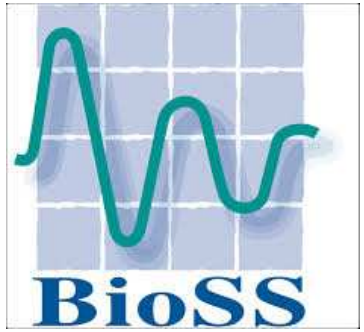
Some Bayesian applications in applied ecology and environmental sciences

Adam Butler (BioSS), joint work with:



Case Study 1 - Kate Searle & Bethan Purse (CEH)

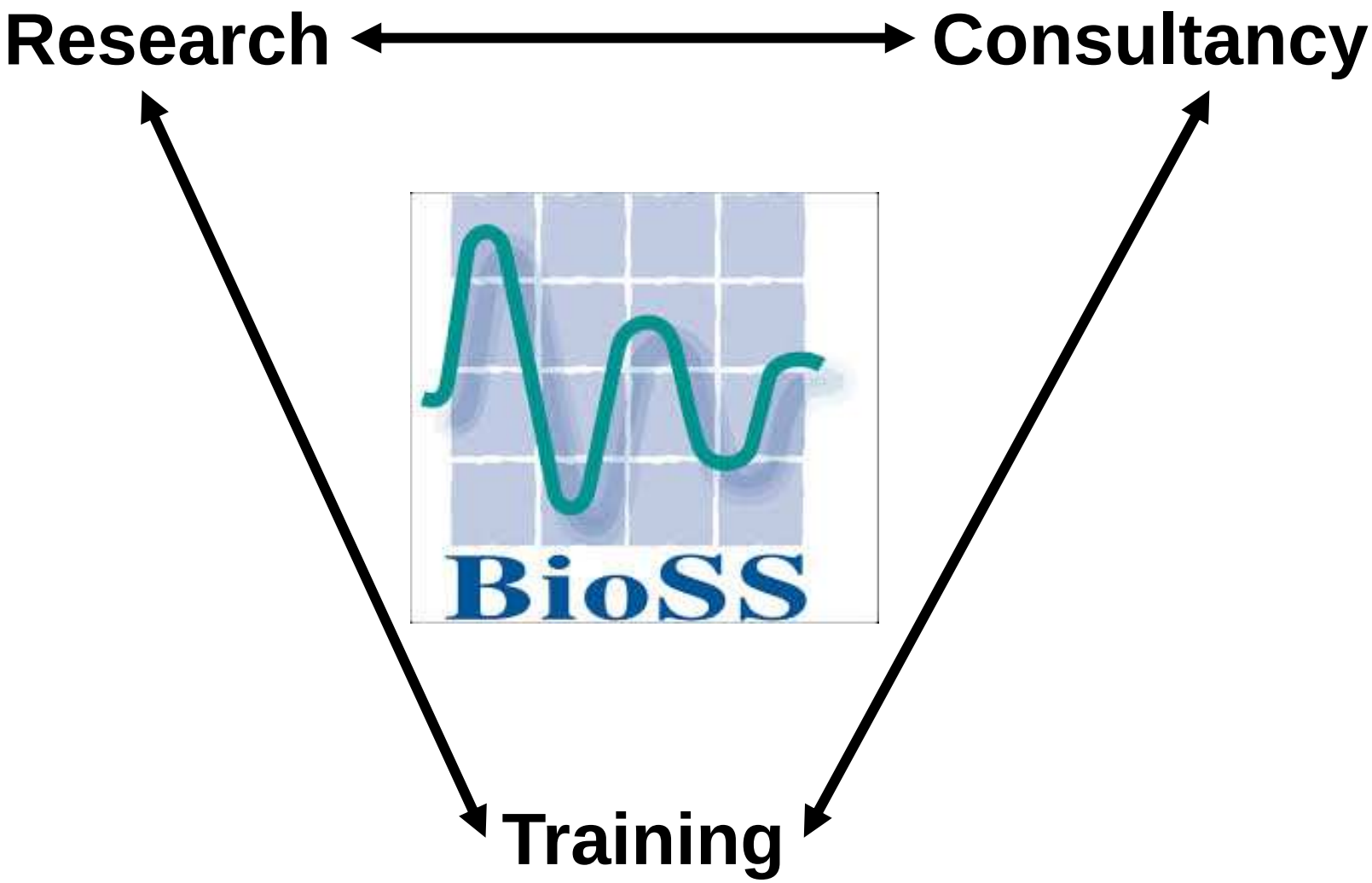
Case Study 2 - Stephen Catterall & Glenn Marion (BioSS), Alex Cook (National University of Singapore), Gavin Gibson (Heriot-Watt)

Case Study 3 - Vera Eory & Kairsty Topp (SRUC), Jyldyz Tabyldy Kyzy & Clare Bond (University of Aberdeen)

Case study 1: Modelling spatial & temporal trends in midge abundance

Case study 2: Modelling the spread of an invasive species

If time - **Case study 3:** Uptake of greenhouse gas mitigation measures of greenhouse gases


Research
Consultancy
BioSS
Training

Bayesian?

Not always!

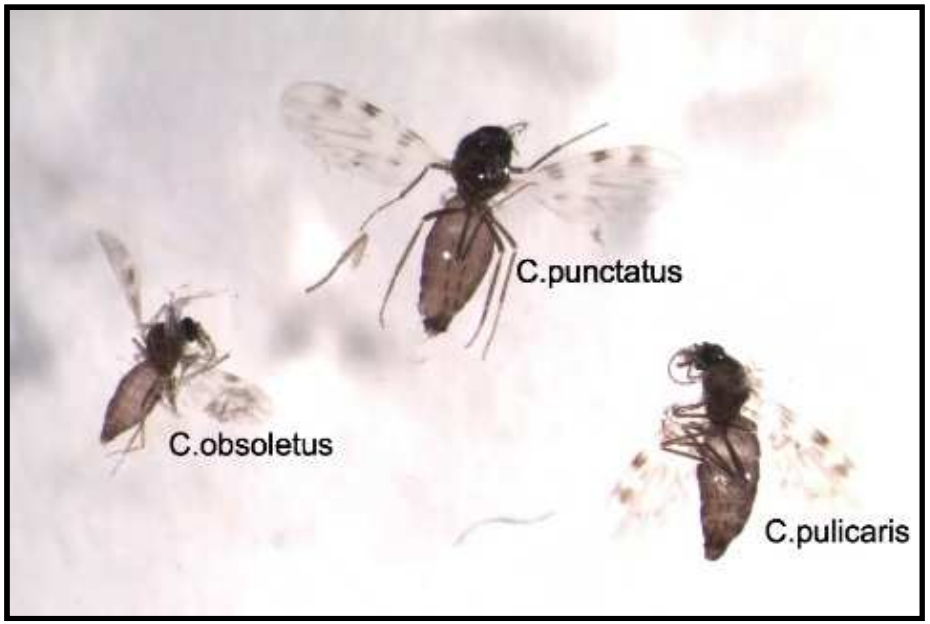
…but increasingly often

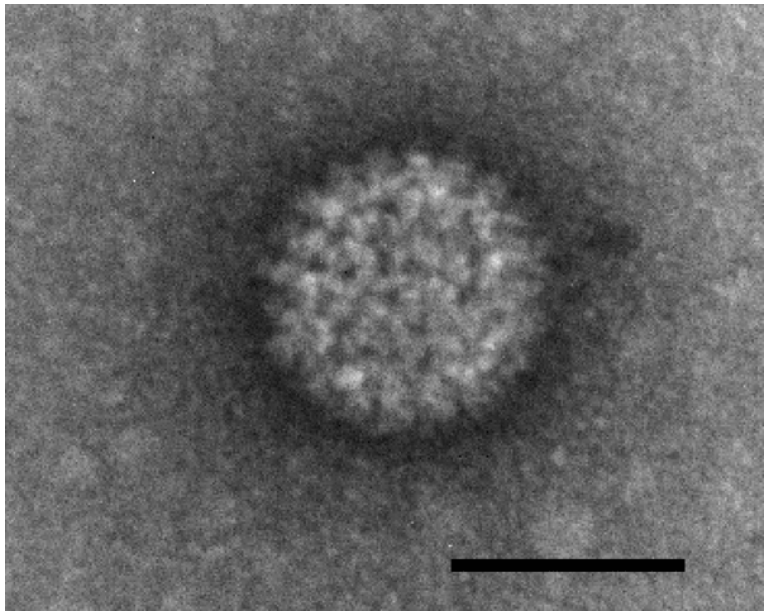
…this is largely because the scientists that I work with are dealing with increasingly complicated forms of data, and trying to answer increasingly challenging problems, for which appropriate non-Bayesian approaches are simply not available…

Case Study 1: Modelling spatial and temporal trends in midge abundance

Background

- Rapid appearance of new vector-borne diseases worldwide, with huge impacts on livelihoods and economies
- **Midge (*Culicoides*)** species are increasingly being identified as a key vector in transmission – e.g. of the **bluetongue virus**





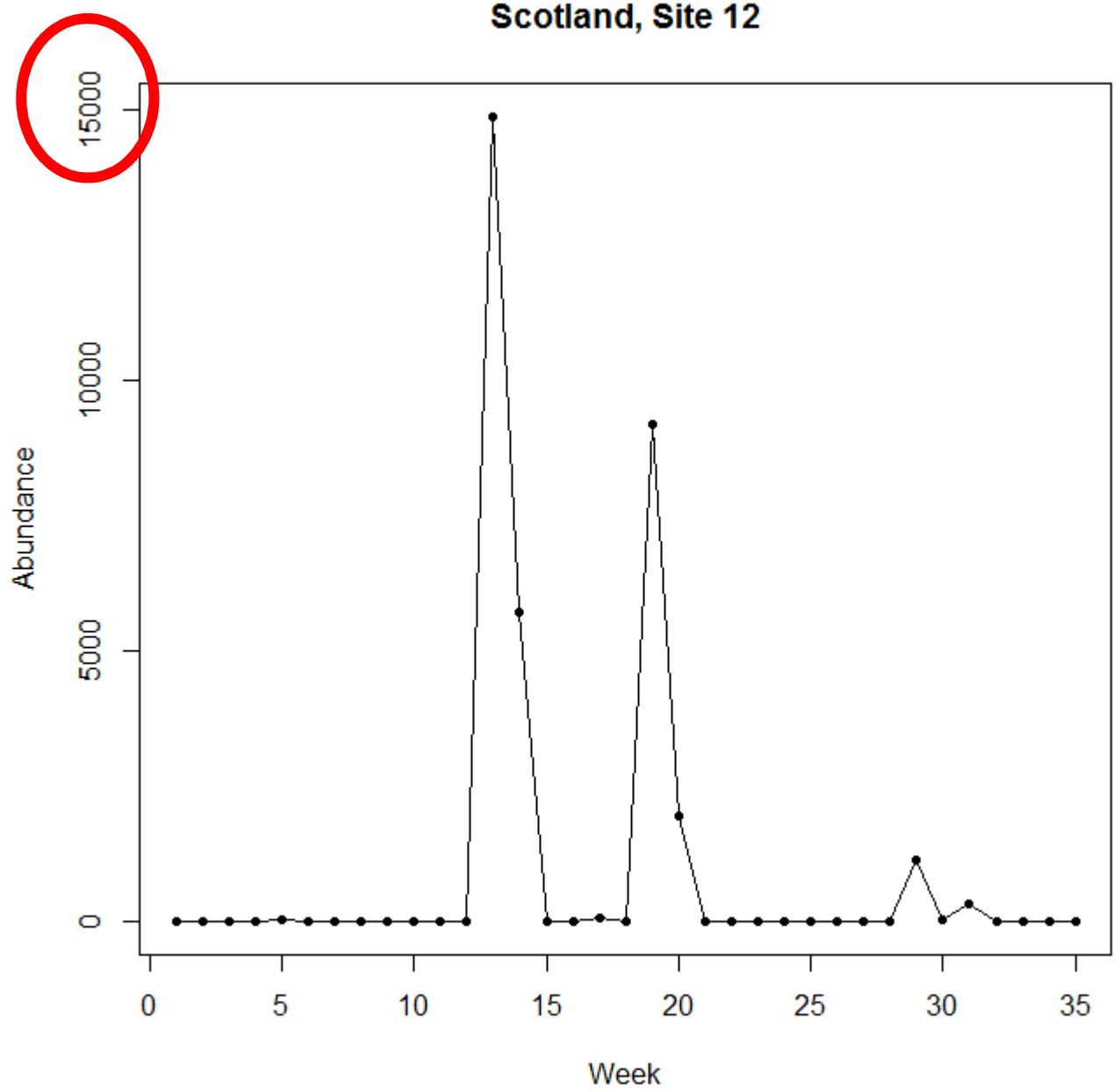
- A key determinant of risk is the contact rate between vectors (e.g. midges) and hosts (e.g. cattle)
- **It is therefore crucial to understand spatio-temporal trends in the abundance of midge species**

