Sampling, modelling and uncertainty, statistical approaches to some environmental questions.

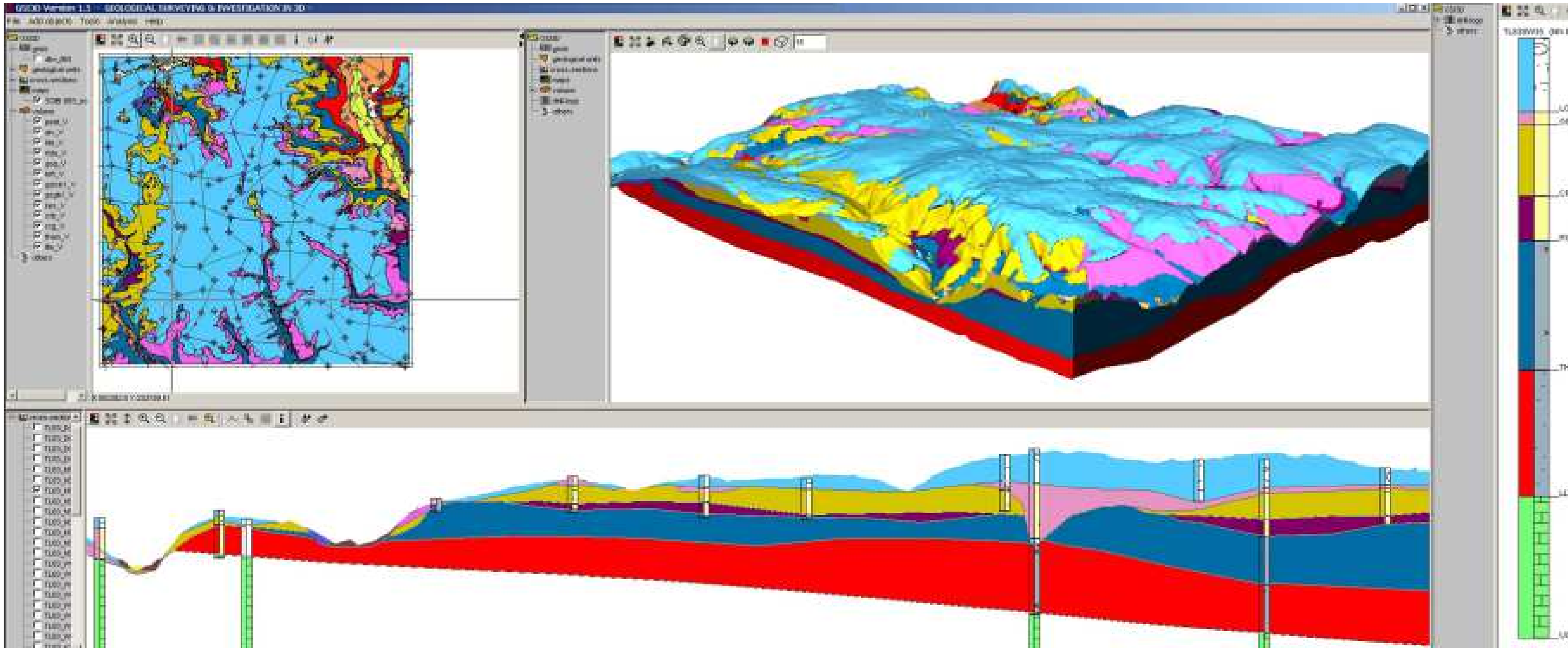
Murray Lark

Some generic issues in the life of an environmental statistican at BGS

- Quantifying uncertainty in interpretations
- Efficient sampling of variables at different spatial scales
- How best to use ad hoc data sets

Uncertainty in expert interpretation of a geological cross-section

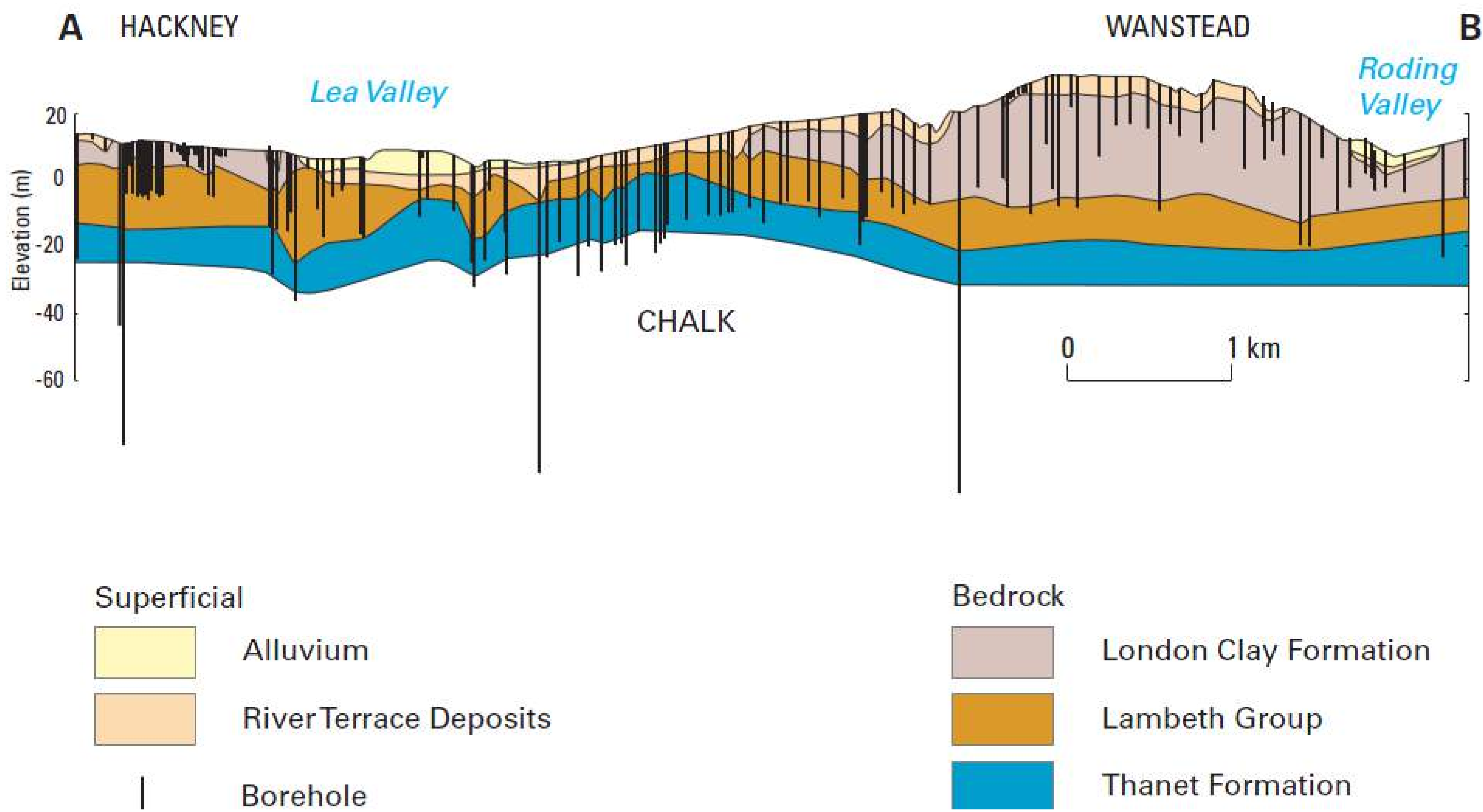
The problem



Site and design

- An 8-km section in London.
- Four bedrock units the London Clay, the Lambeth Group, the Thanet Sand (all Palaeogene) and the Chalk (Cretaceous), the latter appearing in about 10% of boreholes.
- Target feature here is the base of the London Clay (Eocene) proven by 51 boreholes.
- 28 geologists participated.
- Ten batches each with a unique set of 5 validation boreholes witheld was prepared by independent random sampling.
- Each geologist was allocated to one batch at random.
- A common interpretation of superficials and information on outcrop (as provide by a map) given to each participant.

