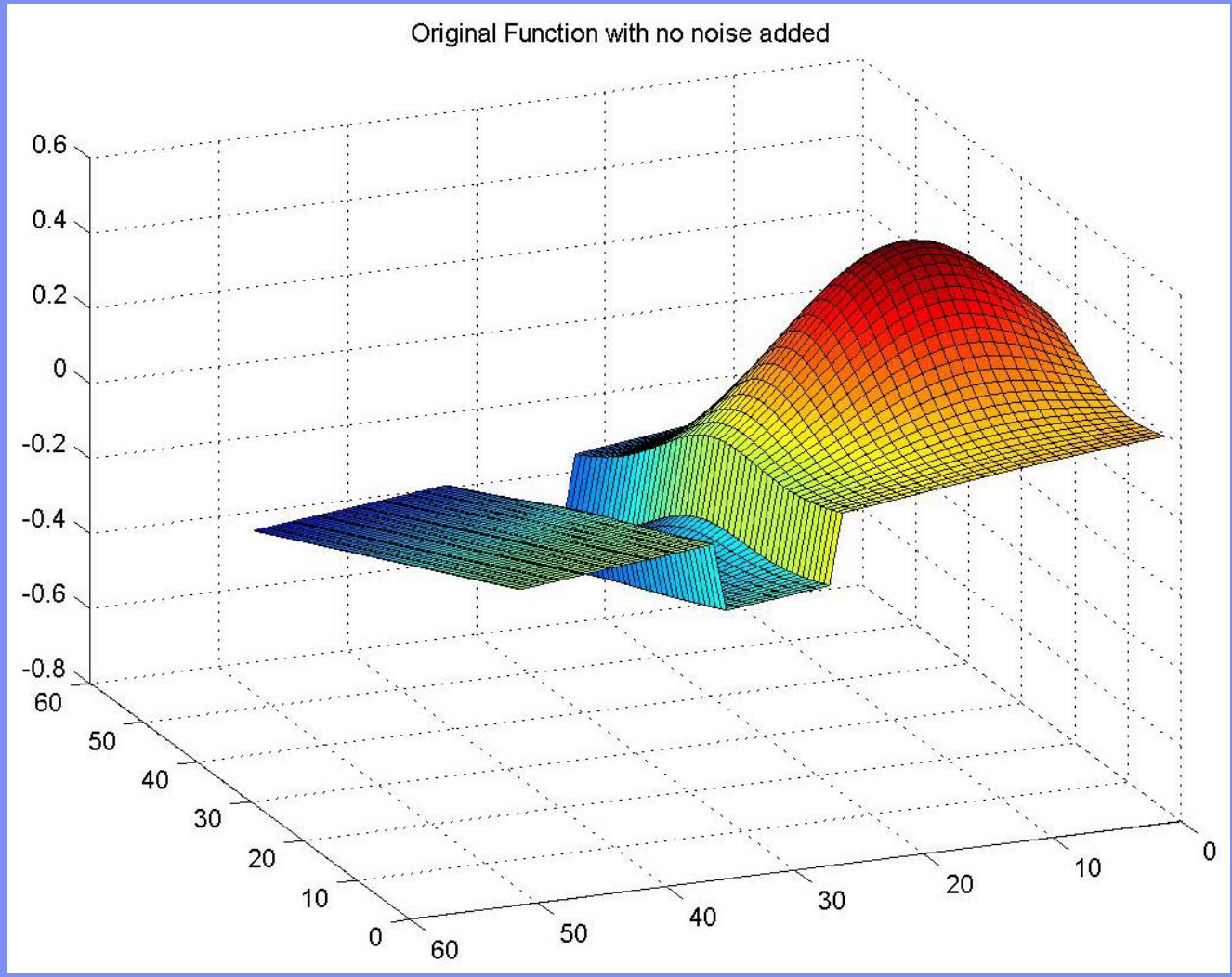
1D partition models for data interpolation