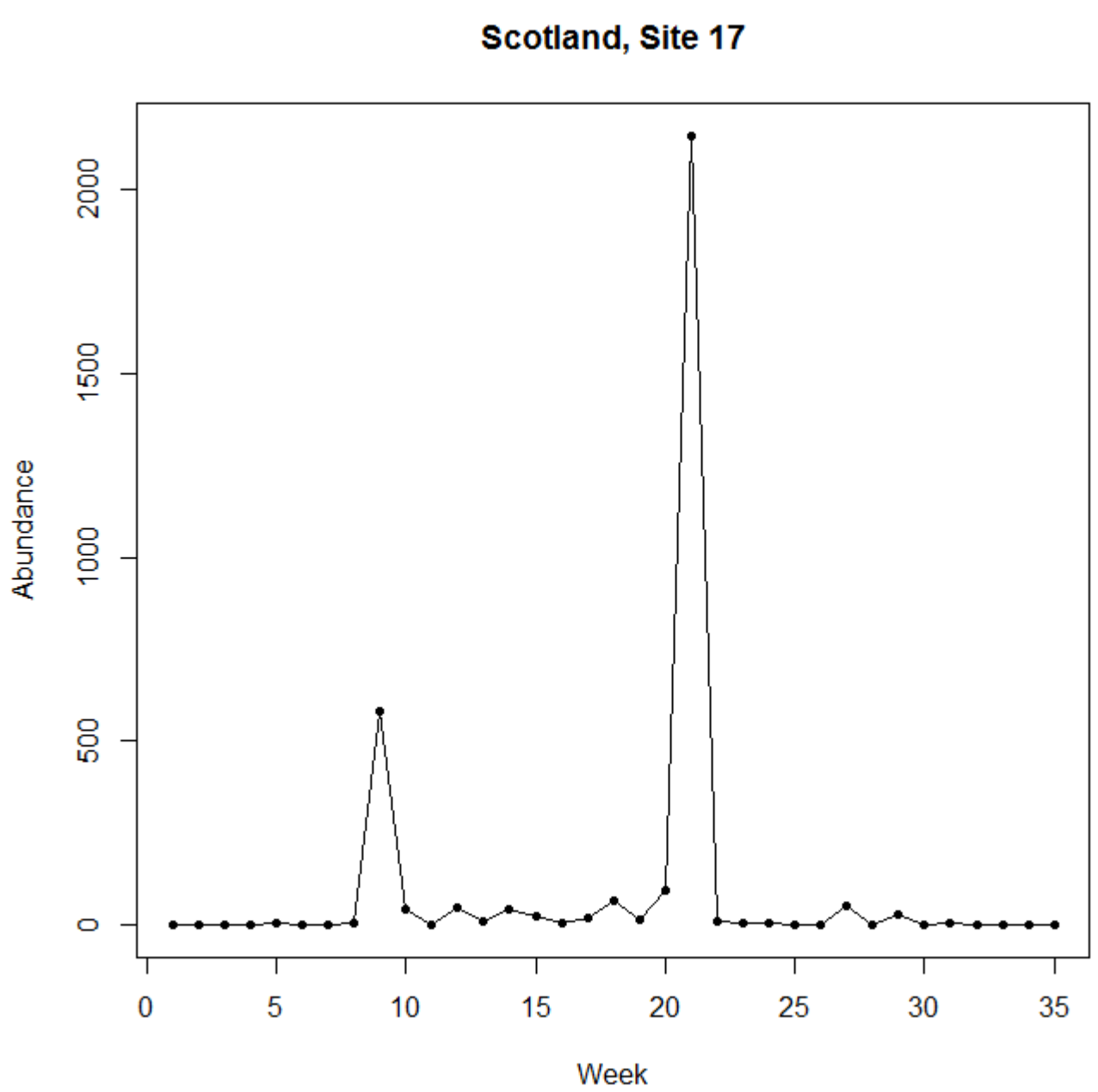
Data

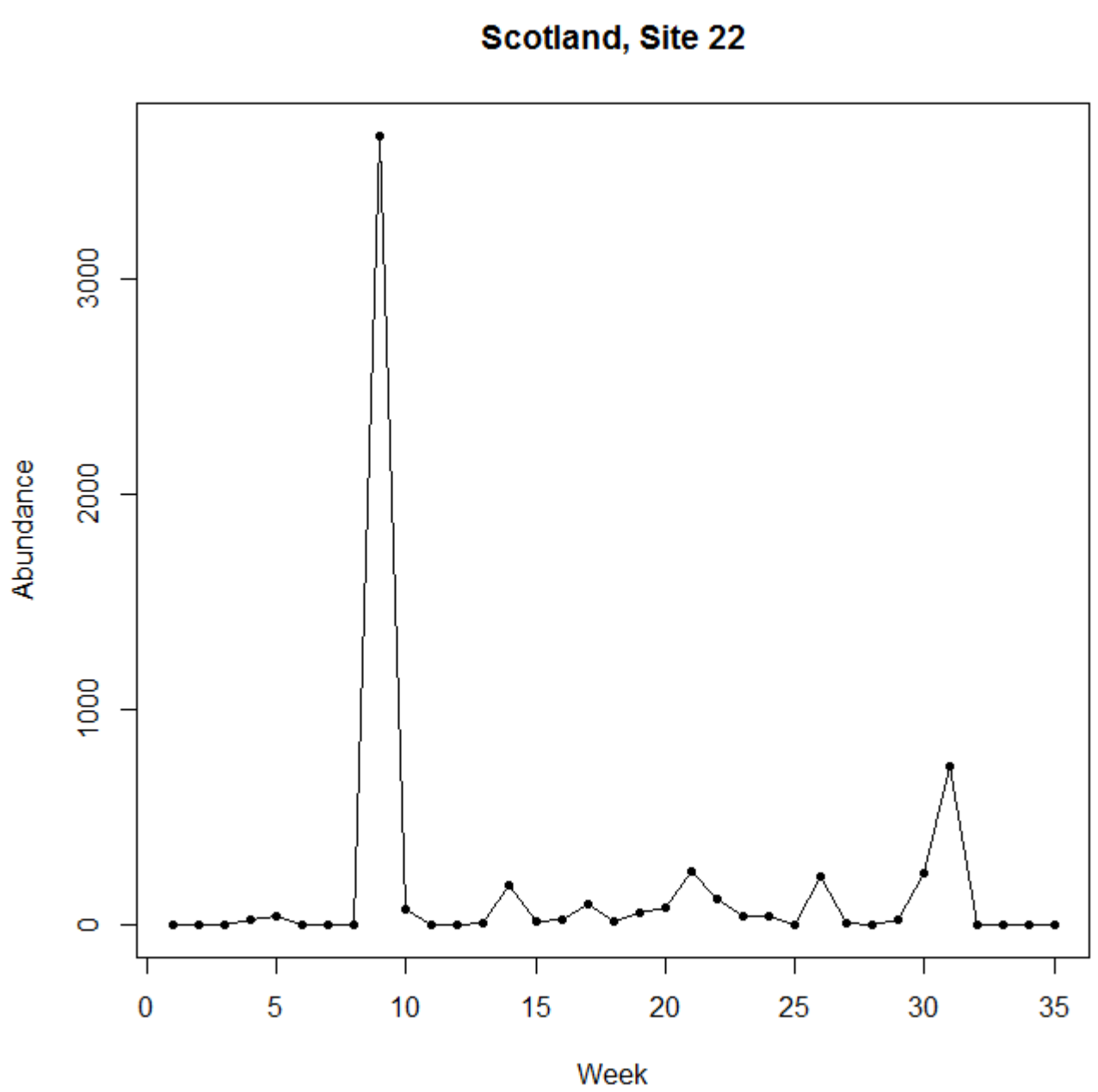
- **Midge abundance monitored weekly over multiple years at 500+ trapping sites across Europe**

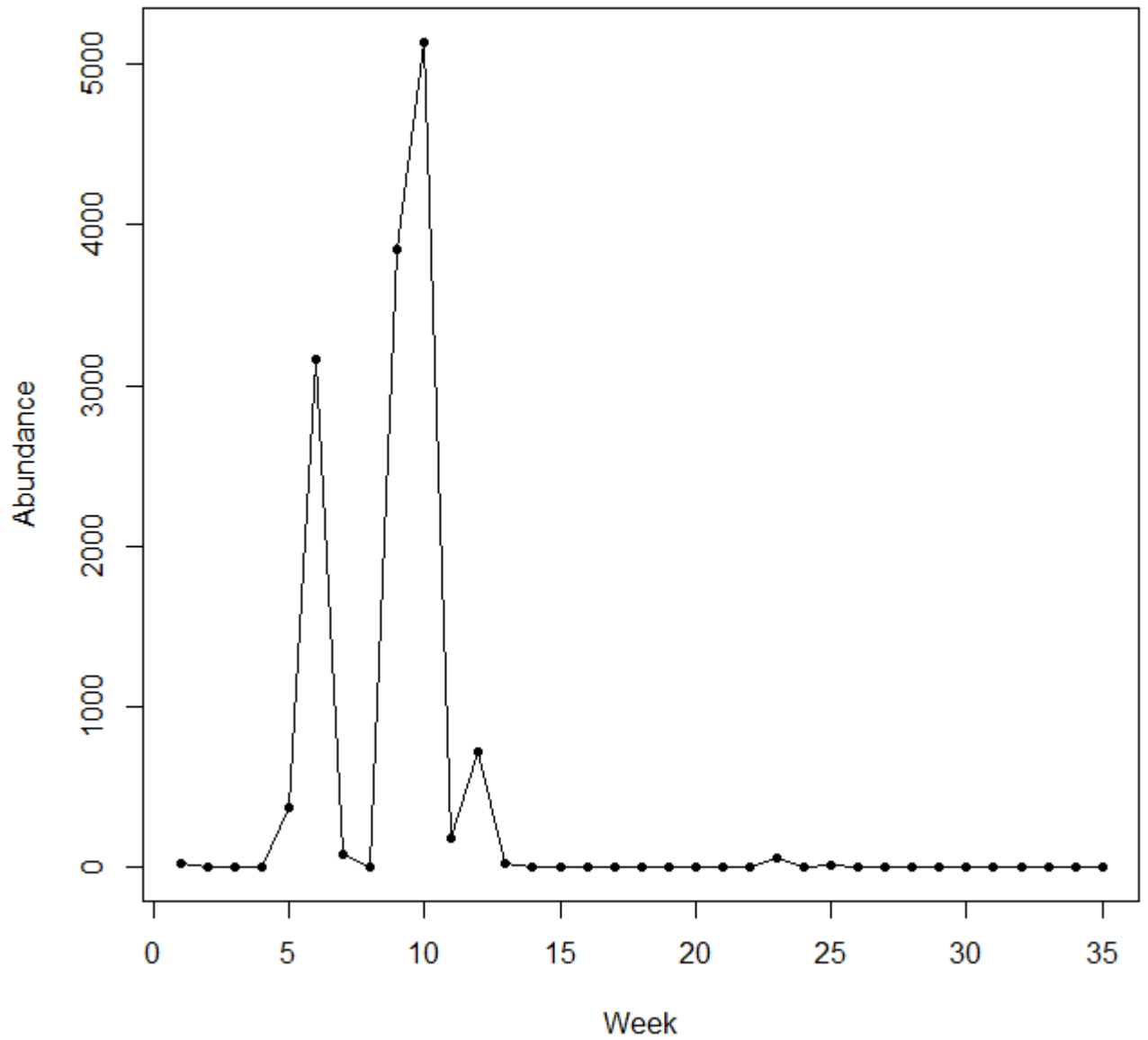
- We have access to data for 400 site-year combinations in Spain and the UK