Site and design

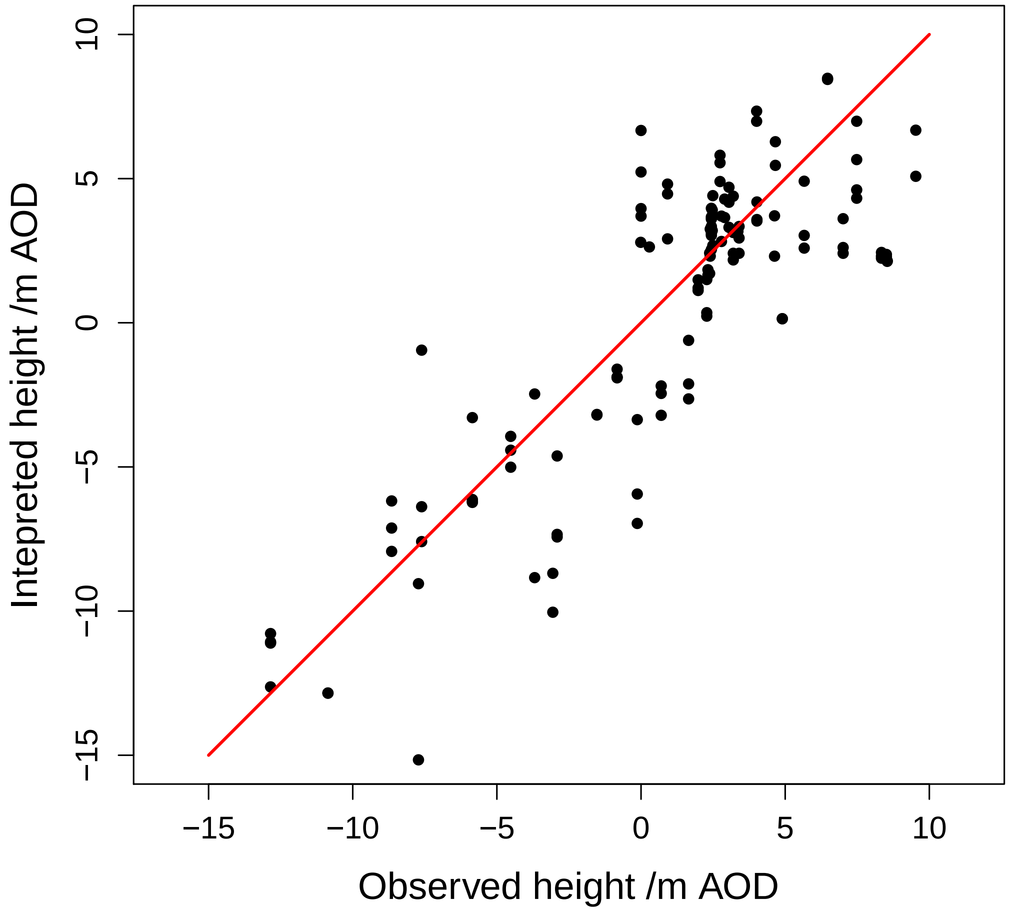


Questionnaire on modelling experience and responses received

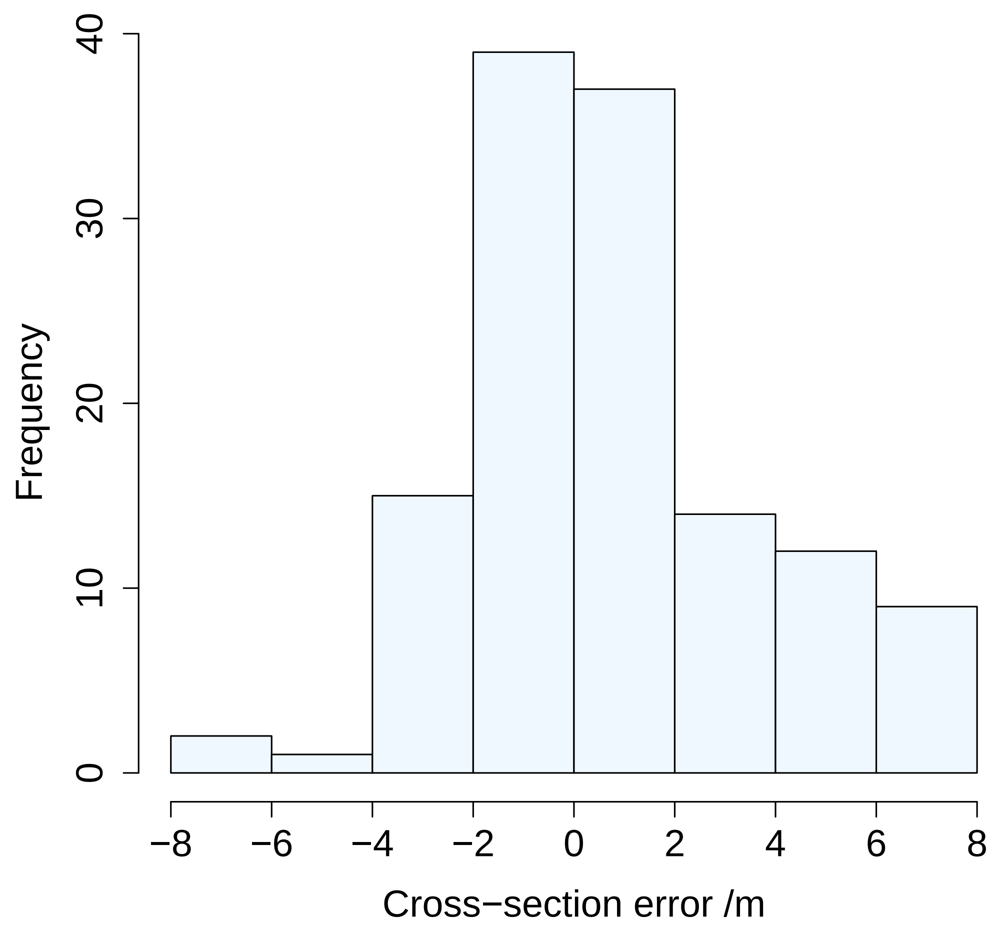
Question: ‘Please indicate with a tick which of the 4 descriptions below best reflects your experience of 3D modelling.’

Description	Number of participants selecting this description
I have no experience of geological modelling in 3D	2
I have some experience of geological modelling in 3D (perhaps through a training course) but little (up to 6 months) or no experience of modelling independently	8
I have moderate experience of geological modelling in 3D (six months to 2 years of modelling independently)	8
I have substantial experience of geological modelling in 3D (more than 2 years of modelling independently)	10

All validation observations of the interpreted and observed height of the base of the London Clay AOD. The red line is the bisector.



Histogram of cross-section errors.