Atmospheric dust input to peat bogs

Looking for common signature in multiple systems



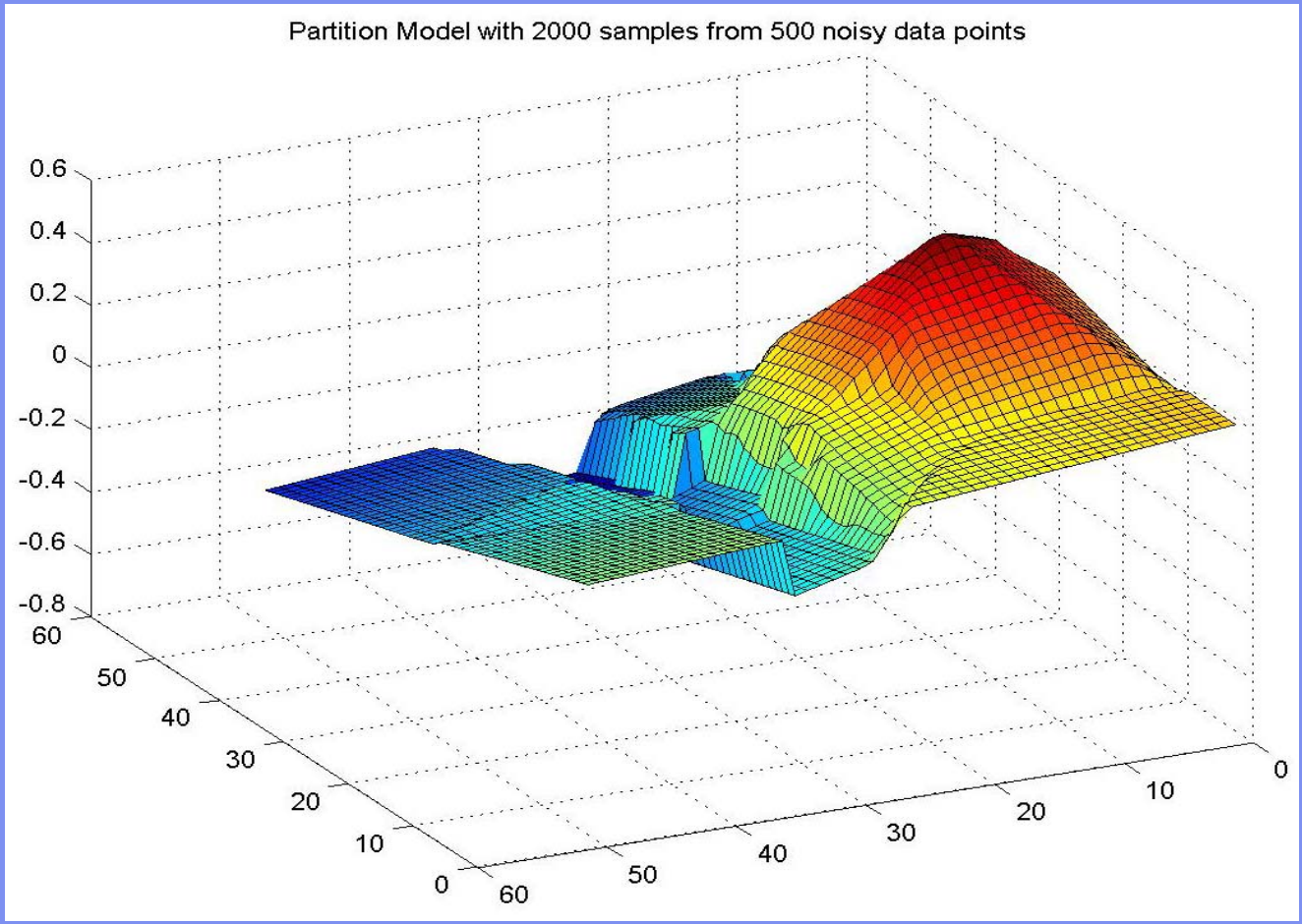
Partition Models – 2D example function



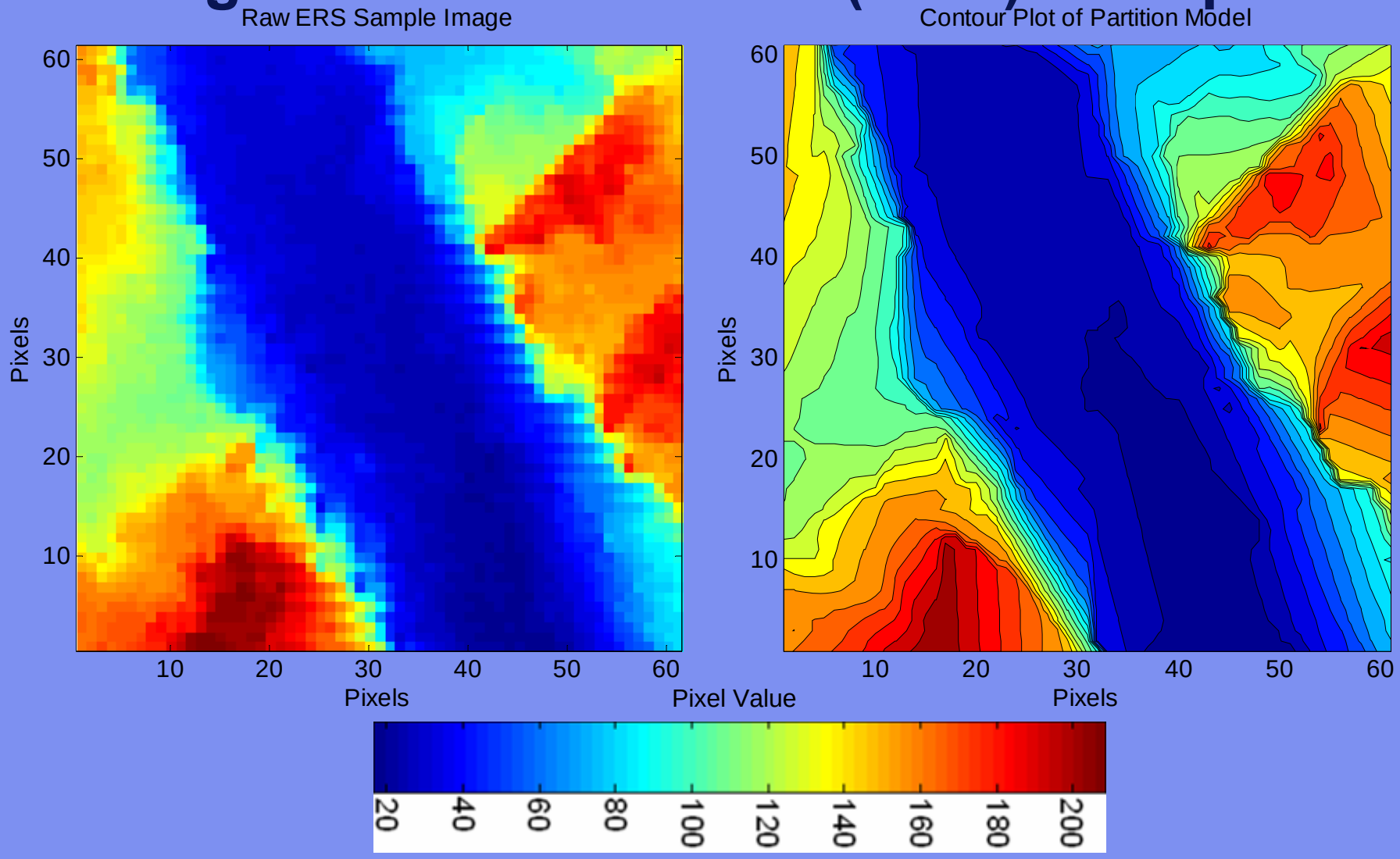
Partition Sampling – 2D single realisation