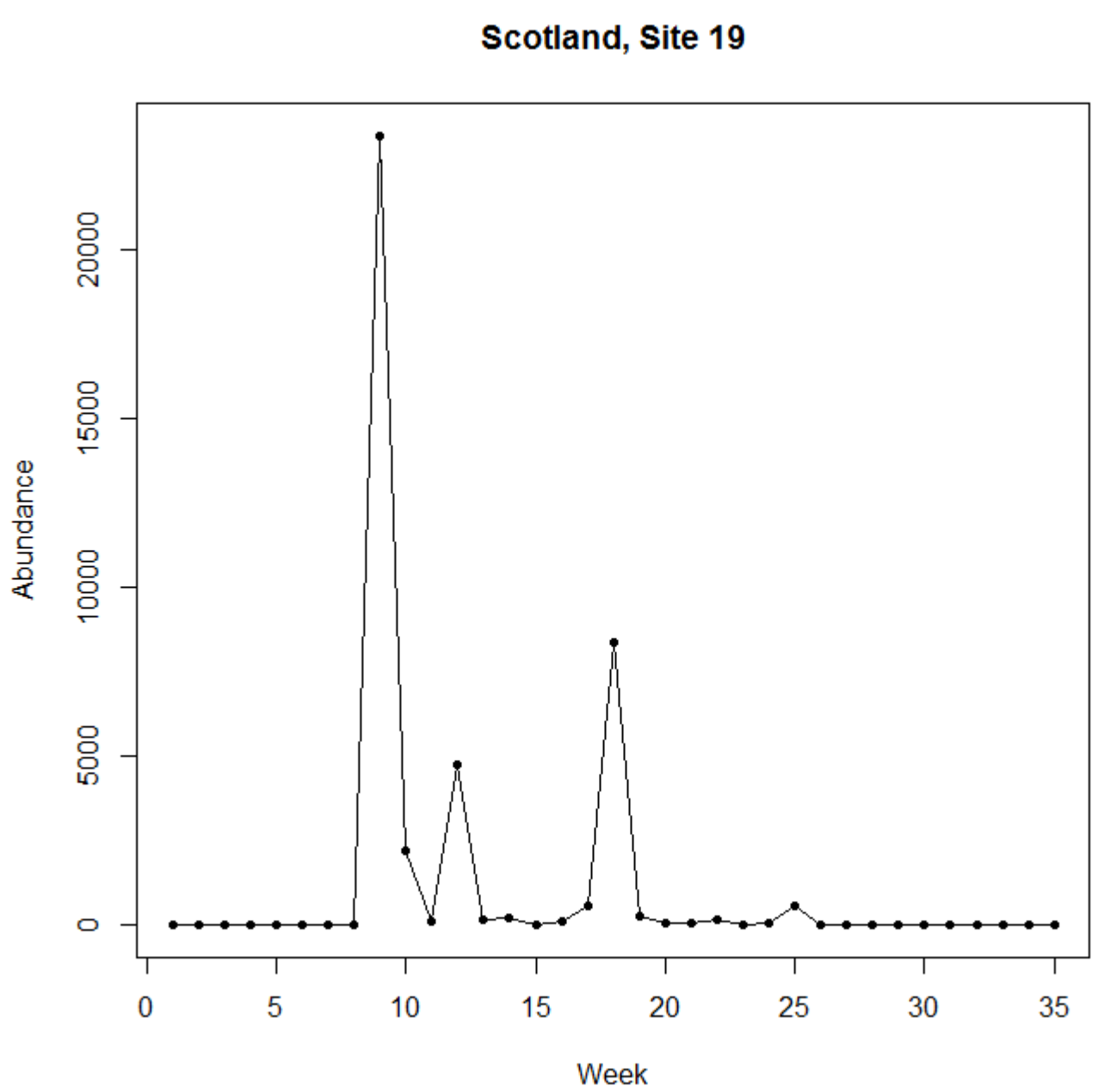
- Data contain many **zero values**
- Abundance **varies** over space and time by orders of magnitude…

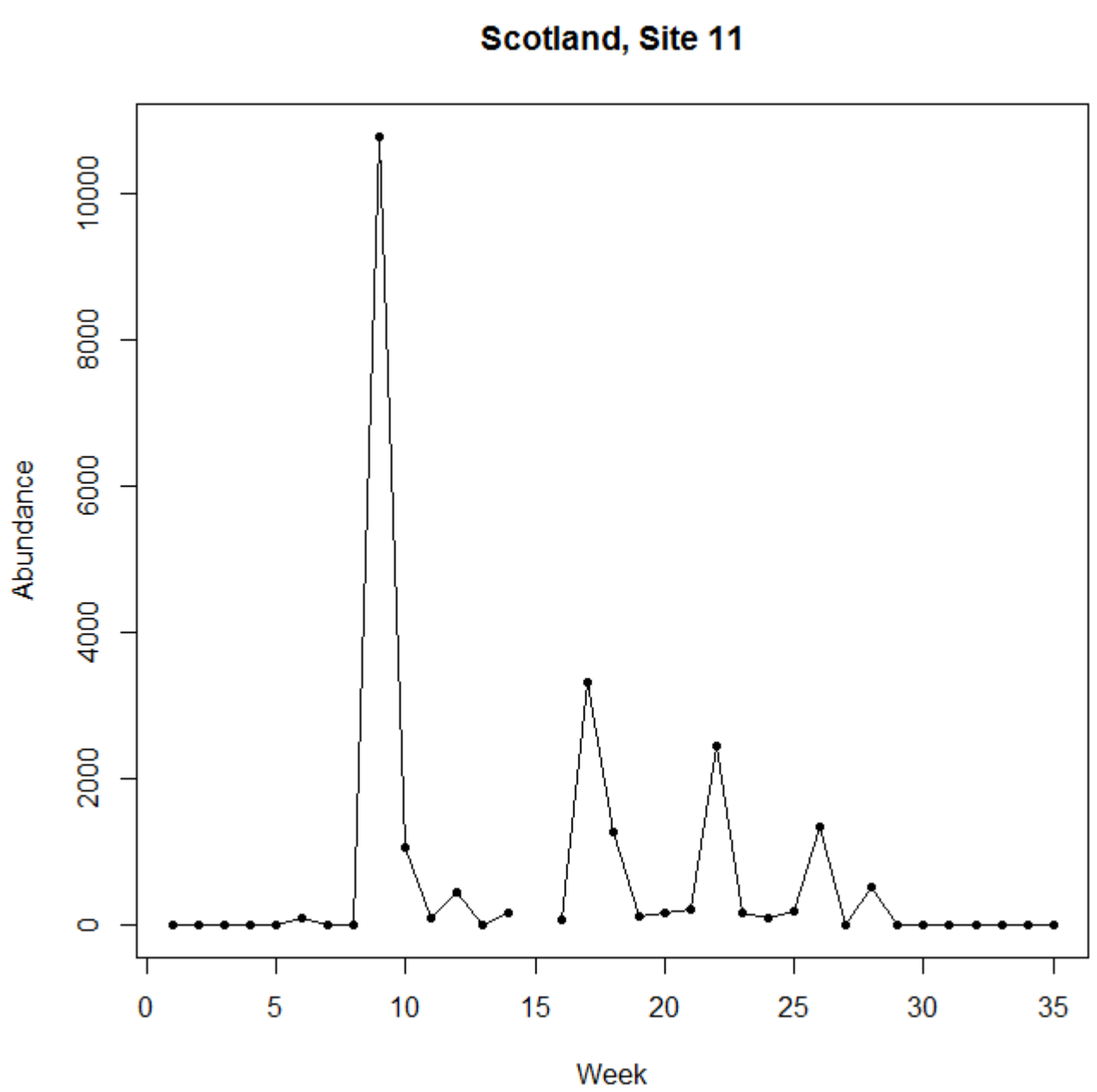

Scotland, Site 12
Abundance
0
5000
10000
15000
0
5
10
15
20
25
30
35
Week


Scotland, Site 17
Abundance
0
500
1000
1500
2000
0
5
10
15
20
25
30
35
Week


Scotland, Site 22
Abundance
0
1000
2000
3000
0
5
10
15
20
25
30
35
Week


Scotland, Site 32
Abundance
0
1000
2000
3000
4000
5000
0
5
10
15
20
25
30
35
Week


Scotland, Site 19
Abundance
0
5000
10000
15000
20000
0
5
10
15
20
25
30
35
Week


Scotland, Site 11
Abundance
0
2000
4000
6000
8000
10000
0
5
10
15
20
25
30
35
Week

y_{ij} = count for day t_j at site-year combination i

$\boldsymbol{x}_i$ = covariates for site−year combination i

Objectives

- Identify characteristics of the **seasonal phenology** – when are the peaks, how big are they, and how long do they last?

- Construct a model that describes these phenological characteristics in terms of **environmental covariates**

- Use the model to **predict** the seasonal phenology at all unmonitored locations throughout Europe

Previous work

Consultancy…

A multi-stage modelling approach (Searle *et al*., 2013):

- Model abundance at each site using a **GAM**
- Extract phenological characteristics (e.g. timing, magnitude and width of peaks) from each fitted model
- Use a regression model to describe the **spatial variations** in these characteristics in terms of environmental covariates

Issues:

(1) GAMs are not designed to provide a robust basis for estimating phenology – i.e. the number of peaks, and the characteristics of these peaks

(2) GAMs tend to deal poorly with the "**burstiness**" of the data (abundance is low and fairly constant for a long time, then changes rapidly once an outbreak has begun)

(3) Full quantification of uncertainty is difficult within this approach

Strategy

It would seem to be preferable to:

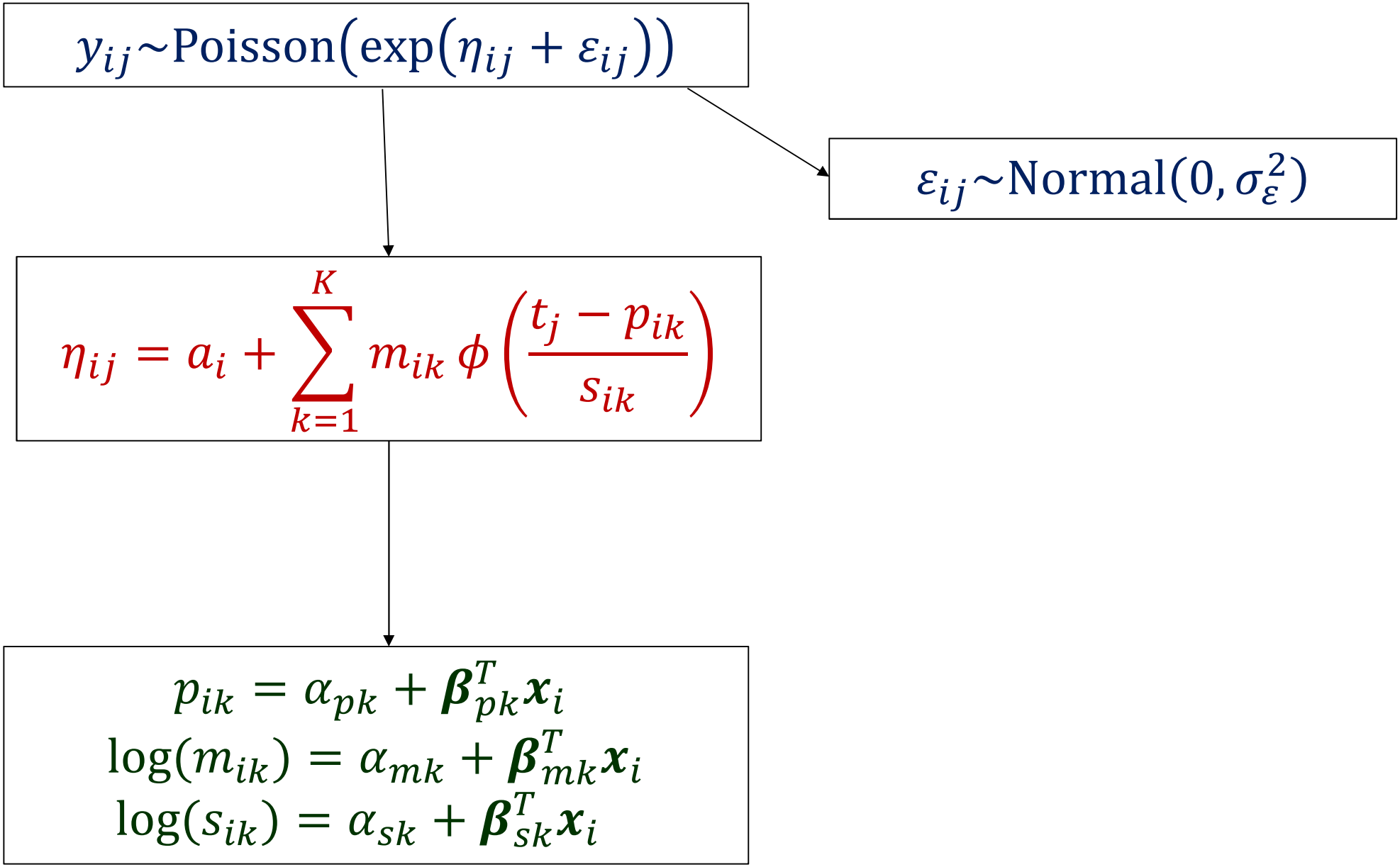
(a) have a **parametric model** for each site that describes trends of abundance in terms of seasonal phenology