The basic model

$$\boldsymbol{\varepsilon} = \boldsymbol{\mu} + \mathbf{X}_{\mathrm{b}}\boldsymbol{\beta}_{\mathrm{b}} + \mathbf{X}_{\mathrm{s}}\mathbf{Z}_{\mathrm{s}} + \boldsymbol{\eta}_{\mathrm{g}},$$

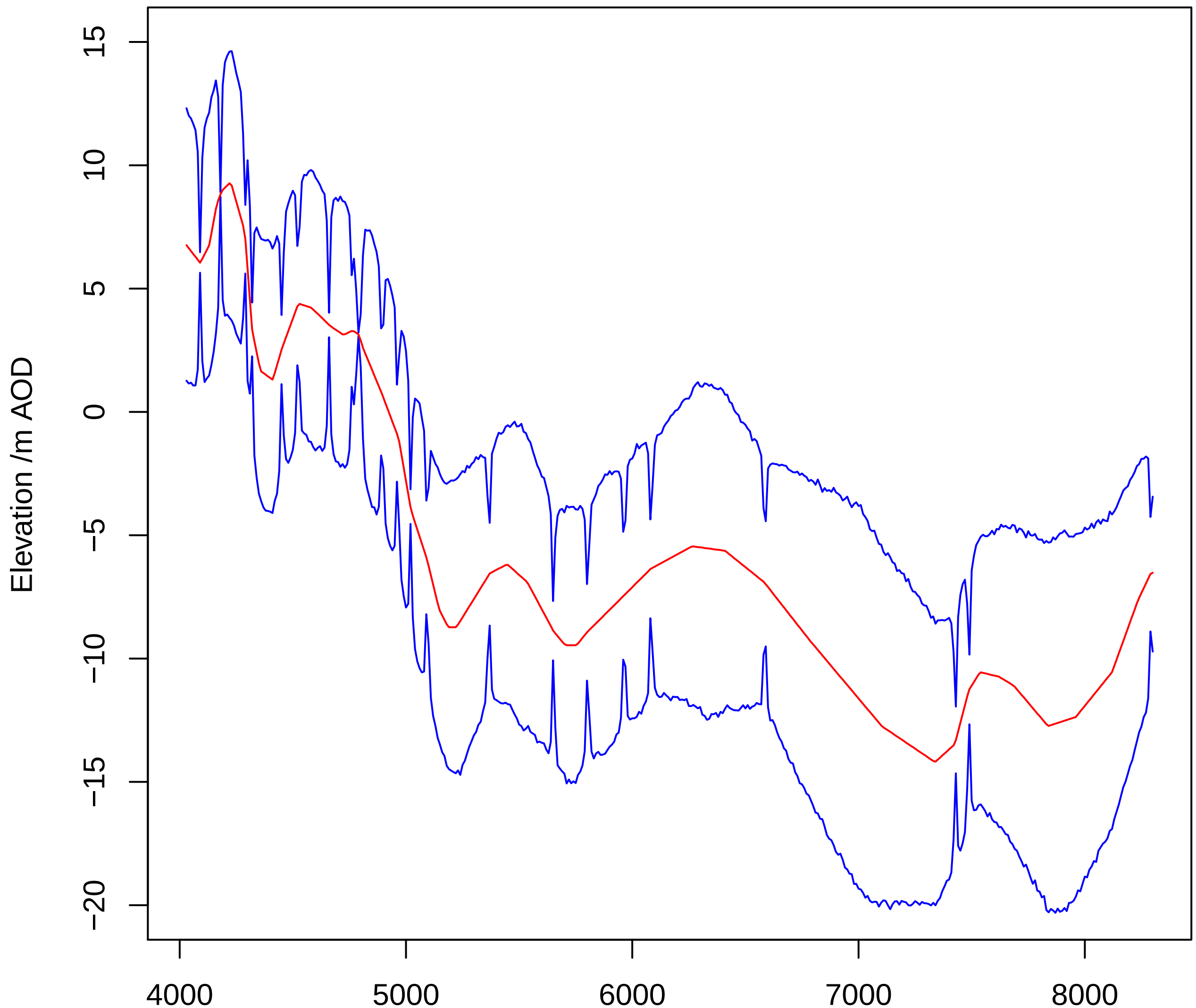
- Between-batch effect (negligible and dropped)
- Between-site within batch effect (spatially dependent, Matérn correlation).
- Between-geologist within-site effect (correlated within geologists at different sites).

Refinements

1. We tested the hypothesis that the between-geologist variance depends on distance to the nearest borehole.
2. We also tested the hypothesis that the between-geologist variance depends on the geologists experience (below).
3. Both hypotheses were supported, but (1) most strongly, and a model combining the effects did not improve on it.

	Between-geologist variance			
Experience	None	Some, <6 months	6 months – 2 years	>2 years
	4.44	2.25	1.32	0.46

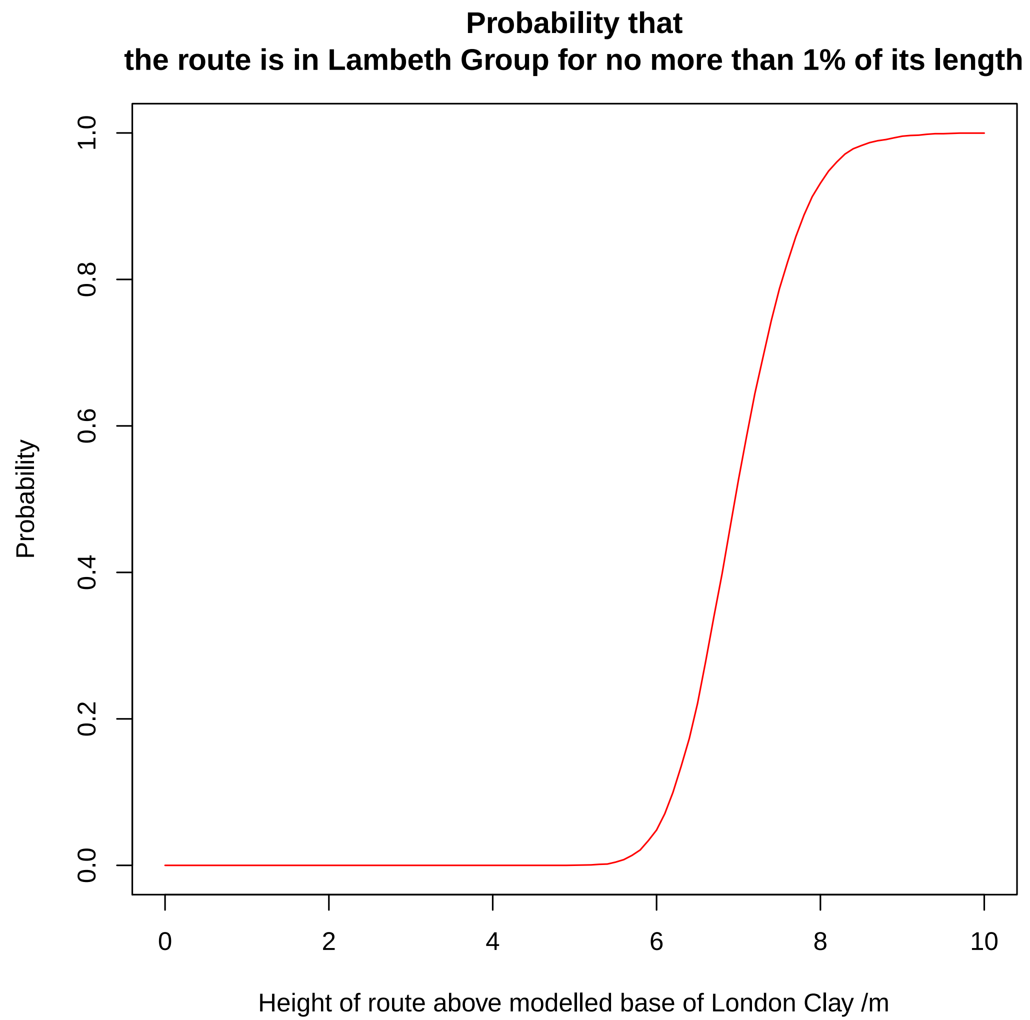
Confidence interval for an interpretation of the base of the London Clay.