Multiple realisations …

ensemble average (smooth, but maintain discontinuities)



Partition Model
Digital Elevation Model (DEM) example

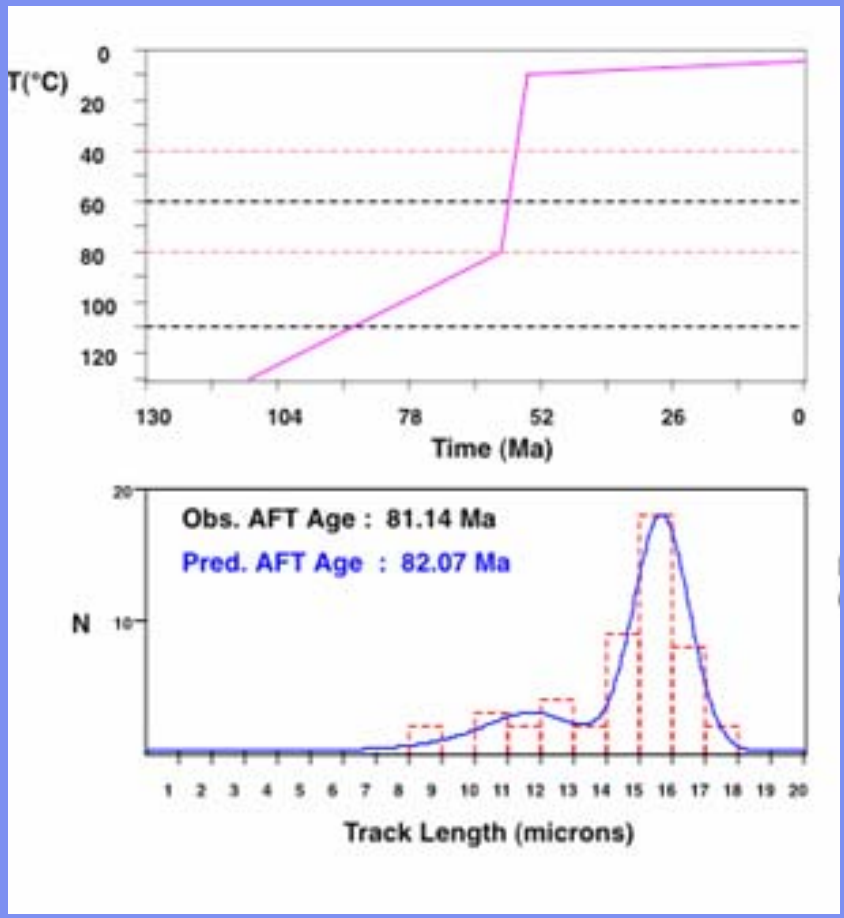


Partition Models
Application to spatially variable physical processes and parameters

Example from thermochronology

Thermochronology : data are sensitived to temperature history experience by host rock