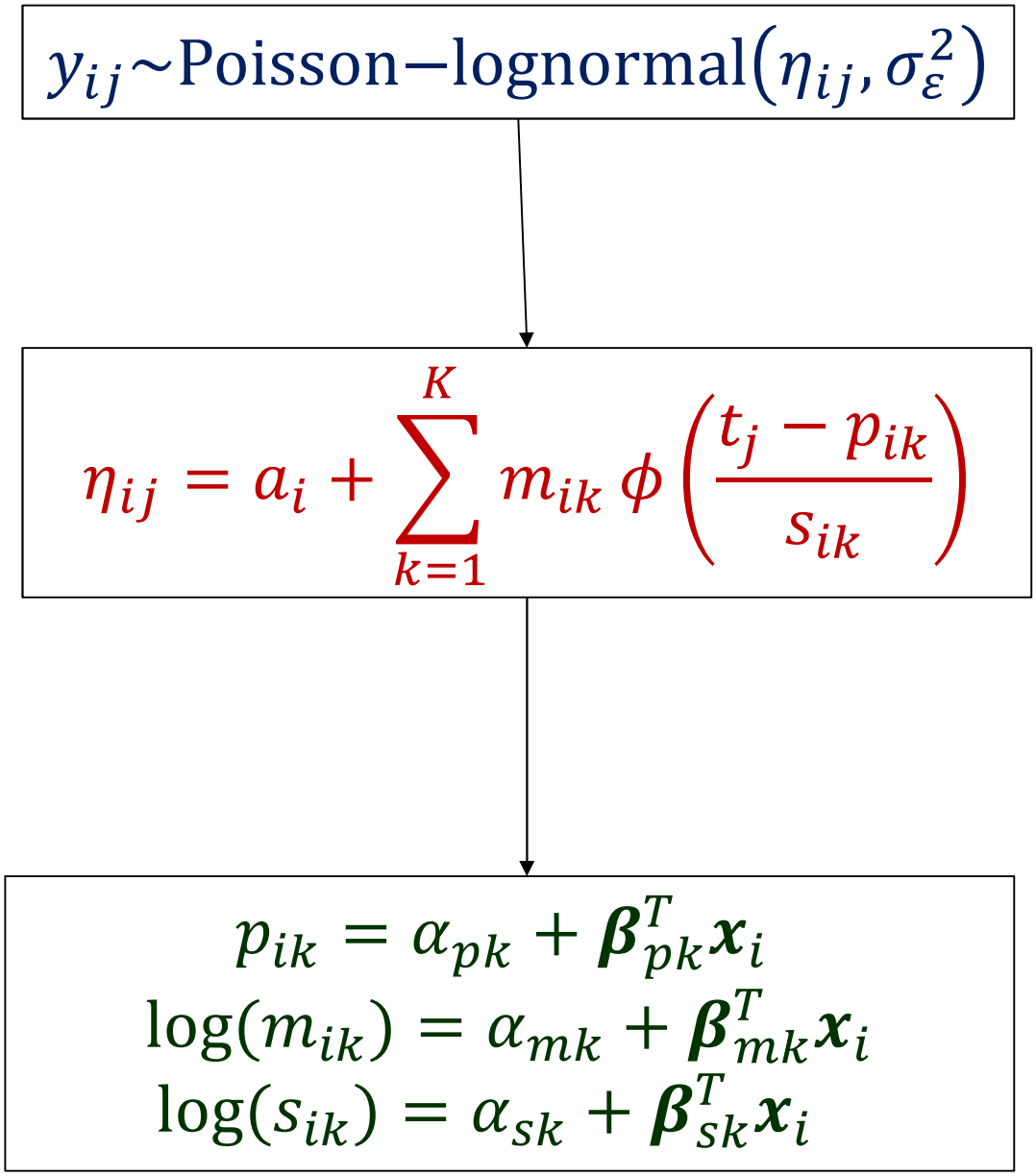
(b) combine the two parts of the modelling into a single, **joint model**

Bayesian hierarchical modelling seems to provide a natural framework for this

The most obvious approach to (a) would be to use a **harmonic model**, but this does not perform well - it fails to capture the burstiness

A **Gaussian mixture model** is more promising…

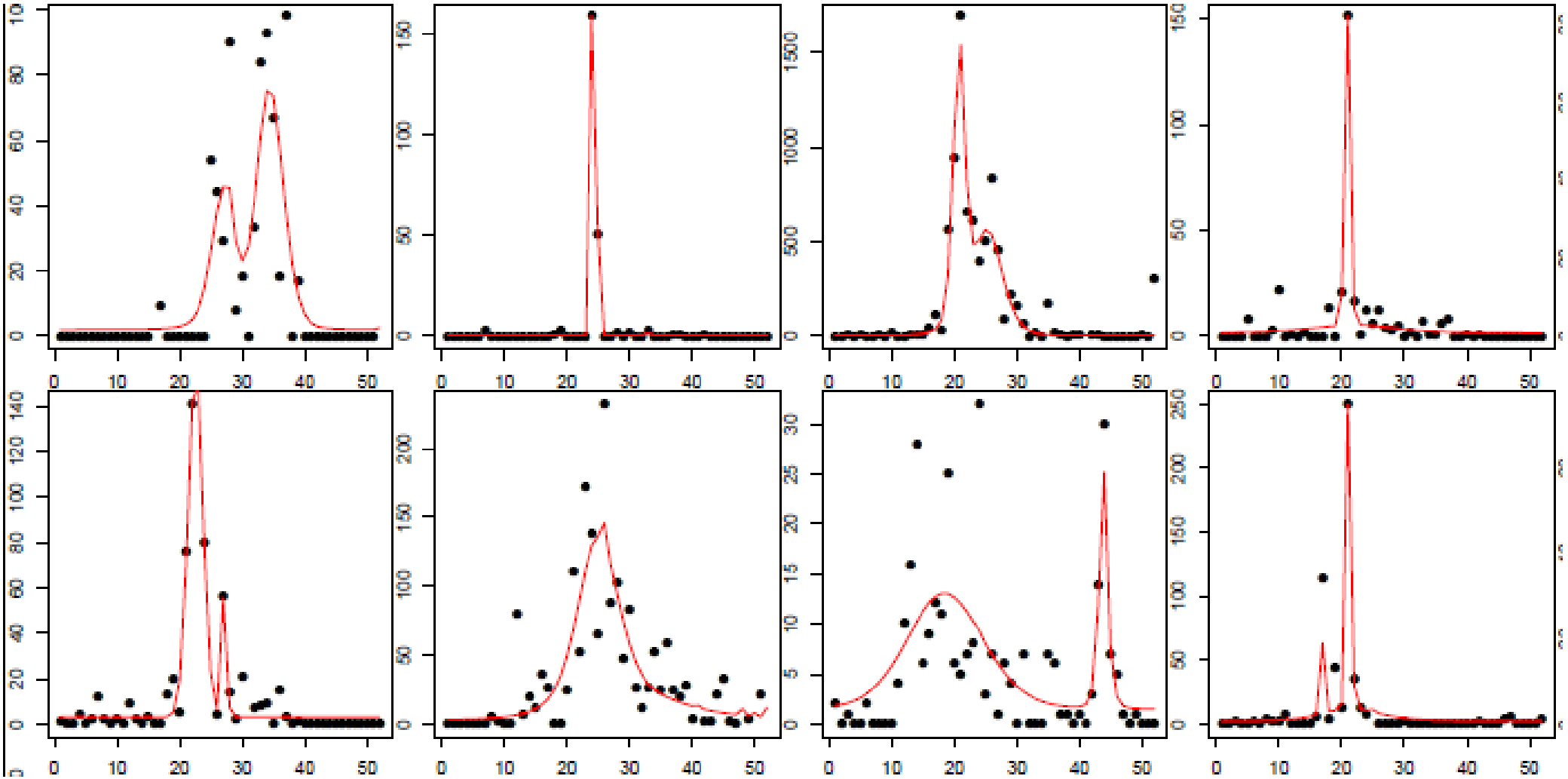

$y_{ij} \sim \text{Poisson}(\exp(\eta_{ij} + \varepsilon_{ij}))$
$\varepsilon_{ij} \sim \text{Normal}(0, \sigma_\varepsilon^2)$
$\eta_{ij} = a_i + \sum_{k=1}^{K} m_{ik}\, \phi\left(\frac{t_j - p_{ik}}{s_{ik}}\right)$
$p_{ik} = \alpha_{pk} + \boldsymbol{\beta}_{pk}^T \boldsymbol{x}_i$
$\log(m_{ik}) = \alpha_{mk} + \boldsymbol{\beta}_{mk}^T \boldsymbol{x}_i$
$\log(s_{ik}) = \alpha_{sk} + \boldsymbol{\beta}_{sk}^T \boldsymbol{x}_i$


$y_{ij} \sim \text{Poisson-lognormal}(\eta_{ij}, \sigma_\varepsilon^2)$
$\eta_{ij} = a_i + \sum_{k=1}^{K} m_{ik}\, \phi\left(\frac{t_j - p_{ik}}{s_{ik}}\right)$
$p_{ik} = \alpha_{pk} + \boldsymbol{\beta}_{pk}^T \boldsymbol{x}_i$
$\log(m_{ik}) = \alpha_{mk} + \boldsymbol{\beta}_{mk}^T \boldsymbol{x}_i$
$\log(s_{ik}) = \alpha_{sk} + \boldsymbol{\beta}_{sk}^T \boldsymbol{x}_i$

Implementation - issues

We assign diffuse priors, and attempt to fit using OpenBUGS…	…Extremely poor convergence, results are not sensible
Attempt to separate out the two parts of the model, and fit separately (with propagation of uncertainty)…	…Convergence is somewhat better, but fitted model is meaningless – predictions are essentially flat, apparently because the estimated level of OD is *extremely* high
Attempt to assign informative priors to the overdispersion variance parameter, σ_ε^2	…Posterior heavily influenced by the top of the prior distribution for the overdispersion
Replace the log link with an identity link	…Somewhat better performance…
Replace the Poisson-lognormal distribution with a normal distribution	…Much better performance…

Some results – 2 peak model



Conclusions

Gaussian Mixture models fit the abundance data moderately well, and provide a biologically interpretable way of modelling spatial variations in phenology

Implementation is problematic, and is linked to issues that arise frequently with Bayesian modelling of abundance data:

1) difficulty in fitting **joint models** simultaneously;

2) difficulty in estimating levels of **overdispersion** empirically;

3) **identity link** often seems to work better than a log link, even though it is conceptually unsatisfactorily…

Further work

Improved **model selection**, and multi-model inference

Allowing the **number of peaks** to vary between sites, and to be estimated empirically for each site

Incorporating **time-varying covariates** (whose value changes over time *within* each year)

Case Study 2:
Modelling the spread of invasive species

Species atlas data

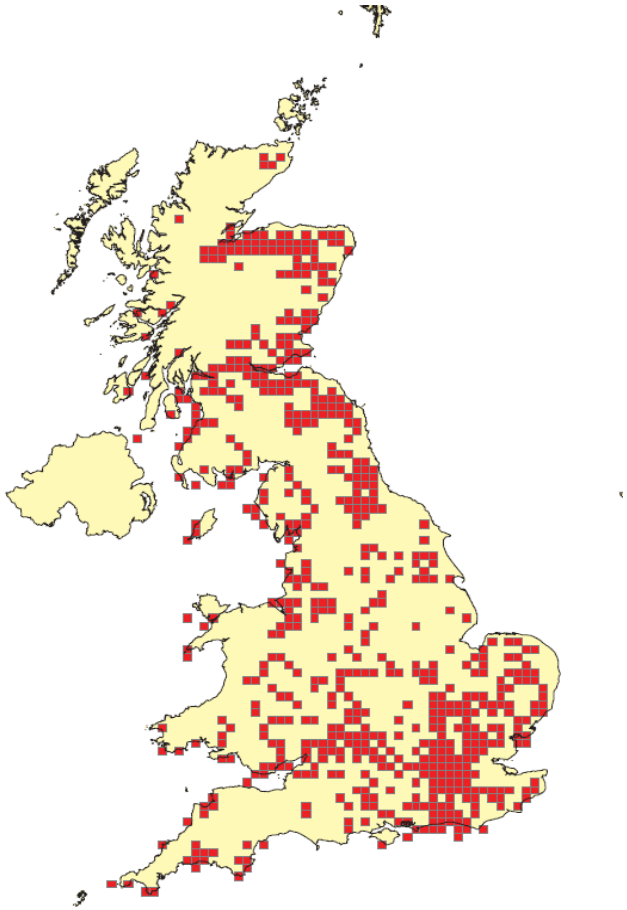
Species atlas data…

…are widely used in ecology
…are available for many taxa within
countries such as Germany, the UK,
Slovenia & the Czech Republic, and
for some taxa over much wider areas

The data that are usually available consist of the presence/absence of a species within cells of a regular grid

Atlases typically refer to a given year/decade

The raw data consist of **presence records**: sightings of a species at a specific location



Species atlas data are usually used to model the **static** spatial distribution of a species

This is often extended to modelling the species at a small set of discrete timepoints ("**snapshots**")

But can we go further, and use them to model the **spread** of a species in space and time?