Useful measures of uncertainty: decision analysis

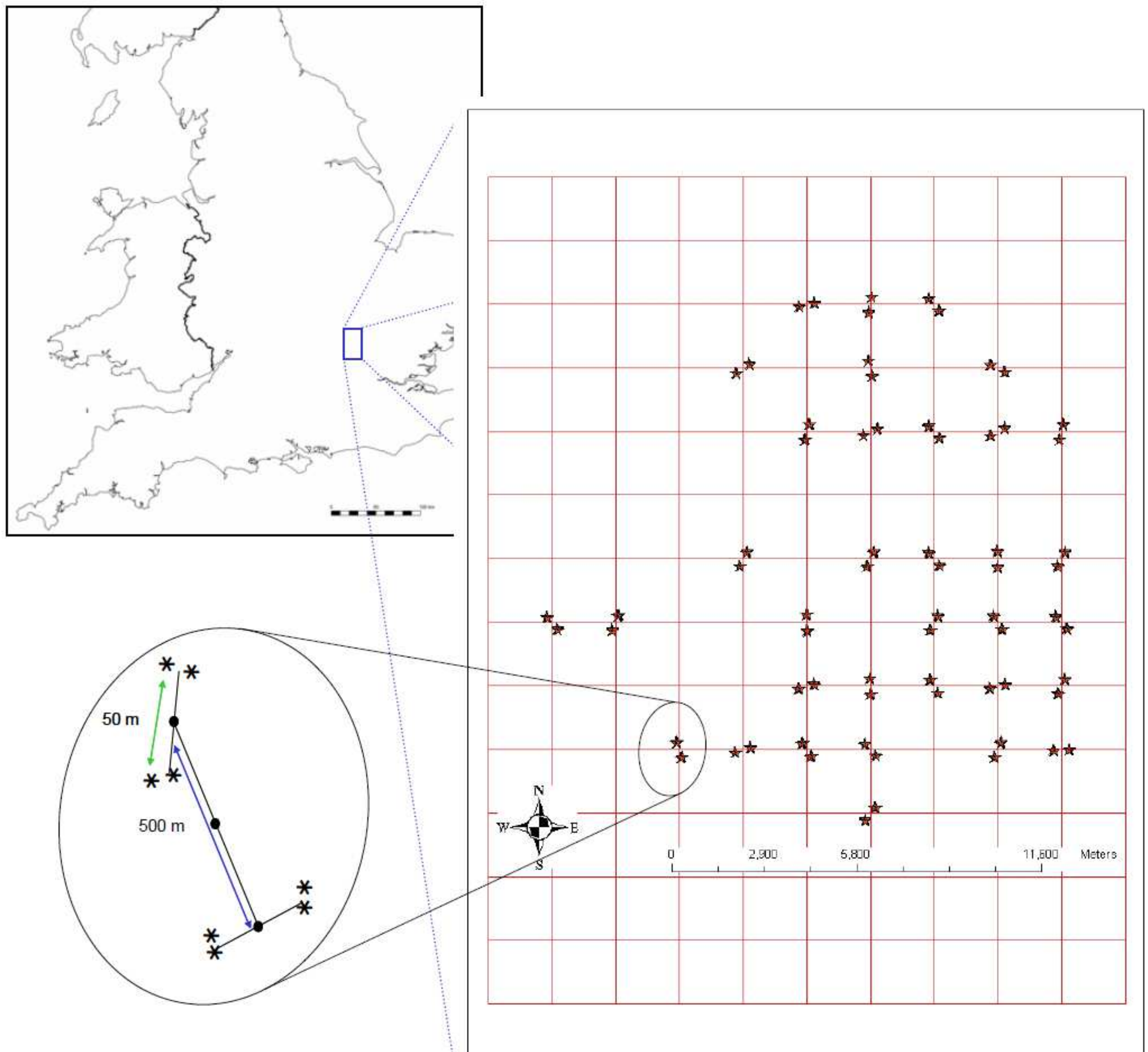
- Point-wise measures of uncertainty (e.g. the confidence intervals) are of limited value
- Users may often be interested in what the model predicts at more than one location (e.g. a route, piers).
- We are therefore interested in the joint uncertainty at these locations.