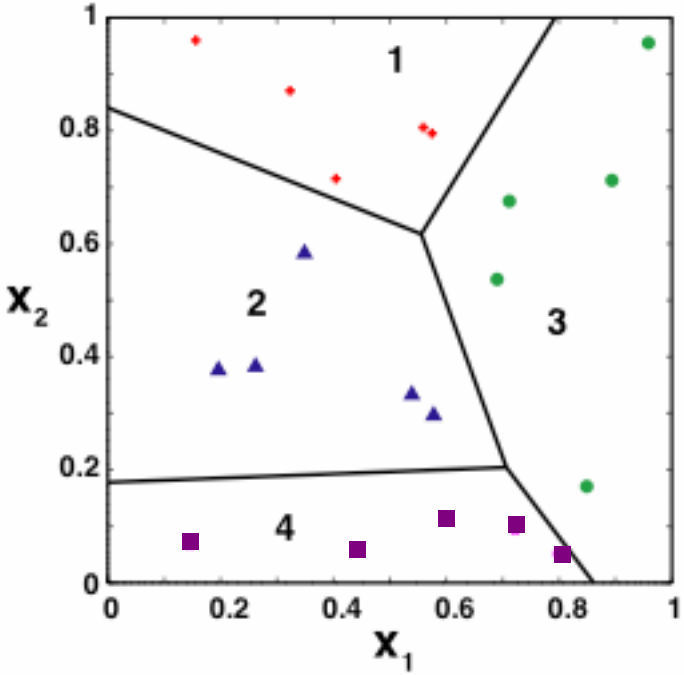
e.g. apatite fission track analysis

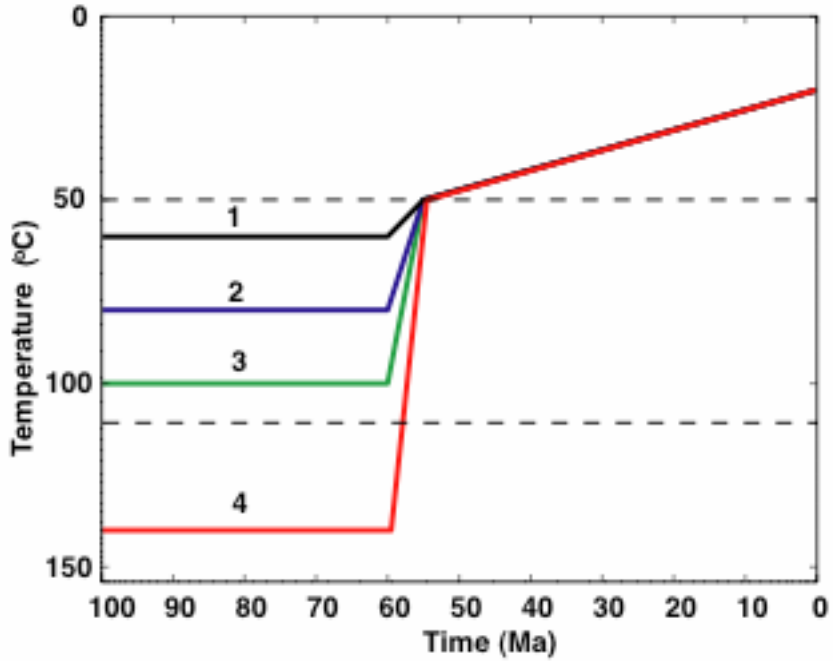


$$p(D|\theta) = f(T(t), \phi)$$

Likelihood is a non-linear function of unknown parameters at each location within each partition