This is primarily of interest for **invasive species**, because

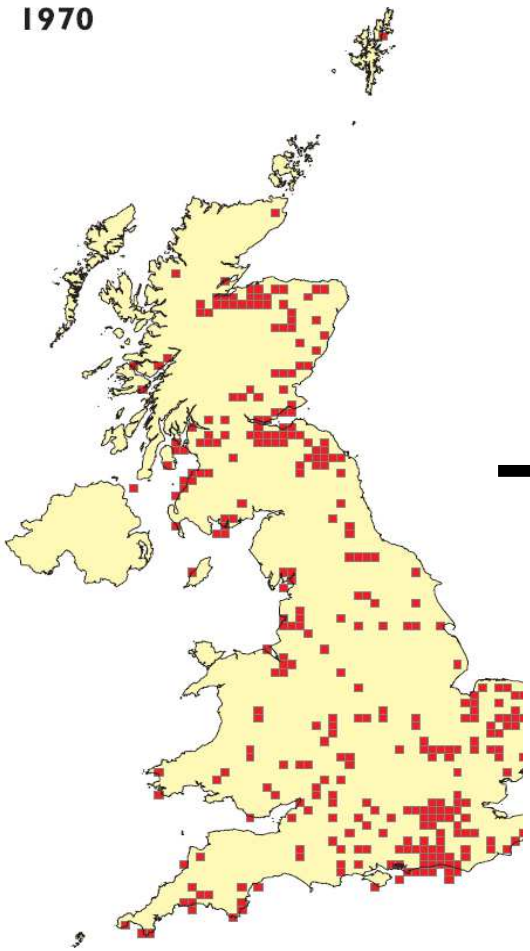
a) their distributions are often changing rapidly;

b) there is interest in predicting their future distribution

Application: the spread of giant hogweed across Great Britain

- **Giant hogweed** (*Heracleum Mantegazzianum*)
- A highly invasive species that is non-native within Great Britain, which is difficult to remove, and which poses risks to human health
- Initially introduced into botanic gardens (1893), but escaped early in the 20th century and now widespread across large parts of GB
- Interest is in understanding how the species has spread, and calculating the probability that it will colonize the remaining parts of GB

Data



z_{i1}

Grid cells:

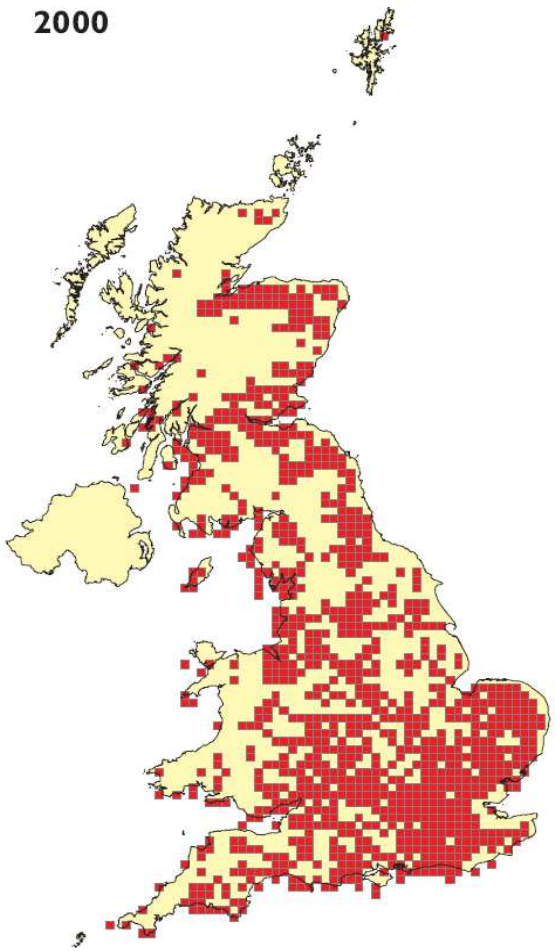
$i = 1, \ldots, n$

Covariates:

x_i

Distances:

d_{ij}



z_{i2}

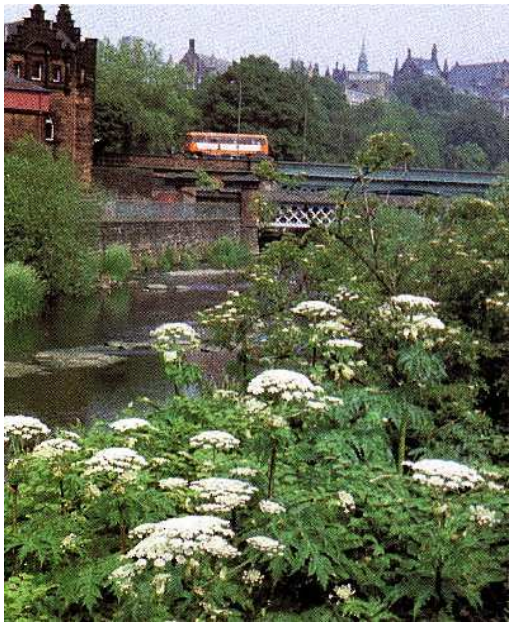
Modelling assumptions

The species is very difficult to **eradicate**

It is therefore reasonable to assume that, especially for data on a relatively coarse grid, it can be described by an **SI** (susceptible-infected) model

i.e. once a grid cell has been colonised, it cannot be decolonised

The covariates $\boldsymbol{x}_i$ include temperature, altitude and habitat proportions



http://www.gla.ac.uk/ibls/DEEB/jd/apfg.html

Model – species distribution models

- Could potentially model using a **species distribution model** for the second time period, with the distribution in the first time period as a covariate

$$z_{i2} \sim \text{Binomial}\left((p_i)^{z_{i1}}\right)$$

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta^T \boldsymbol{x}_i + \log(\omega_i)$$

$$\omega_i = \sum_{j=1}^{n} h_{ij}(\lambda) z_{j1}$$

Model – species distribution models

- Could potentially model using a **species distribution model** for the second time period, with the distribution in the first time period as a covariate

$$z_{i2} \sim \text{Binomial}\left((p_i)^{z_{i1}}\right)$$

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta^T \boldsymbol{x}_i + \log(\omega_i)$$

$$\omega_i = \sum_{j=1}^{n} h_{ij}(\lambda) z_{j1}$$

Spatial kernel: $h_{ij}(\lambda) = d_{ij}^{-2\lambda}$

Issues

Spatial and temporal effects are confounded within the ω_i term

This makes biological interpretation of the parameters difficult, as they are fundamentally linked to the time period that separates the atlases

Could avoid this by formulating model at higher frequency in time than the data, and treating values of z_{it} for intermediate times as **latent**

…or could go further, and formulate as a **continuous time** model e.g. a **discrete state-space continuous time Markov process**

General formulation

State of system: $\underline{z}(t)$ at time t

q **event types** $\{e_i : i = 1,\ldots,q\}$

Event type e_i induces a **change** δ in the system state: $\underline{z}(t) \rightarrow \underline{z}(t) + \underline{\delta}$

$r_i(t)$: function determining the **rate** at which event e_i occurs

State of system: $\underline{z}(t)$ at time t

q **event types** $\{e_i : i = 1,\ldots,q\}$

Event type e_i induces a **change** δ in the system state: $\underline{z}(t) \rightarrow \underline{z}(t) + \underline{\delta}$

$r_i(t)$: function determining the **rate** at which event e_i occurs

State of system: $\underline{z}(t)$ at time t ← **Presence/absence map**

q **event types** $\{e_i : i = 1,\ldots,q\}$

Event type e_i induces a **change** δ in the system state: $\underline{z}(t) \rightarrow \underline{z}(t) + \underline{\delta}$

$r_i(t)$: function determining the **rate** at which event e_i occurs

State of system: $\underline{z}(t)$ at time t

q **event types** $\{e_i : i = 1,\ldots,q\}$

Event type e_i induces a **change** δ in the system state: $\underline{z}(t) \rightarrow \underline{z}(t) + \underline{\delta}$

$r_i(t)$: function determining the **rate** at which event e_i occurs

State of system: $\underline{z}(t)$ at time t

***q* event types** $\{e_i : i = 1,\ldots,q\}$ ← **cell *i* becomes colonized**

Event type e_i induces a **change** δ in the system state: $\underline{z}(t) \rightarrow \underline{z}(t) + \underline{\delta}$

$r_i(t)$: function determining the **rate** at which event e_i occurs

State of system: $\underline{z}(t)$ at time t

q **event types** $\{e_i : i = 1,\ldots,q\}$

Event type e_i induces a **change** δ in the system state: $\underline{z}(t) \rightarrow \underline{z}(t) + \underline{\delta}$

$r_i(t)$: function determining the **rate** at which event e_i occurs

State of system: $\underline{z}(t)$ at time t

q **event types** $\{e_i : i = 1,\ldots,q\}$

Event type e_i induces a **change** δ in the system state: $\underline{z}(t) \rightarrow \underline{z}(t) + \underline{\delta}$

$r_i(t)$: function determining the **rate** at which event e_i occurs

$\underline{\delta}$ = (0,…,0,1,0,…,0): absence becomes presence

State of system: $\underline{z}(t)$ at time t

q **event types** $\{e_i : i = 1,\ldots,q\}$

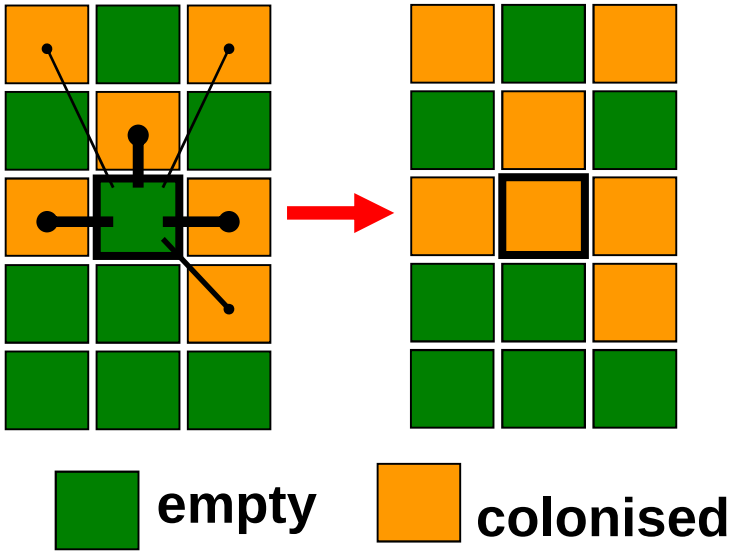
Event type e_i induces a **change** δ in the system state: $\underline{z}(t) \rightarrow \underline{z}(t) + \underline{\delta}$

$r_i(t)$: function determining the **rate** at which event e_i occurs

Parametric model for the rate function

Sequence of colonisations of empty grid cells

Colonisation rate of uncolonised location:

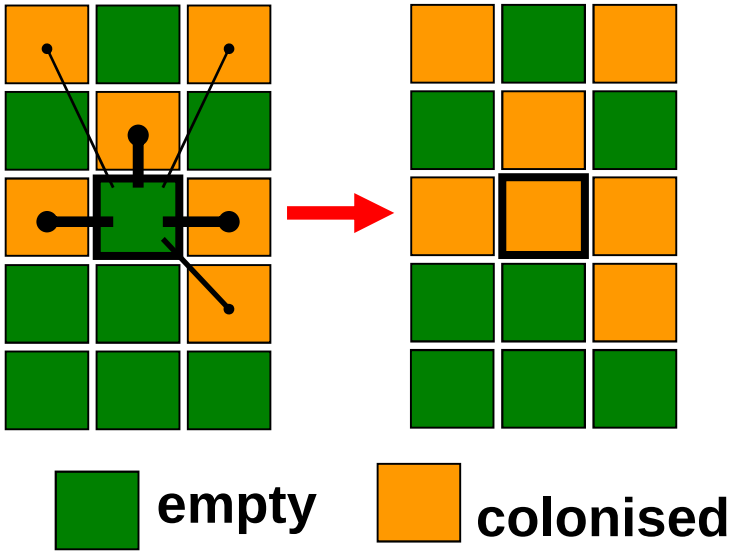


$$\log\left(r_i(t)\right) = \alpha + \boldsymbol{\beta}^T \boldsymbol{x}_i + \sum_{j=1}^{n} h_{ij}(\lambda) z_{jt}$$

Parametric model for the rate function

Sequence of colonisations of empty grid cells

Colonisation rate of uncolonised location:



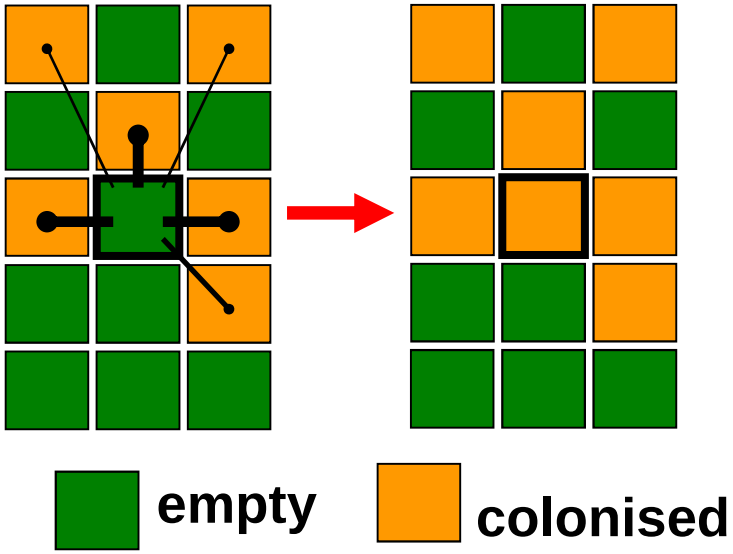
$$\log\left(r_i(t)\right) = \alpha + \boldsymbol{\beta}^T \boldsymbol{x}_i + \sum_{j=1}^{n} h_{ij}(\lambda) z_{jt}$$

Dispersal rate depends on number of colonised sites and distances between them & the target site

Parametric model for the rate function

Sequence of colonisations of empty grid cells

Colonisation rate of uncolonised location:



$$\log\left(r_i(t)\right) = \alpha + \boldsymbol{\beta}^T \boldsymbol{x}_i + \sum_{j=1}^{n} h_{ij}(\lambda) z_{jt}$$

Suitability depends on environmental covariates

A novel aspect of our model, relative to other similar continuous time models, is that it allows for **inhomogeneous suitability**

i.e. replaces α with $\alpha + \boldsymbol{\beta}^T \boldsymbol{x}_i$

Inference

The model parameters are $(\alpha, \boldsymbol{\beta}, \lambda)$

The complete **likelihood** is:

$$\prod_{k=1}^{n} r_k(t_{k-1}) \exp\{-(t_k - t_{k-1}) R_k(t_{k-1})\}$$

where t_i is the colonisation time for site i

In practice the **state space** $\underline{z}(t)$ is known for some timepoints and unknown for others

MCMC is used to infer the joint distribution of (θ,ε)