Useful measures of uncertainty: decision analysis

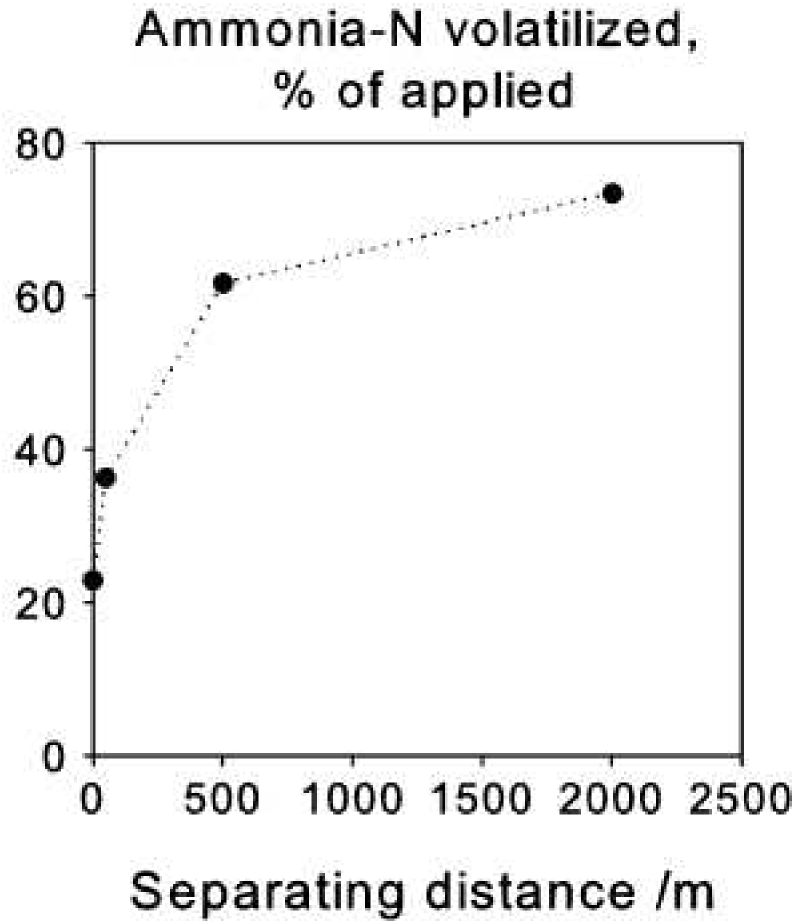


Spatially nested sampling and its optimization

Variation at multiple scales

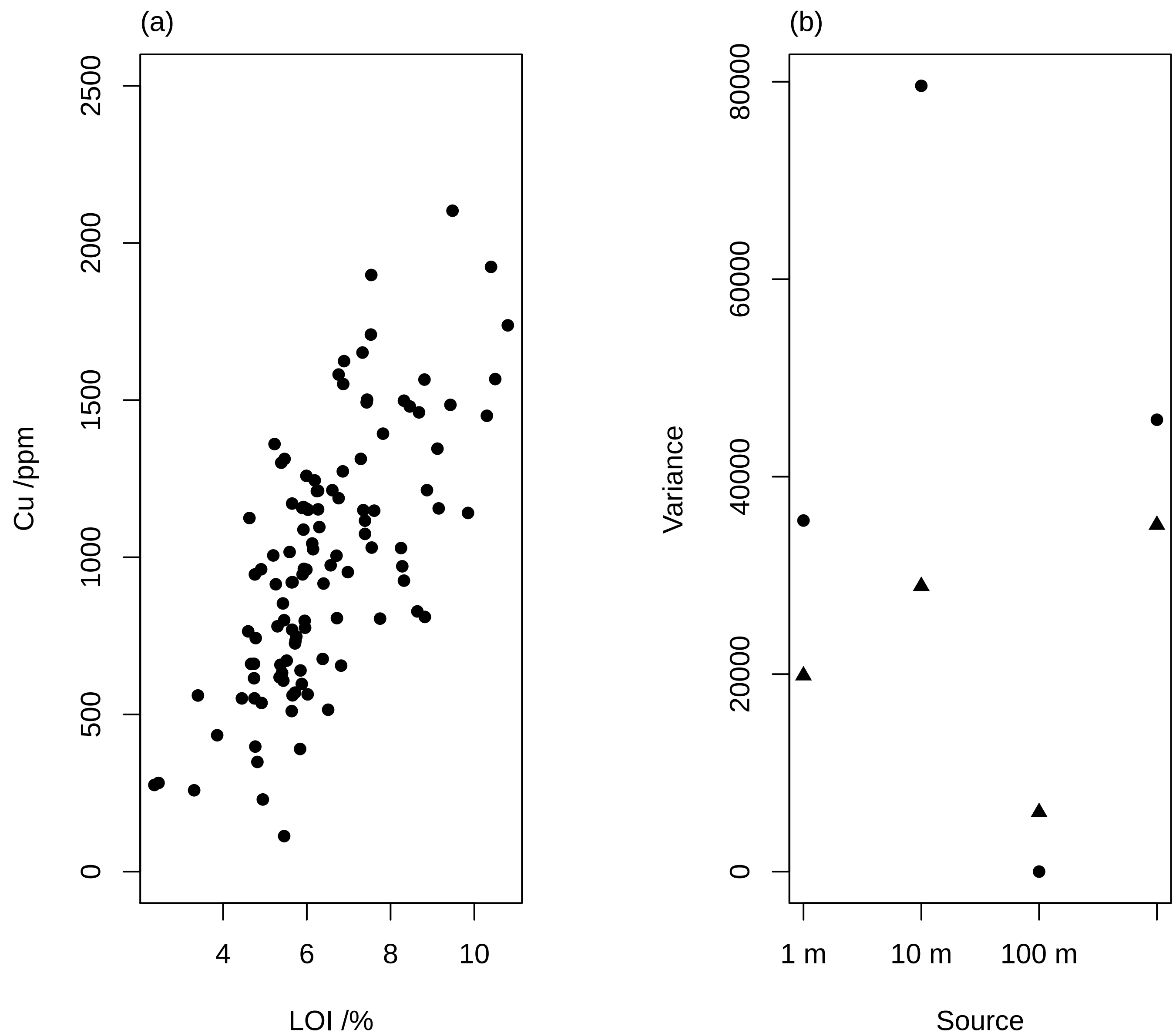


Variation at multiple scales



Variation at multiple scales

Variation at multiple scales



Optimization: Objective function

Given variance components $\boldsymbol{\psi} = \left\{\sigma_1^2, \sigma_2^2, \ldots, \sigma_m^2\right\}^{\mathrm{T}}$, and a set of design matrices $\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_m$, we can specify the elements $\{i, j\}$ of the expected information matrix, $\mathbf{M}$ for the variances:

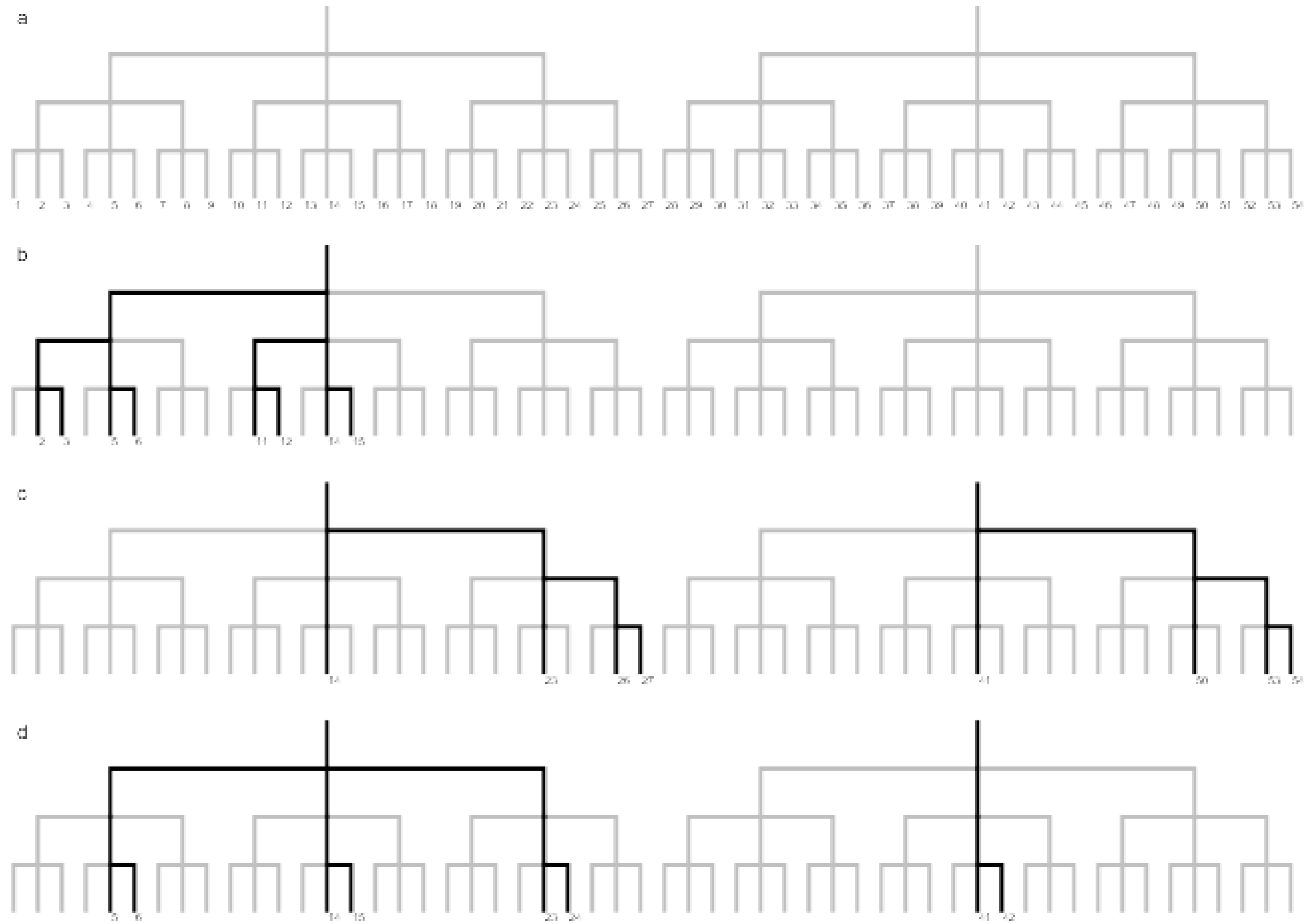
$$\mathrm{E}\left[\mathbf{M}_{i,j}\right] = \mathrm{E}\left[-\frac{\partial^2 \ell_{\mathrm{R}}\left(\boldsymbol{\psi}|\mathbf{z}\right)}{\partial\sigma_i^2\partial\sigma_j^2}\right] = \frac{1}{2}\mathrm{Tr}\left\{\mathbf{P}\mathbf{V}_i\mathbf{P}\mathbf{V}_j\right\}, \qquad (1)$$