Model partition distribution and thermal histories





The problem is to find

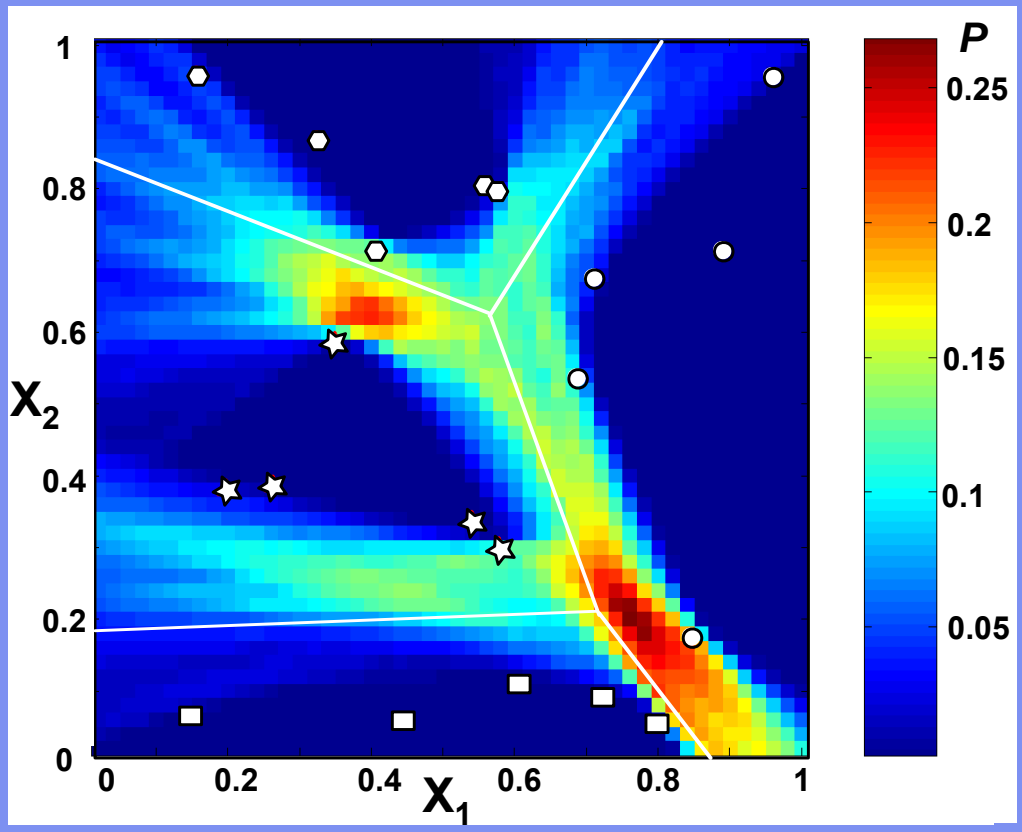
(a) how to partition the samples in 2D

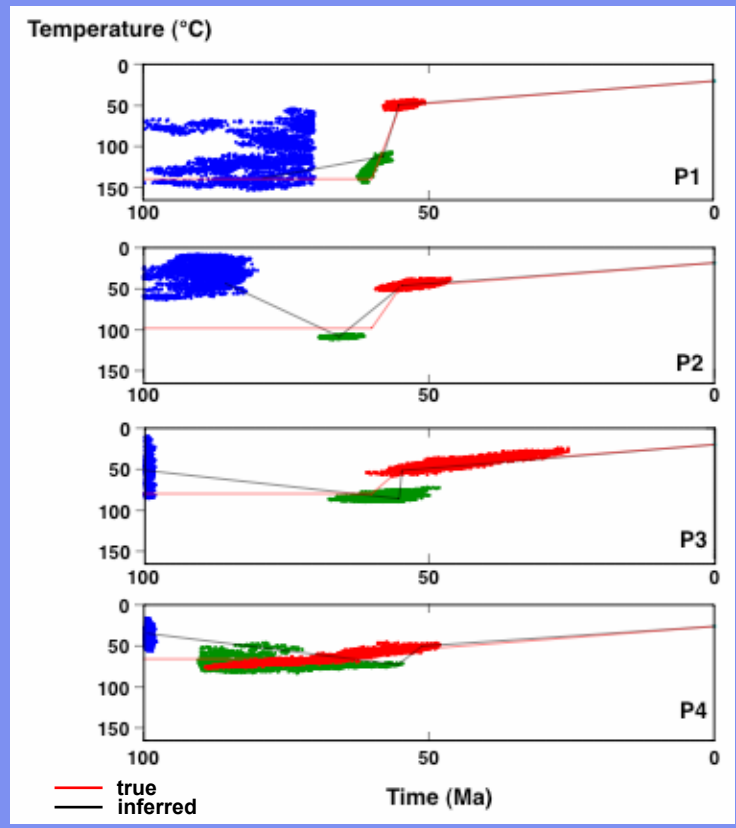
(i) number of partitions

(ii) location of the partitions

(b) the distribution of thermal histories in each partition

Inferred partition distribution and thermal histories



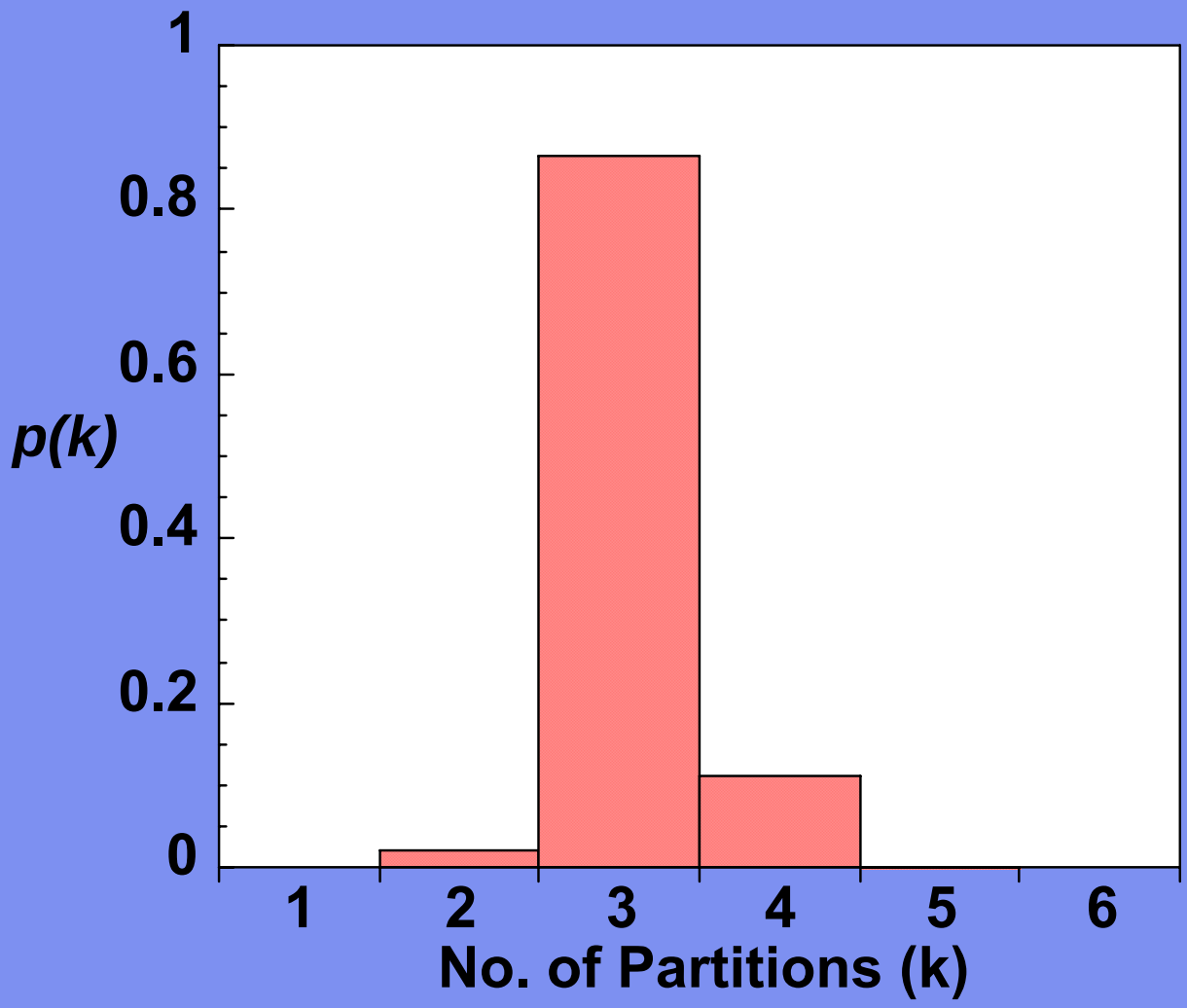


(Stephenson, Gallagher and Holmes 2006)