Algorithm is based on Metropolis-Hastings, with a sequential update of each component of the parameter vector using a normal proposal

Diffuse priors assumed for all priors – event times are assumed to have a uniform distribution

Results

Biased estimate of dispersal without covariates

- longer range dispersal estimated with covariates

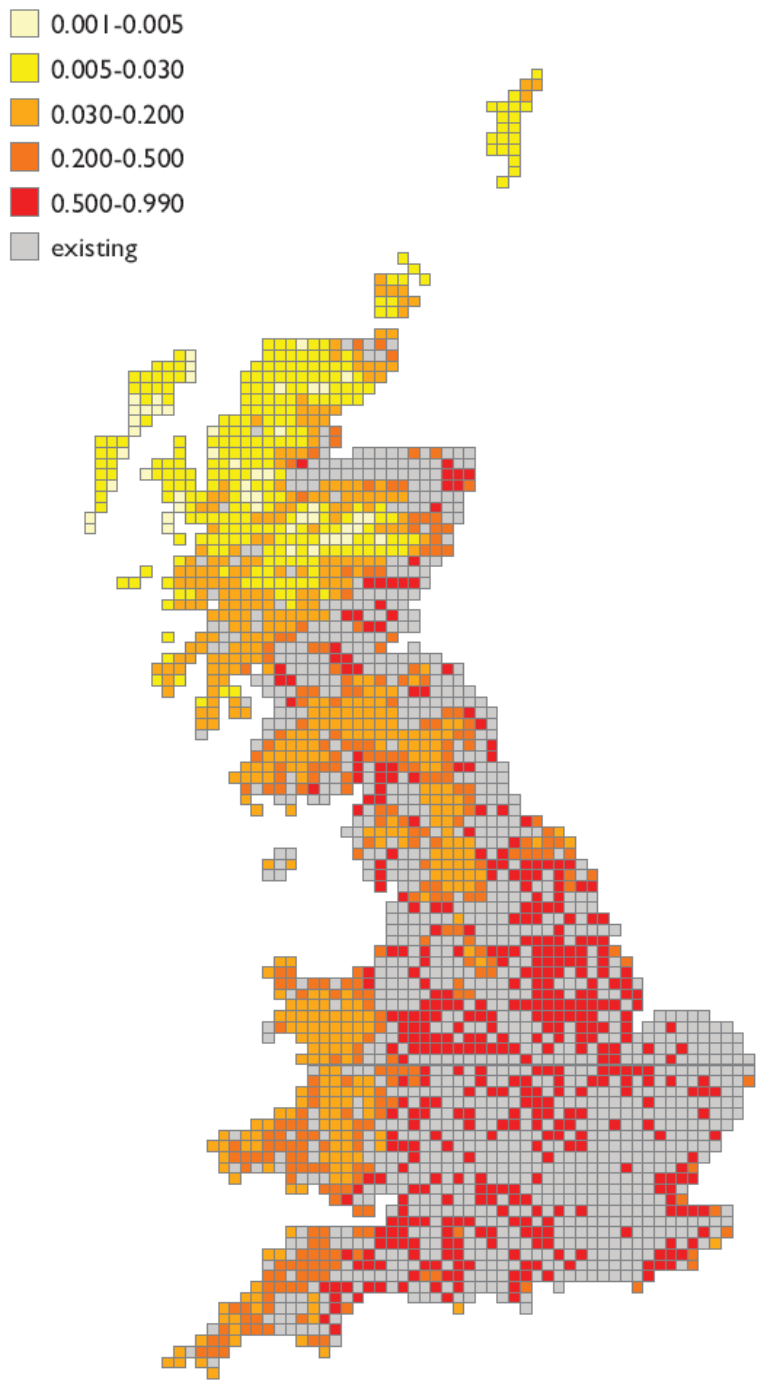
Also biased estimates of covariate effects when ignore dispersal

Estimated suitability for each 10x10km^2 square

(posterior mean)

Probability of colonisation by 2050, assuming no change in climate or land use

(posterior mean)



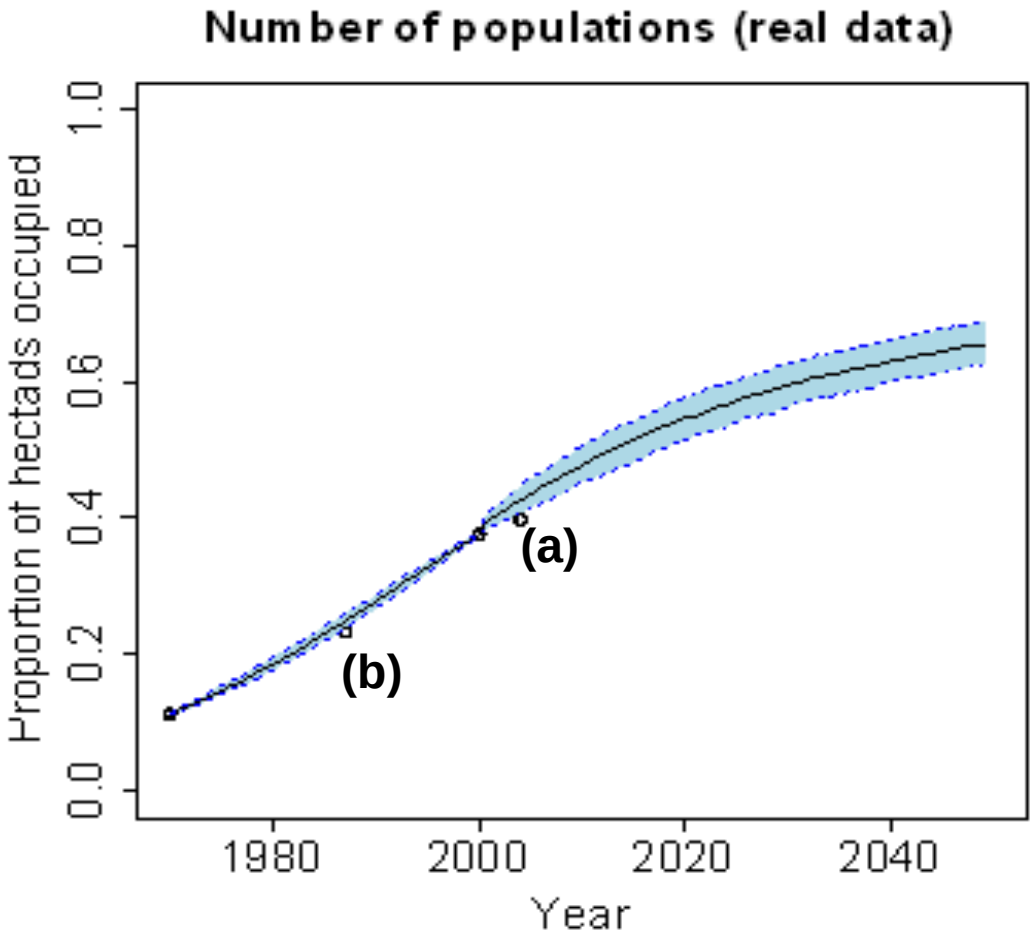
Evaluating model performance

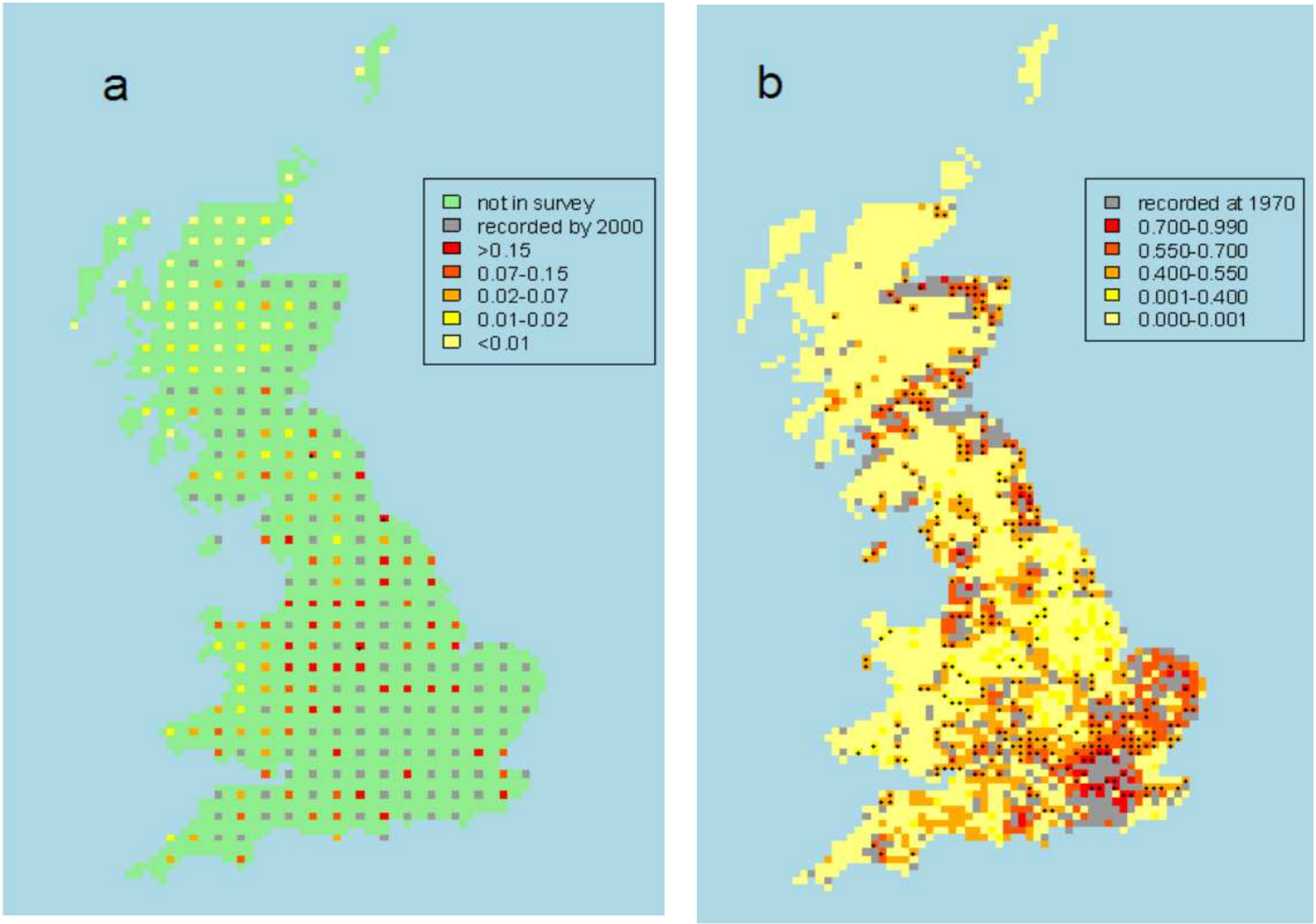
We fit the model to data for two snapshots (1970 and 2000)

We evaluate performance using two independent data sources:

a) Limited survey of one-in-nine 10x10km2 squares in 2004

b) 1987 atlas (which is known to have underreporting bias)




a
not in survey
recorded by 2000
>0.15
0.07-0.15
0.02-0.07
0.01-0.02
<0.01
b
recorded at 1970
0.700-0.990
0.550-0.700
0.400-0.550
0.001-0.400
0.000-0.001

Simulation study

Parameter	Estimate and 95% credible interval	Proportion of credible intervals containing original parameter value
λ (dispersal)	1.0 (0.9, 1.2)	93%
α (altitude)	-0.0053 (-0.0071, -0.0034)	82%
τ (temperature)	-0.24 (-0.37, -0.09)	96%
Suitability parameters:		
β_2 (coast)	0.30 (0.10, 0.54)	98%
β_3 (arable)	0.08 (0.05, 0.12)	97%
β_4 (broadleaf forest)	0.72 (0.37, 1.17)	94%
β_5 (urban)	0.41 (0.24, 0.61)	95%
β_6 (conifer forest)	0.04 (0.00, 0.14)	96%
β_7 (improved grassland)	0.16 (0.09, 0.25)	93%
β_8 (open water)	0.15 (0.00, 0.53)	93%
β_9 (semi-natural)	0.09 (0.01, 0.20)	95%
β_{10} (upland)	0.01 (0.00, 0.05)	98%

Reliability of inference assessed using simulated data generated from posterior mean estimates of parameter values for giant hogweed

Simulation study

Parameter	Estimate and 95% credible interval	Proportion of credible intervals containing original parameter value
λ (dispersal)	1.0 (0.9, 1.2)	93%
α (altitude)	-0.0053 (-0.0071, -0.0034)	82%
τ (temperature)	-0.24 (-0.37, -0.09)	96%
Suitability parameters:		
β_2 (coast)	0.30 (0.10, 0.54)	98%
β_3 (arable)	0.08 (0.05, 0.12)	97%
β_4 (broadleaf forest)	0.72 (0.37, 1.17)	94%
β_5 (urban)	0.41 (0.24, 0.61)	95%
β_6 (conifer forest)	0.04 (0.00, 0.14)	96%
β_7 (improved grassland)	0.16 (0.09, 0.25)	93%
β_8 (open water)	0.15 (0.00, 0.53)	93%
β_9 (semi-natural)	0.09 (0.01, 0.20)	95%
β_{10} (upland)	0.01 (0.00, 0.05)	98%

Reliability of inference assessed using simulated data generated from posterior mean estimates of parameter values for giant hogweed

Summary

Modelling framework allows us to use species atlas data to draw inferences about **spatio-temporal spread**

Based on **SI models** from epidemiology, but extended to allow for environmental heterogeneity

Current work: comparison against **species distribution modelling** approach

Interested in extension to species that have **extinction** & colonization (using SIS rather than SI model)

Approach could be further extended to allow for **non-detection**

Case Study 3: Uptake of greenhouse gas mitigation measures in agriculture

Context

Agriculture is an important source of **greenhouse gas emissions** – particularly **N_2O**

Various measures have been developed or proposed that could potentially reduce emissions

We are interested here in the following five measures:

Improved slurry application

Land drainage

Improved timing of nitrogen (N) application

Avoiding nitrogen (N) excess

Nitrification inhibitors

To what extent to farmers currently adopt these measures, and how would this rate (the "**adoption rate**") change if different policy tools were used?