where

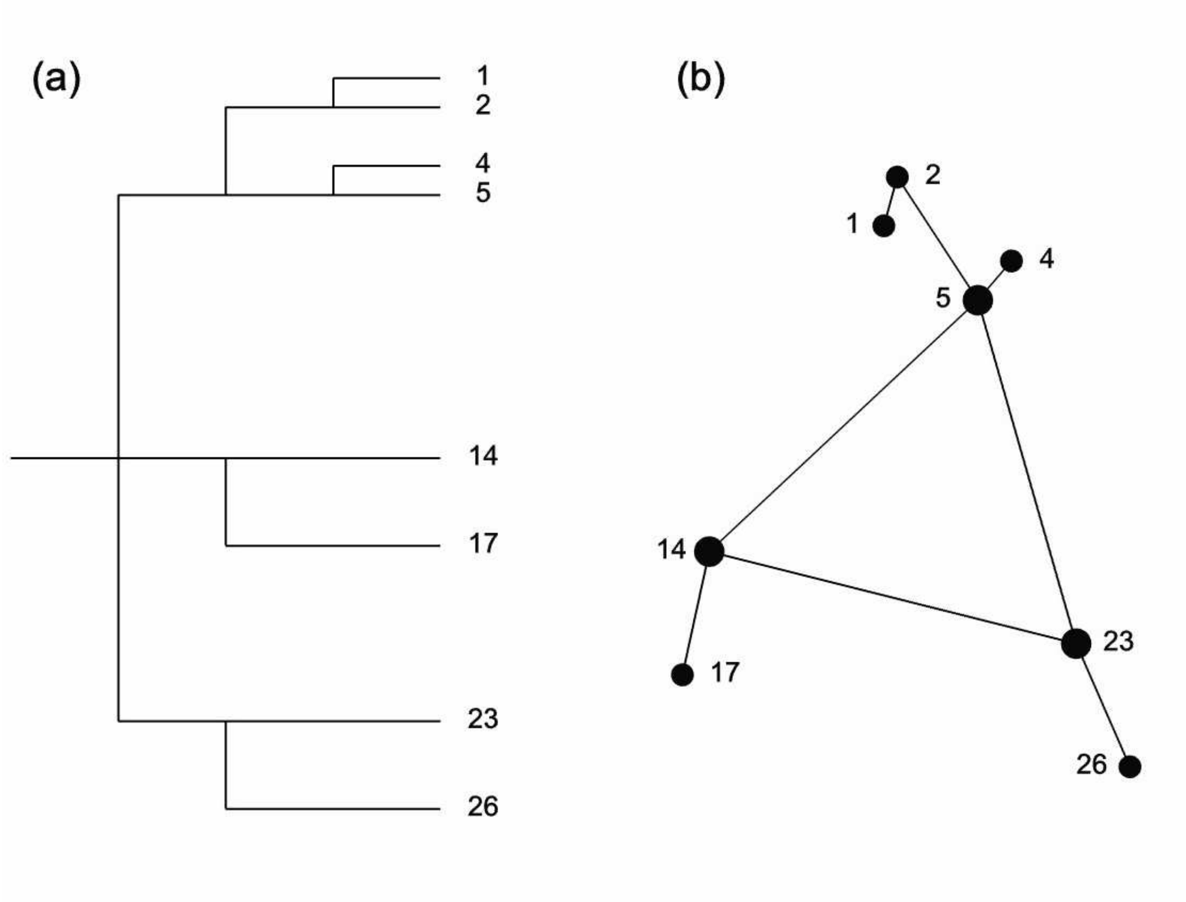
$$\mathbf{V}_i = \frac{\partial}{\partial\sigma_i^2}\mathbf{V} = \mathbf{U}_i\mathbf{U}_i^{\mathrm{T}}.$$

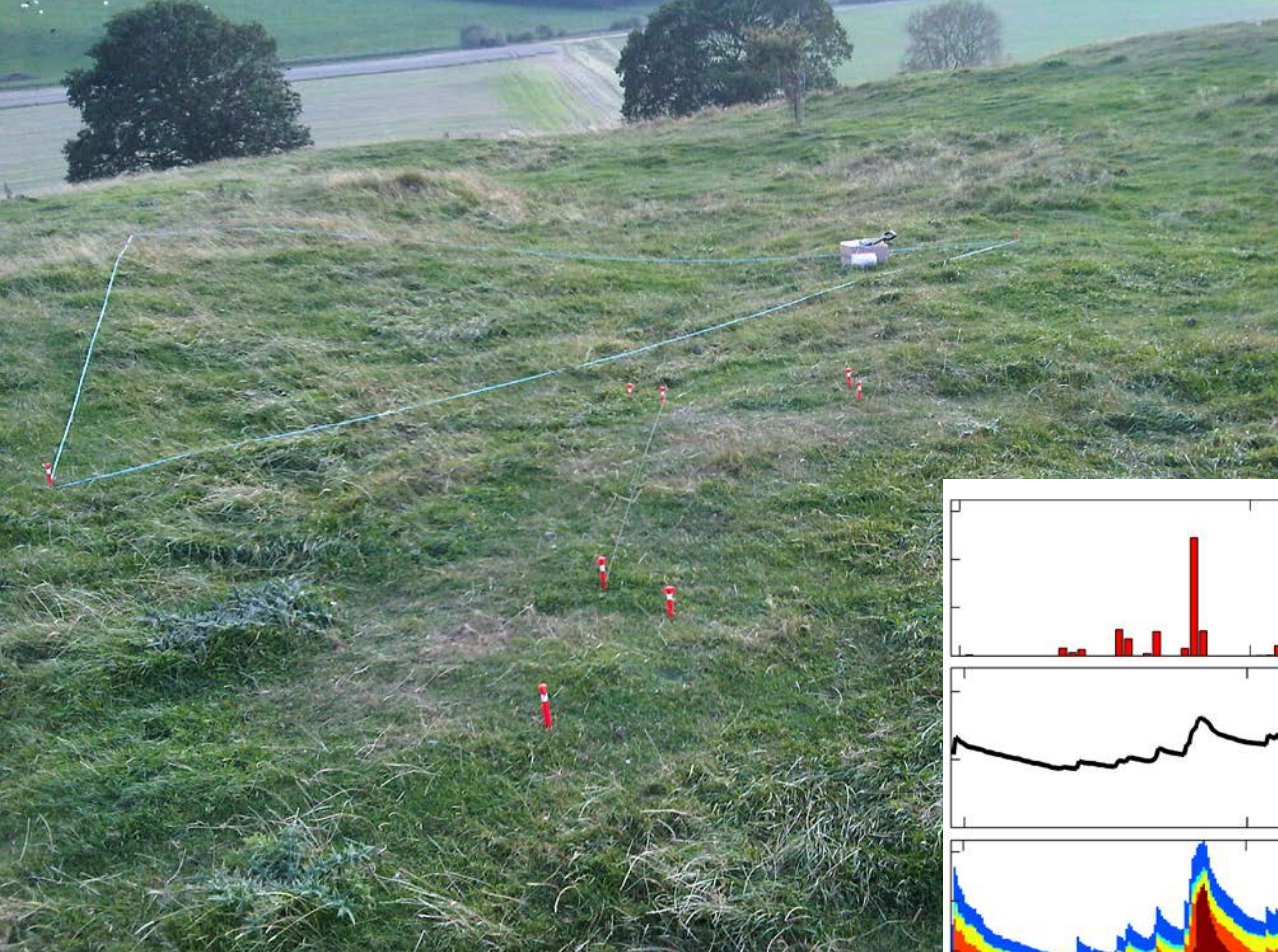
In this study the objective function was $S = \mathrm{Tr}\left\{\mathbf{M}^{-1}\right\}$.