Summary

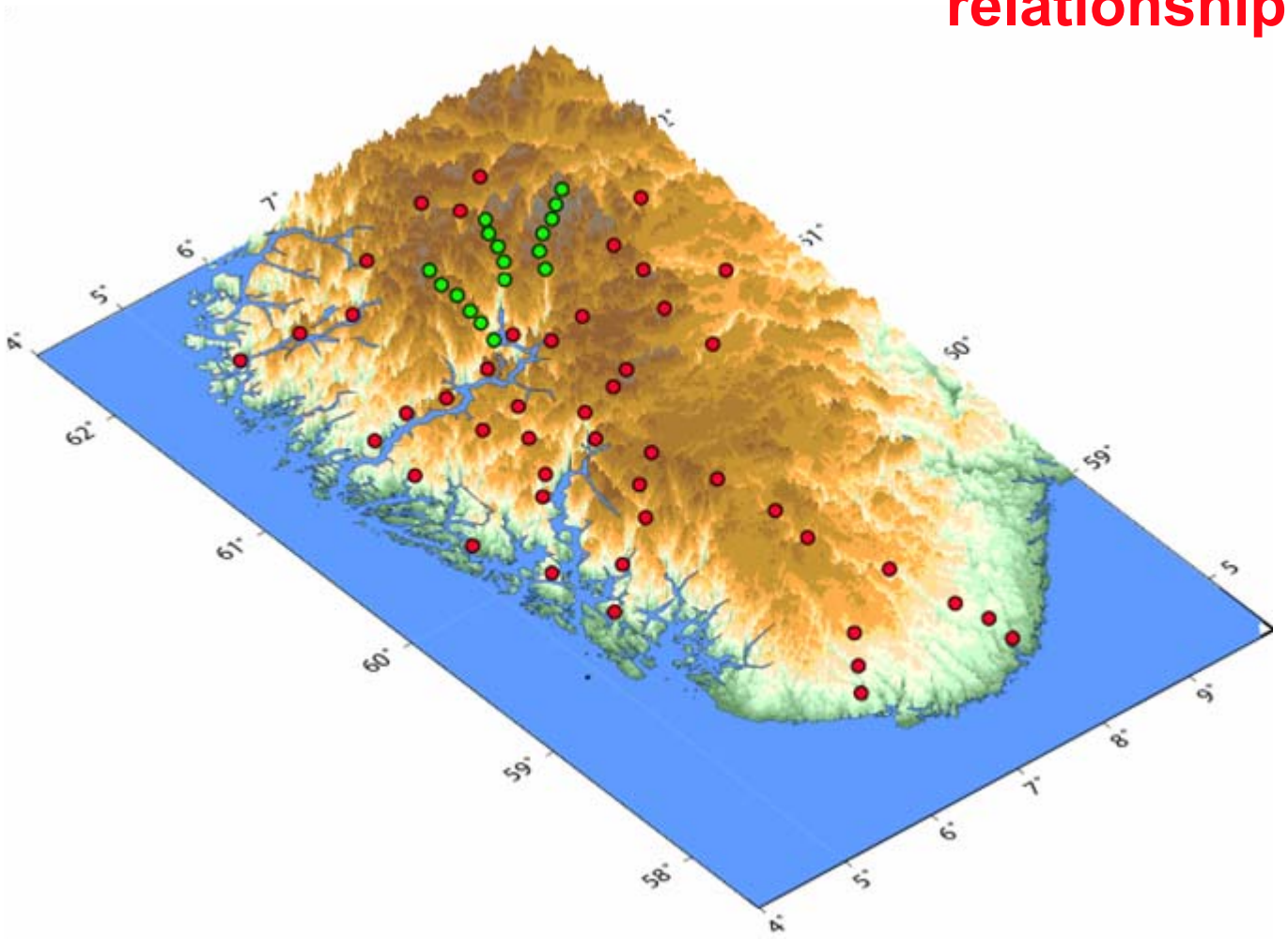
- Partition models allow for unknown number of discontinuities with unknown geometry in variable dimensions
- Bayesian approach deals with the problem in terms of probabilities…intuitive for model choice
- Obtain probability distributions (partitions, model parameters, posterior predictions)
- Bayesian approach is naturally parsimonious
- Potential for self-adaptive/self regularising model parameterisation