Possible **policy tools** include: education, financial incentives, regulation

Empirical data on the current adoption rate are inadequate, and data on the rates that would be obtained under different policy scenarios are clearly unavailable

The use of **expert opinion** is therefore essential…

Elicitation exercise

In each case we are interested in eliciting the adoption rates of these five measures under each of three scenarios:

Current (existing level of uptake)
Year 2023 – no change to current policy
Year 2023 – policies to actively promote mitigation

We are therefore interested in eliciting the distribution of 15 quantities

We elicit three values for each – the 25%, 50% and 75% quantiles

Process

The elicitation took the form of a **one-day workshop**

Involved **six agricultural experts** (agricultural consultants, managers of research farms, agricultural researchers)

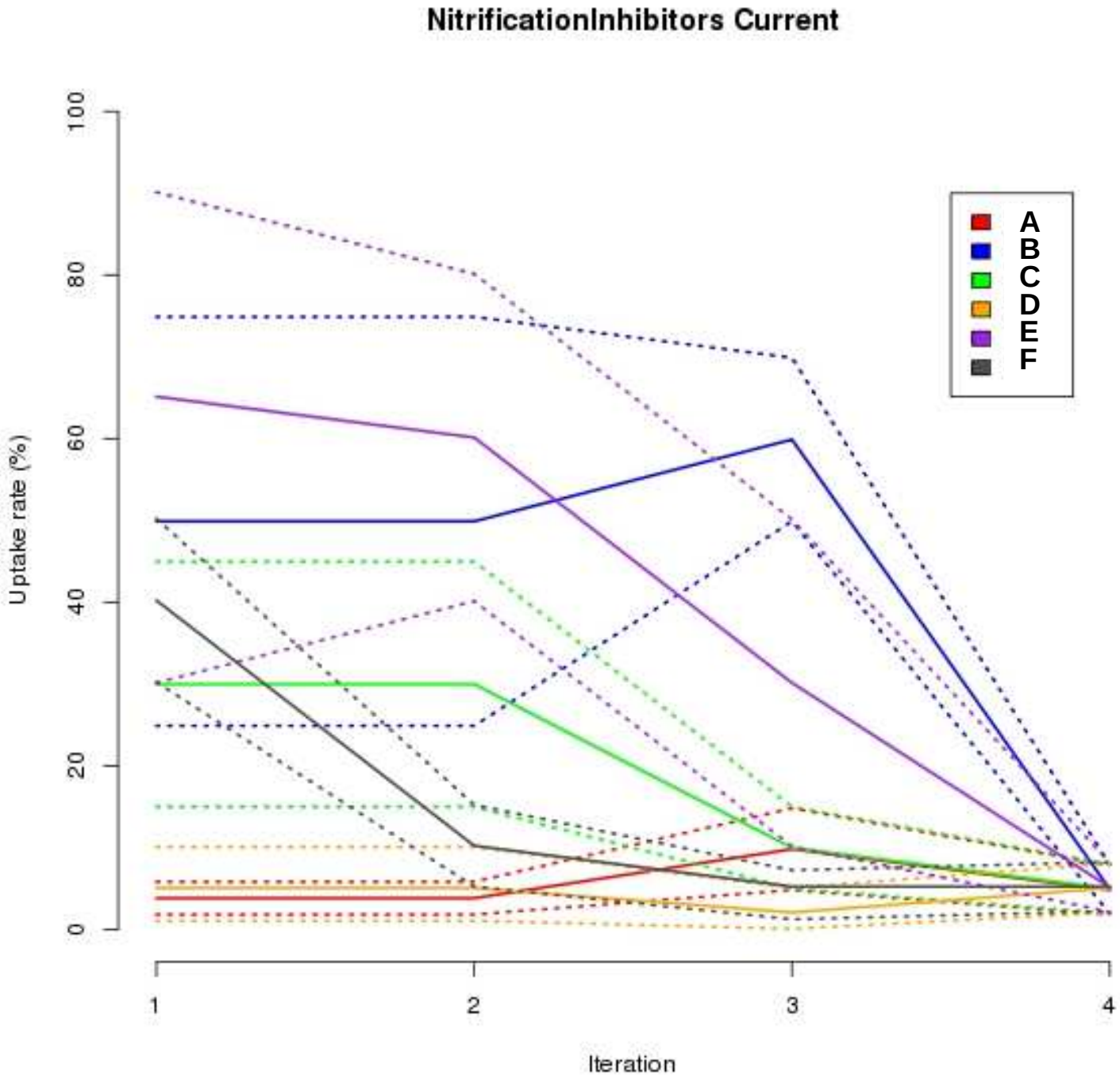
Also involved **five facilitators** (two researchers working on agricultural GG emissions, two socio-economists, one statistician)

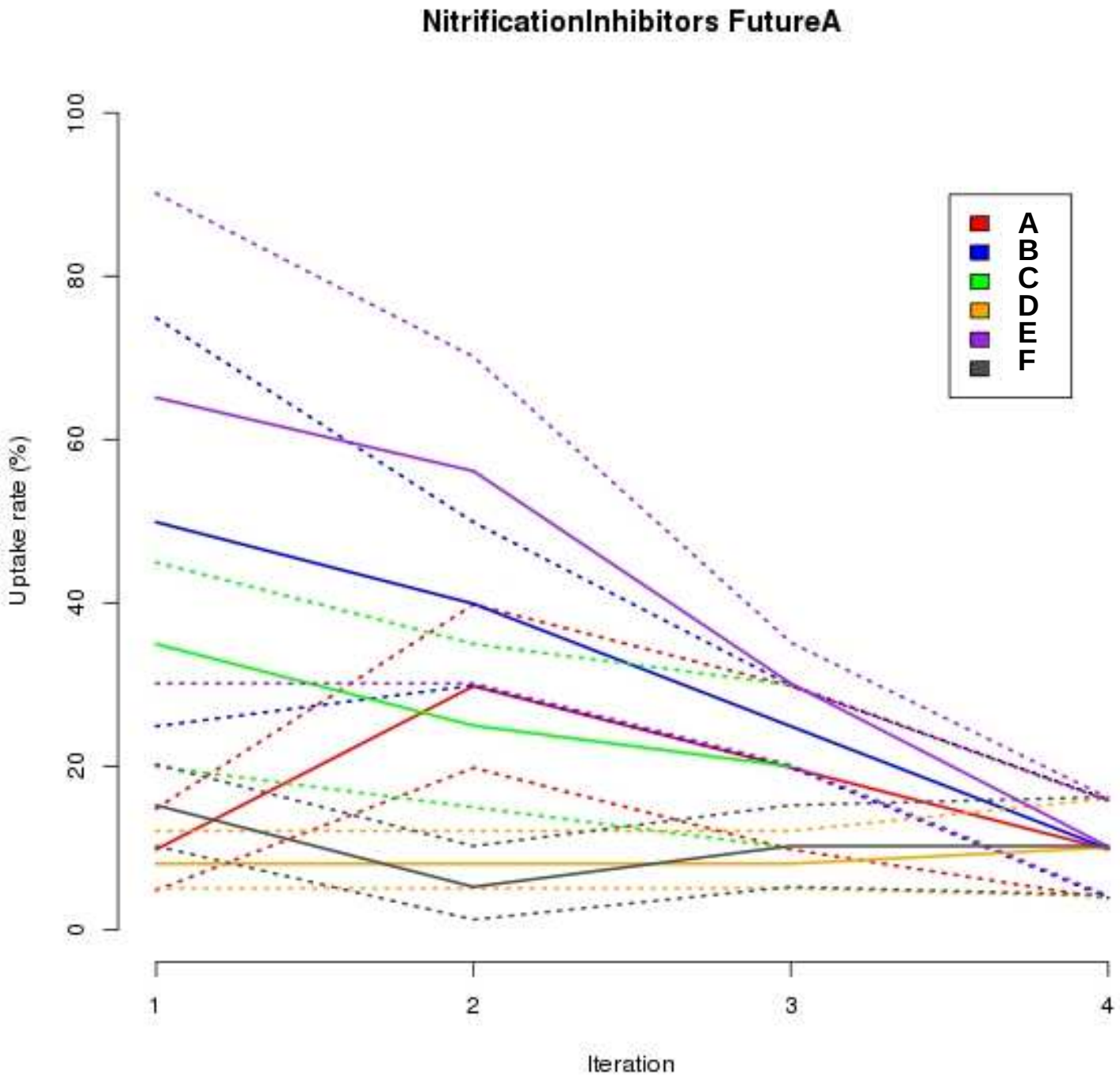
Initial briefings (on the scenarios, on the definition of quantiles, and on the elicitation process), following by individual elicitations

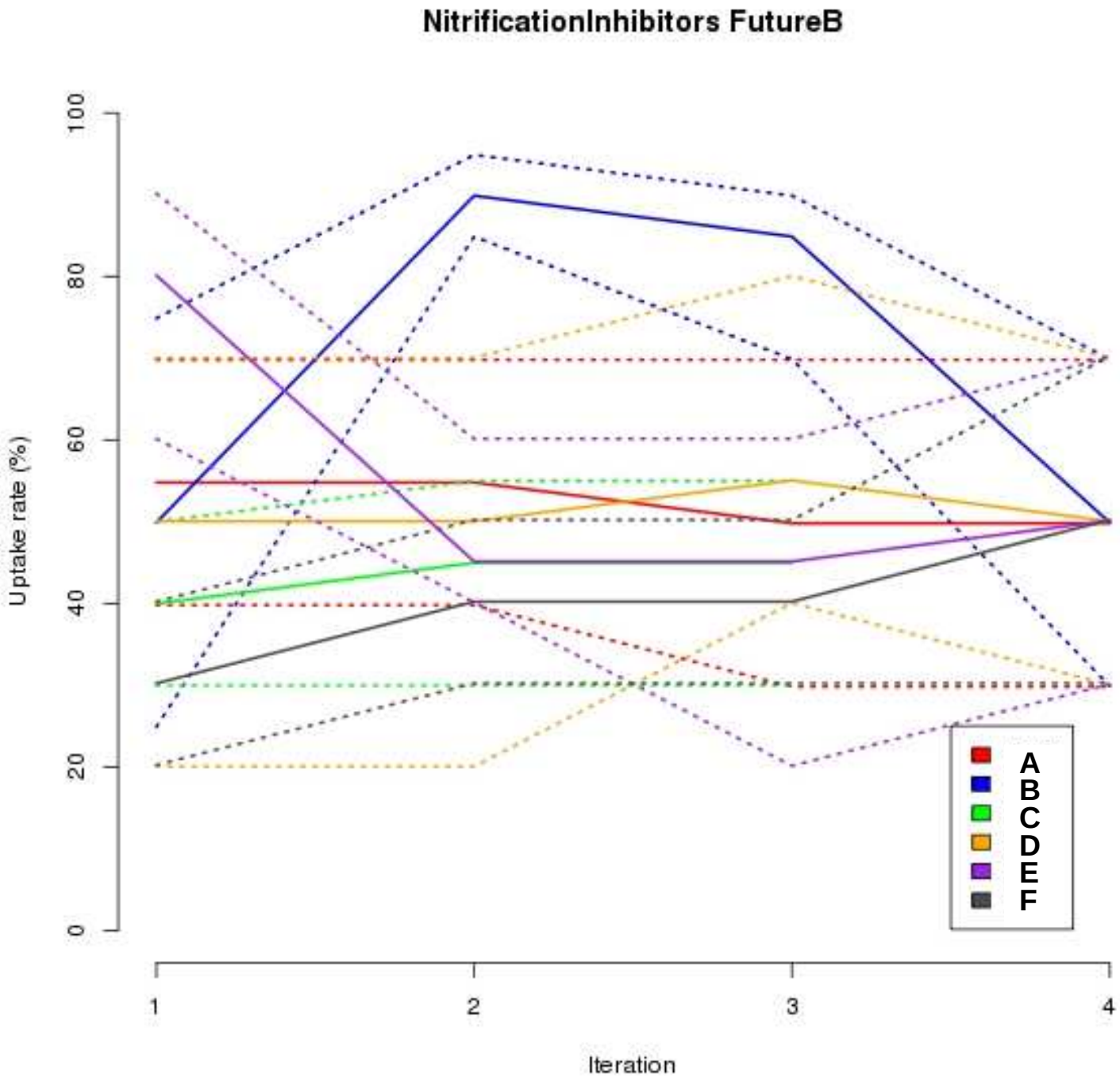
Disclosure of results, followed by discussion and revised assessments – this process was repeated three times