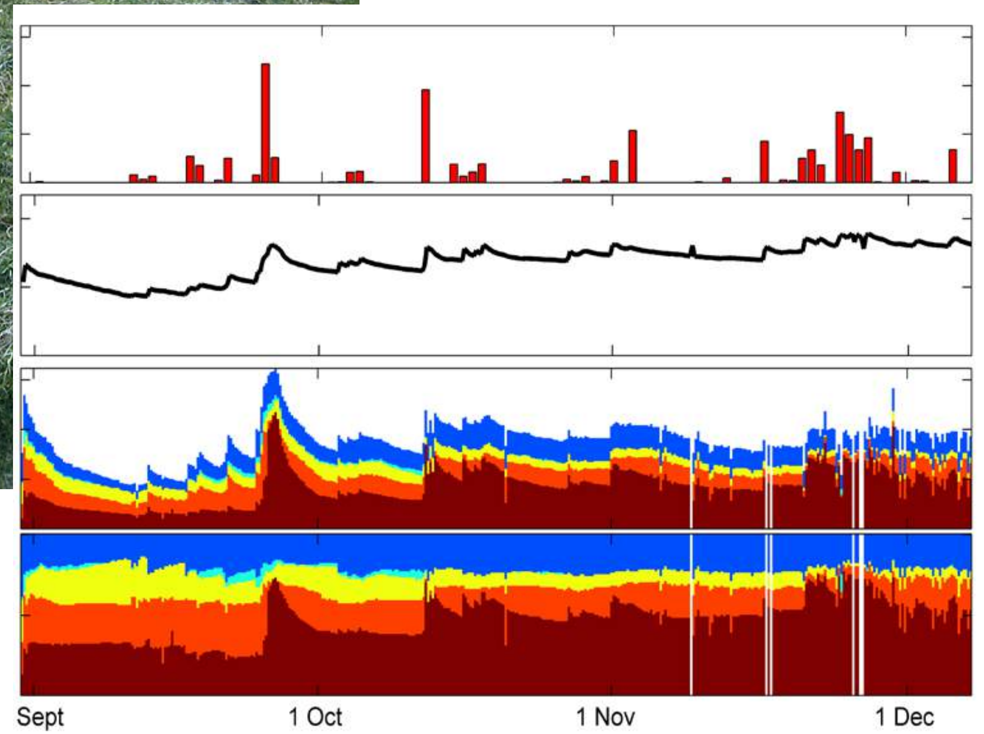
Optimization

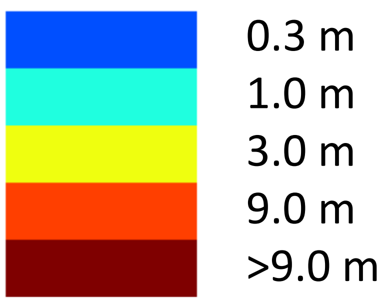


Optimization






Sept
1 Oct
1 Nov
1 Dec


0.3 m
1.0 m
3.0 m
9.0 m
>9.0 m

Continental-scale groundwater recharge: a meta-analysis

Introduction

- There is considerable interest in understanding Africa's groundwater resources, in particular their recharge and contributing factors.
- Colleagues at BGS have consolidated results from studies across the continent but:
  - They use different methodologies
  - They are far from independently distributed in space

Study sites

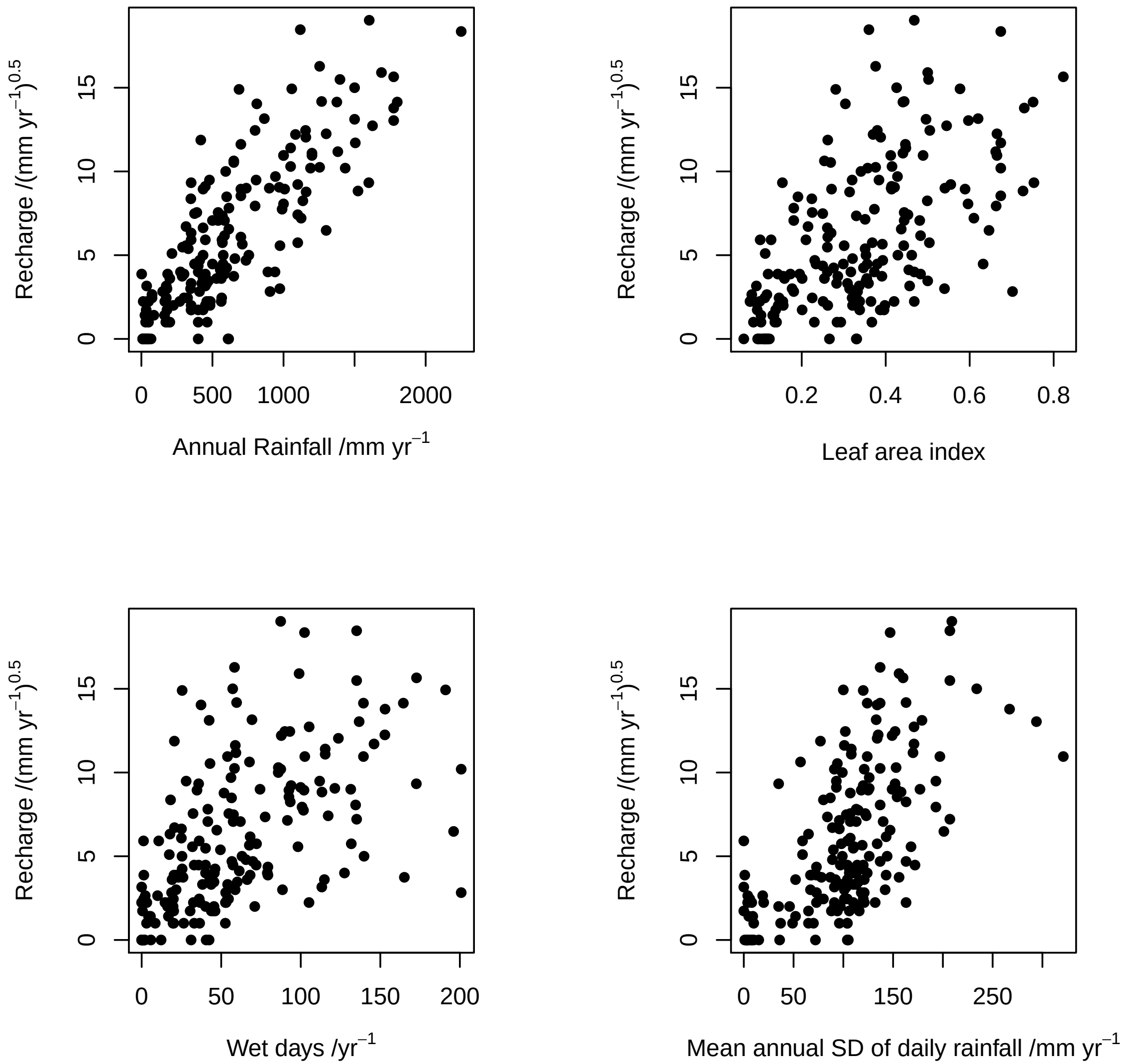
Available explanatory factors

- Mean annual rainfall
- Number of wet days
- Mean annual SD of rainfall
- Leaf Area Index