Sampling Partition Models distribution on number of partitions



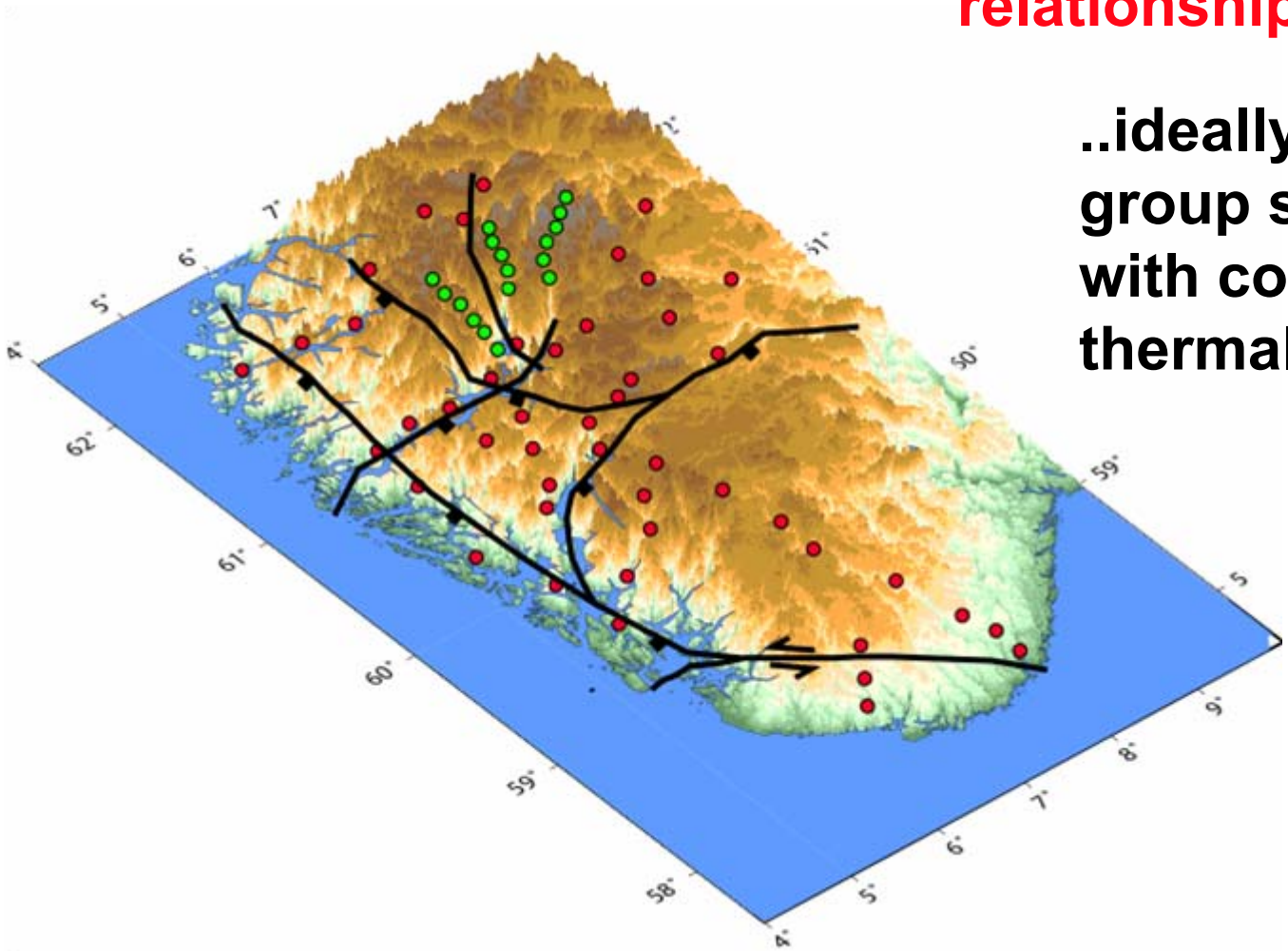
Traditionally, each sample is modelled independently..