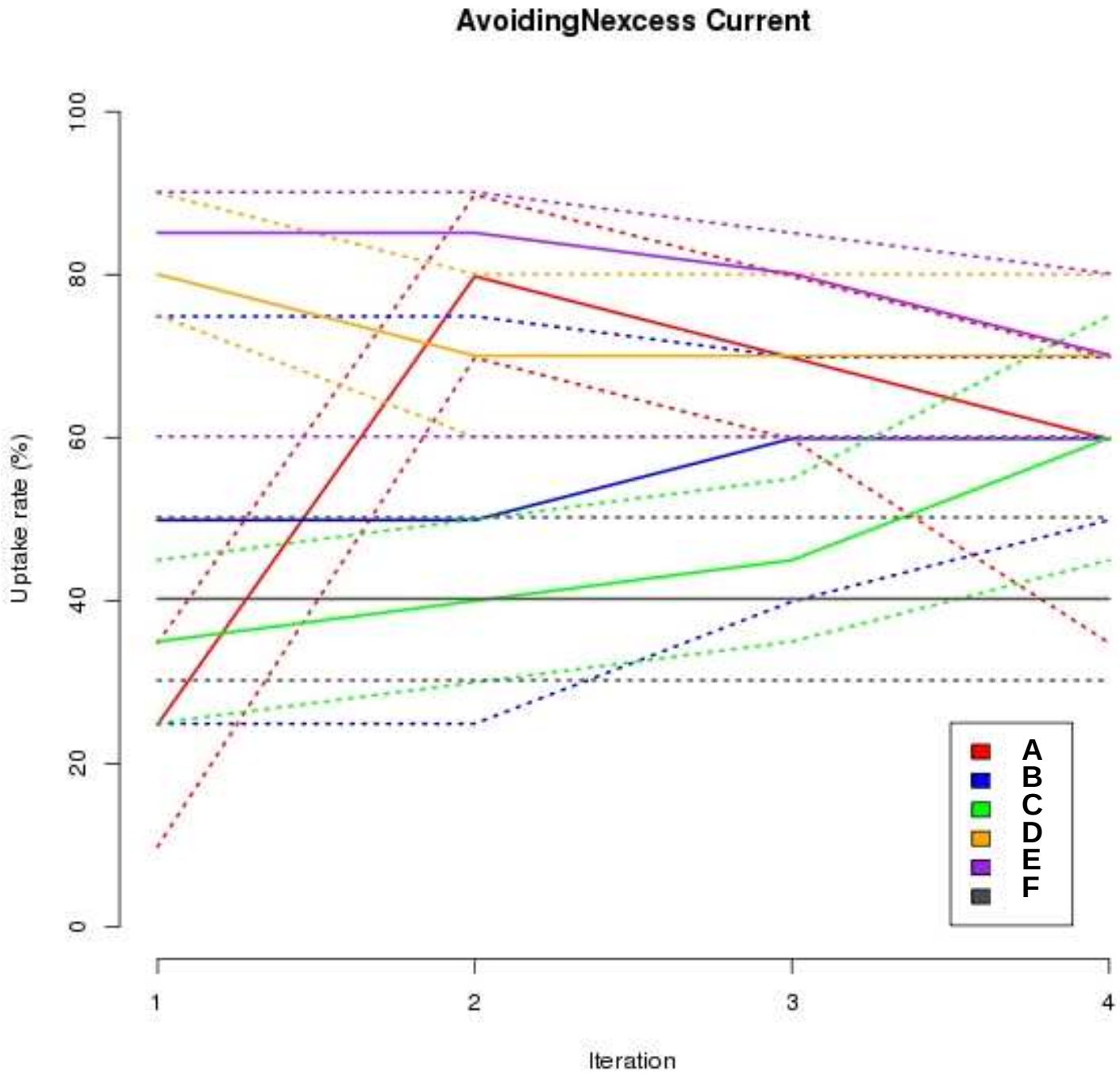
At the final stage the participants attempted to agree on a single distribution

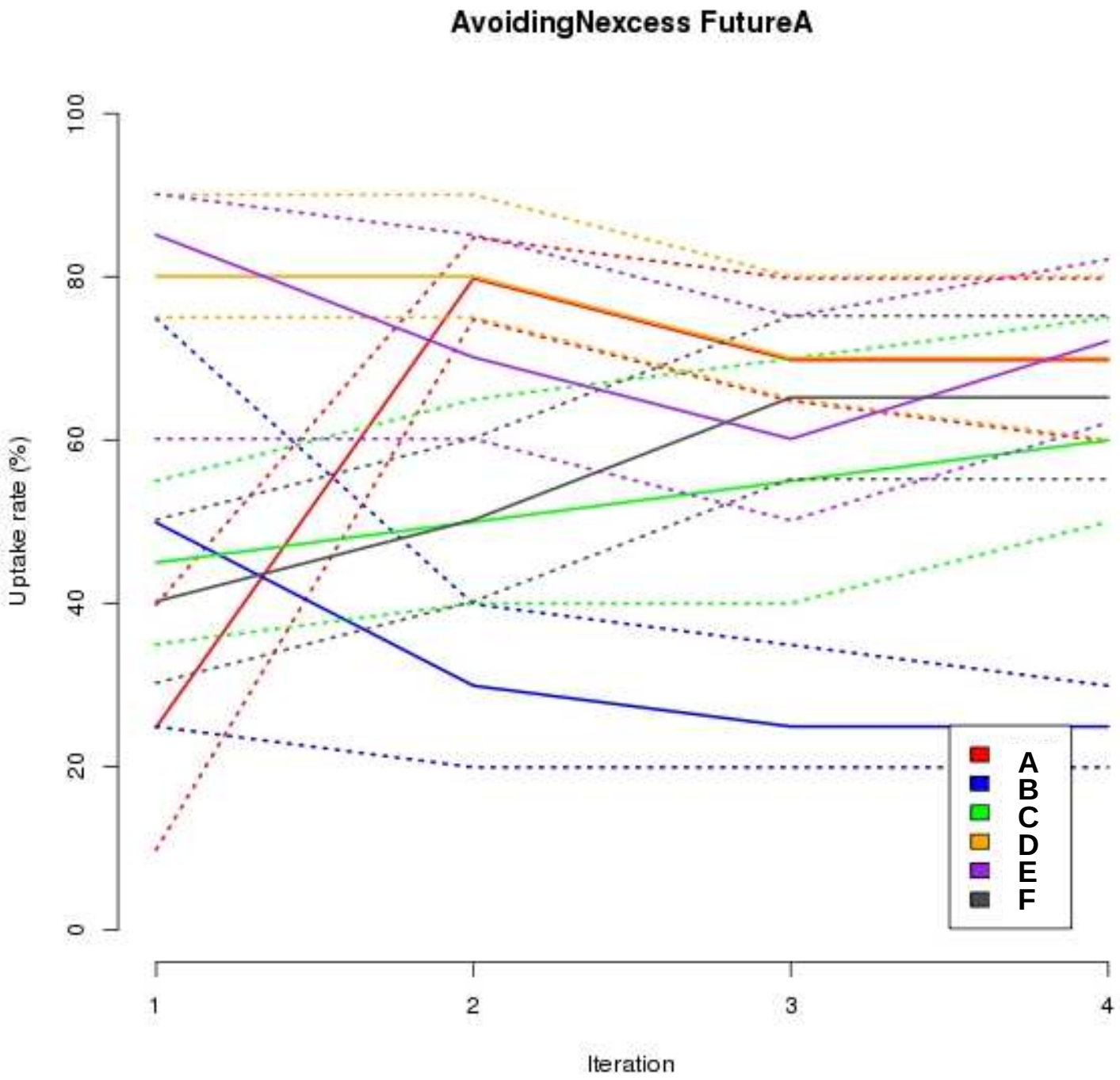
Results

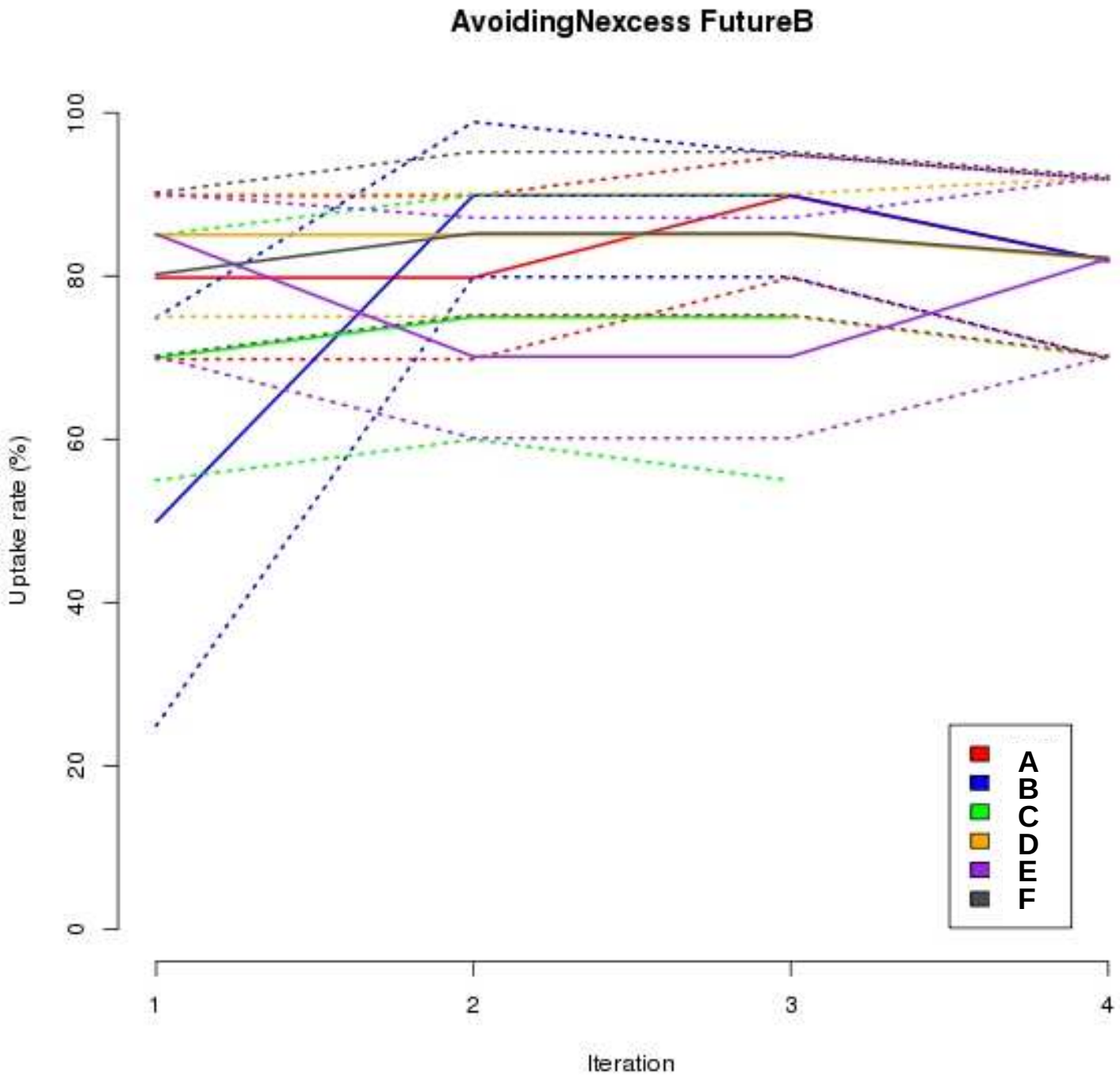

NitrificationInhibitors Current
Uptake rate (%)
0
20
40
60
80
100
1
2
3
4
Iteration
A
B
C
D
E
F


NitrificationInhibitors FutureA
Uptake rate (%)
0
20
40
60
80
100
1
2
3
4
Iteration
A
B
C
D
E
F


NitrificationInhibitors FutureB
Uptake rate (%)
0
20
40
60
80
100
1
2
3
4
Iteration
A
B
C
D
E
F


AvoidingNexcess Current
Uptake rate (%)
0
20
40
60
80
100
1
2
3
4
Iteration
A
B
C
D
E
F


AvoidingNexcess FutureA
Uptake rate (%)
0
20
40
60
80
100
1
2
3
4
Iteration
A
B
C
D
E
F


AvoidingNexcess FutureB
Uptake rate (%)
0
20
40
60
80
100
1
2
3
4
Iteration
A
B
C
D
E
F

Lessons learnt

The participants took the exercise very seriously, and the discussions were detailed and intense

A lot of the disagreements concerned the exact definition of the measures themselves – especially “land drainage”

The participants seemed to understand the concept of quantiles, but there was some ambiguity as to whether the probability distributions related to *variability* or *uncertainty*

Results provide the first meaningful assessment of the impact of government policy upon the adoption of GG mitigation measures by farmers

Concluding thoughts

Lots of potential to use Bayesian methods in ecology and the environmental sciences…

…primary motivation is to allow a richer class of models to be fitted

…but also situations where it is useful to bring in external information via the prior