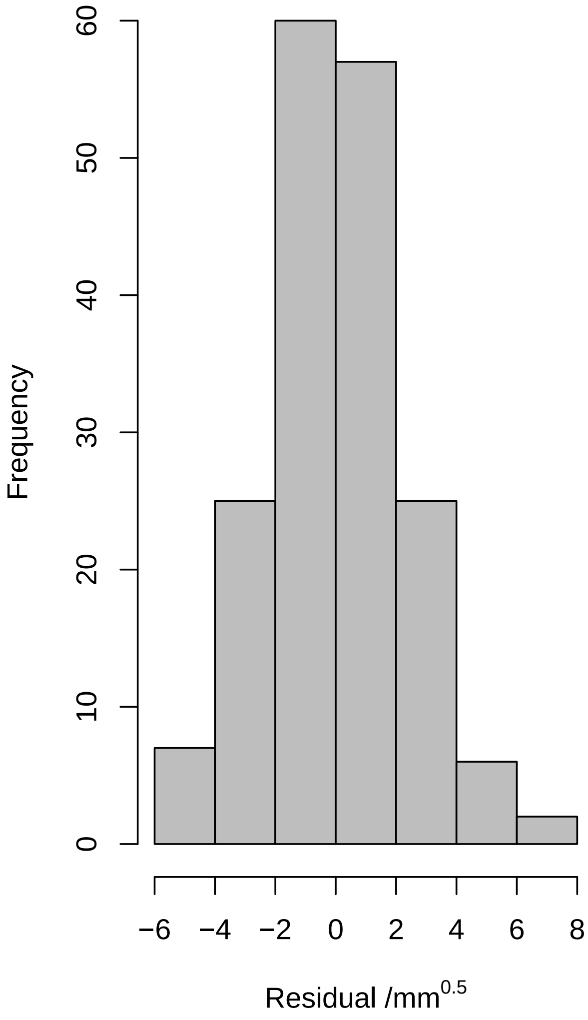
also

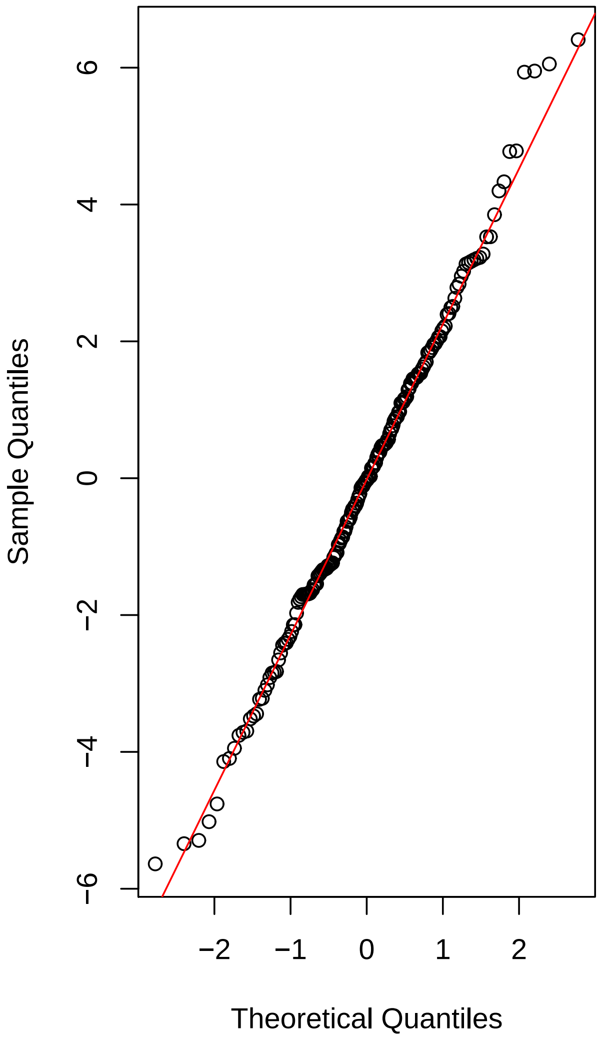
- Hydrogeological units (5 aquifer domains)
- Reference Soil Groups (16 usable, World Reference Base classes)

Exploratory analyses



Exploratory analyses



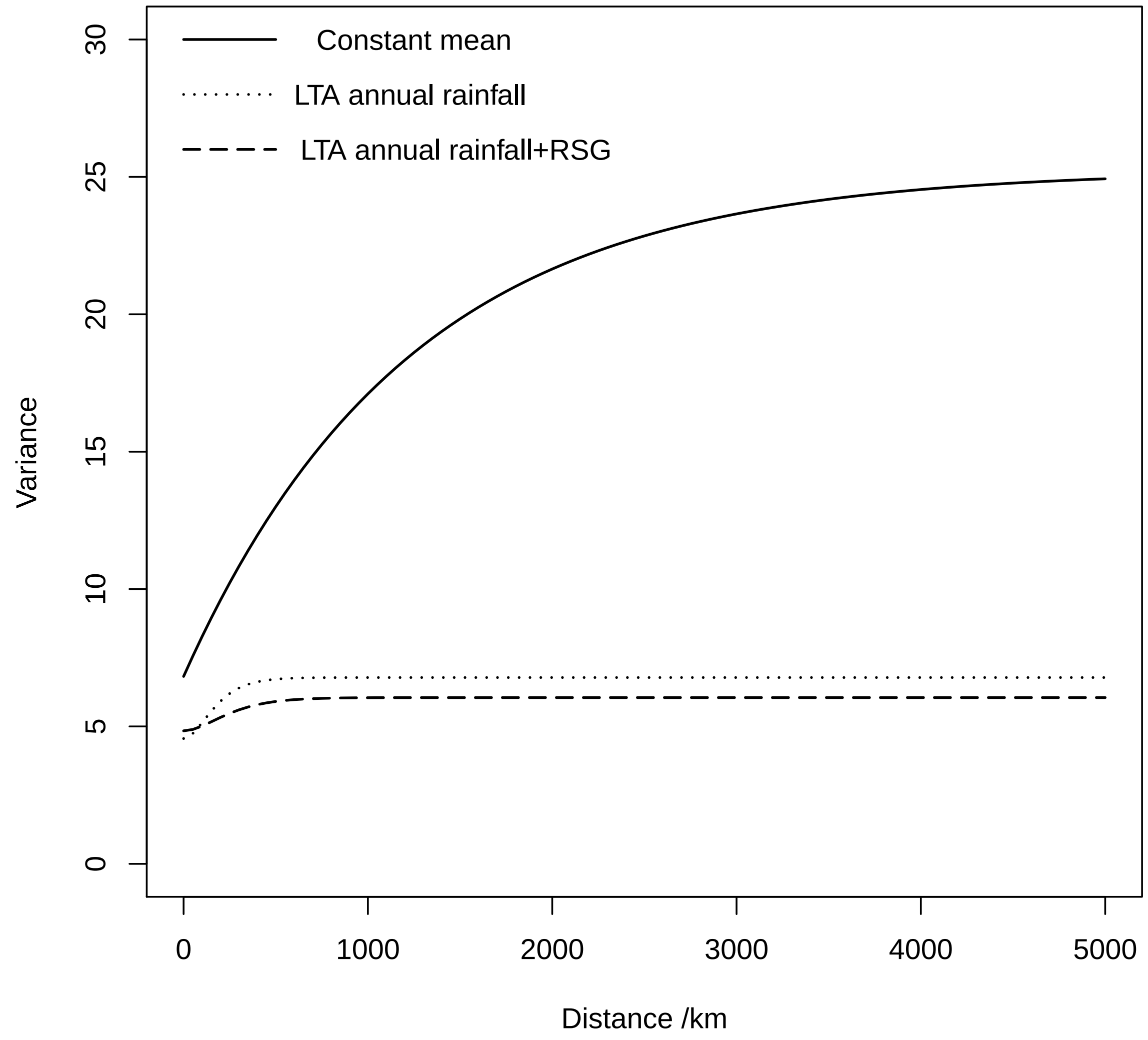


Linear mixed model

- Alternative random effects structures
- Correlated random effect (Matérn function) and ‘nugget’ term
- Estimation by REML.

- Selected model includes annual rainfall and Reference Soil Groups (additive effects only)
- Nugget variance is 80% of the random variation
- Correlated component has an effective range of $\sim$650 km

Linear Mixed Model



Methodologies and uncertainty

- Recharge was determined by a range of methods including chloride mass balance, soil moisture deficit, water table fluctuation
- A confidence class was allocated to each value, depending on method and context
- Not necessarily a linear scale, so treated as a categorical variable (5 levels, 5 is least confidence)
- There was a significant improvement by modelling the random variance as a function of the confidence level:

Confidence level	1	2	3	4	5
σ	1.81	1.92	2.55	2.75	5.27

Methodologies and uncertainty