ignores spatial relationships….



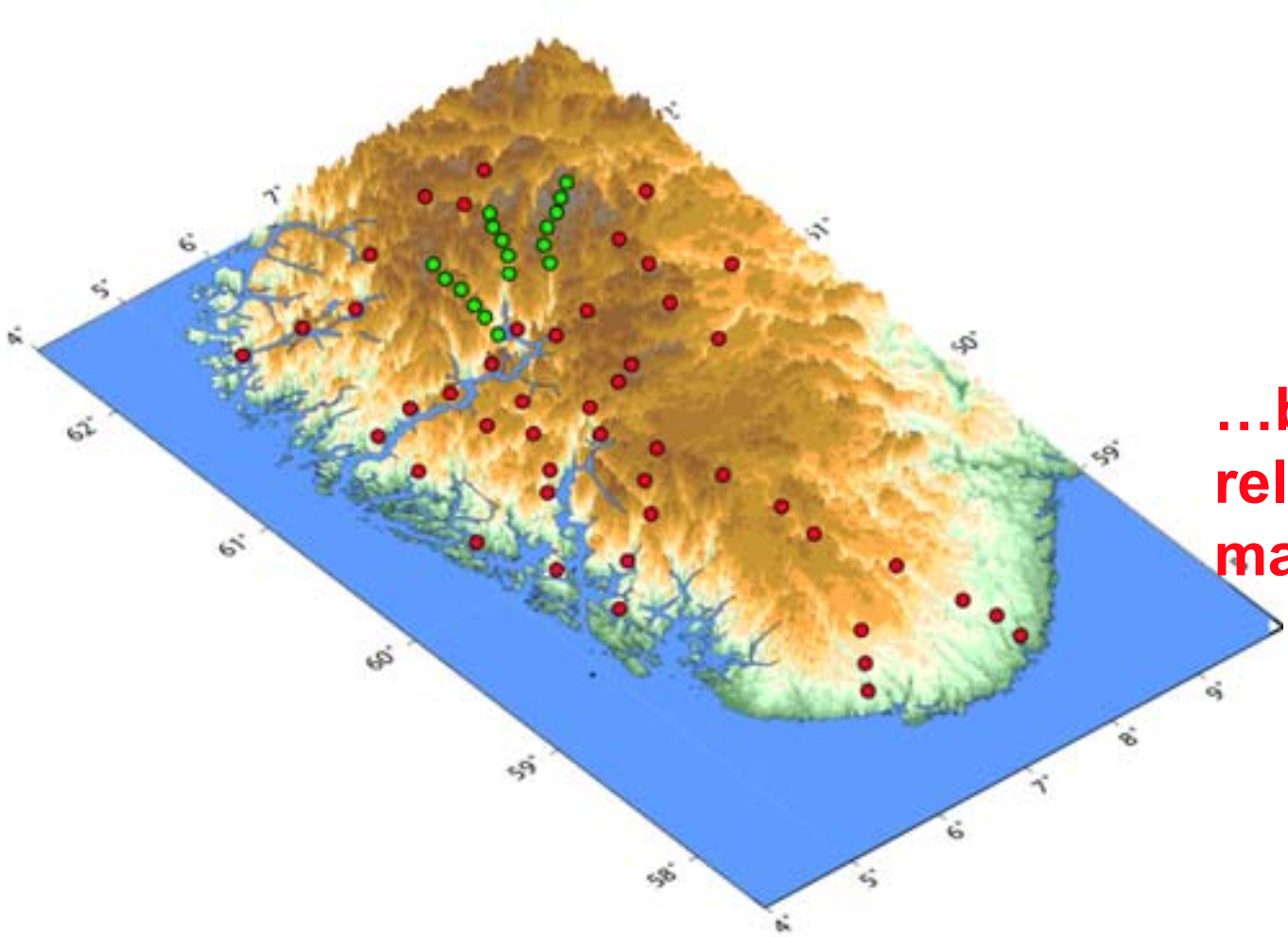
Traditionally, each sample is modelled independently..

ignores spatial relationships….



..ideally want to group samples with common thermal history

Traditionally, each sample is modelled independently..



…but the spatial relationships may be unknown…