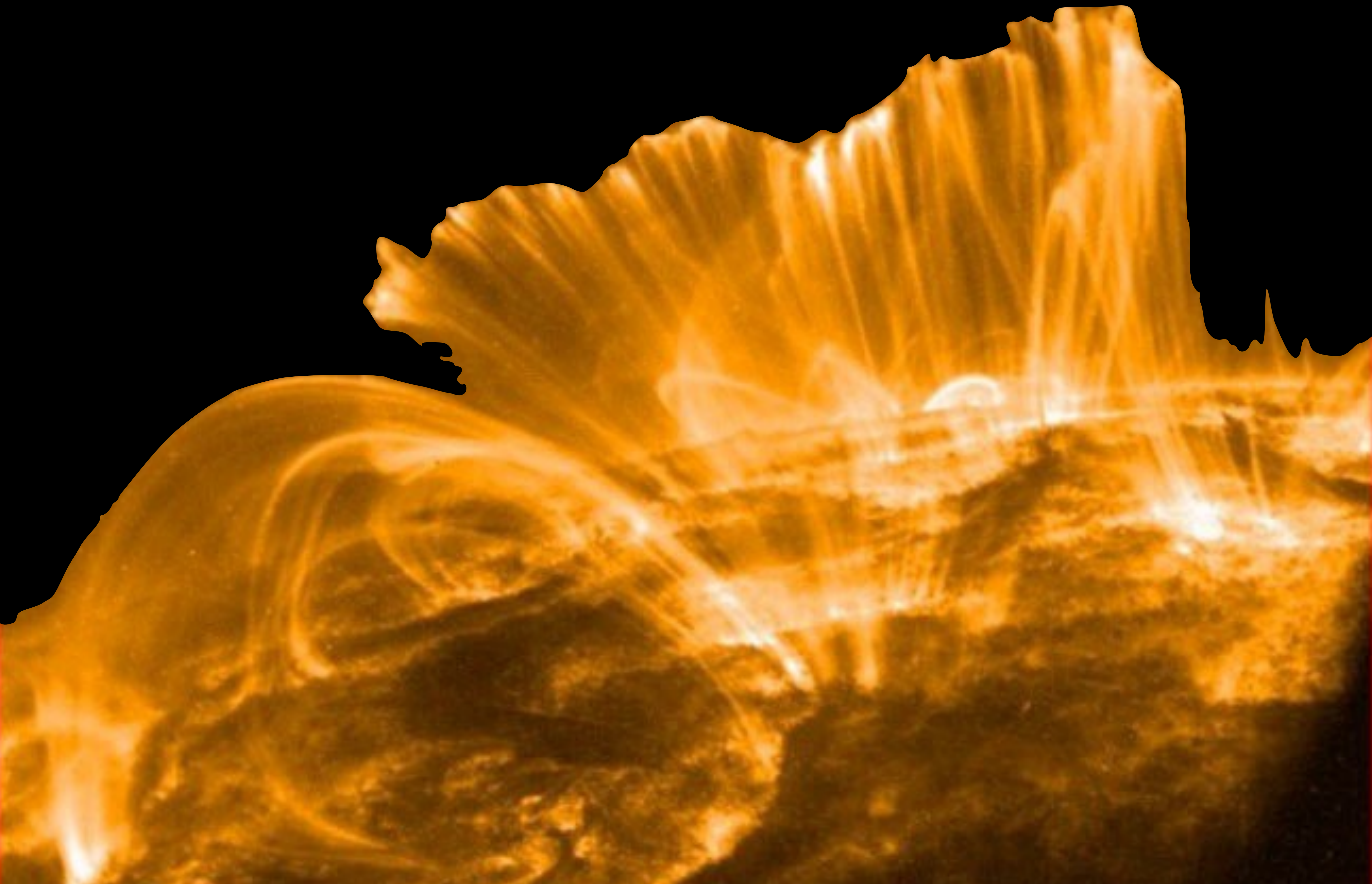
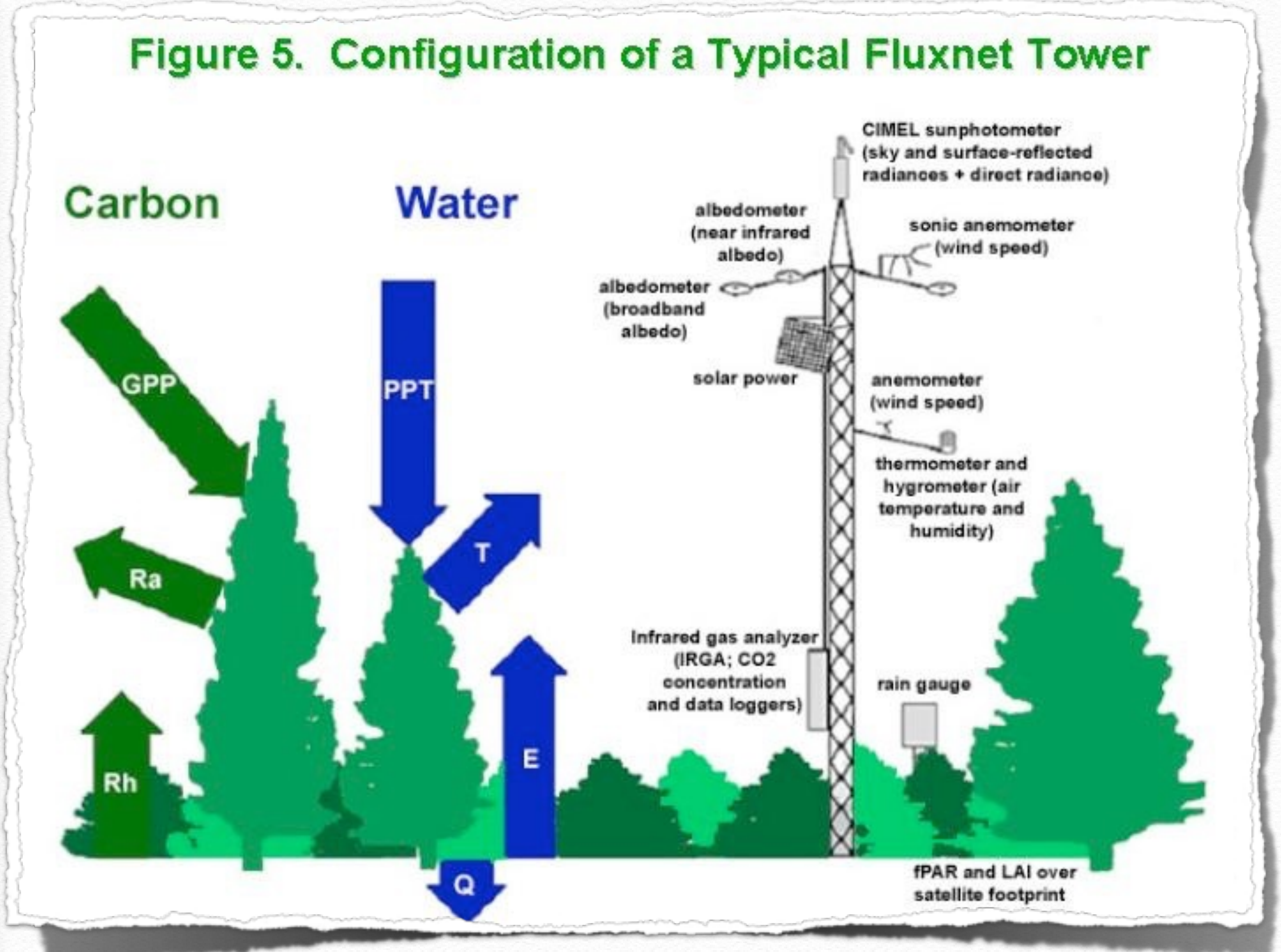
FUSING DATA



Fusion = Synthesis

- Essential complement to reduction
- No single data set provides a complete picture
- Involves more than just concatenating the files together (data harmonization)
- Observed at different scales, uncertainties
- Naïve approaches drop uncertainties, covariances

Identifiability



Tower: NEE, ET

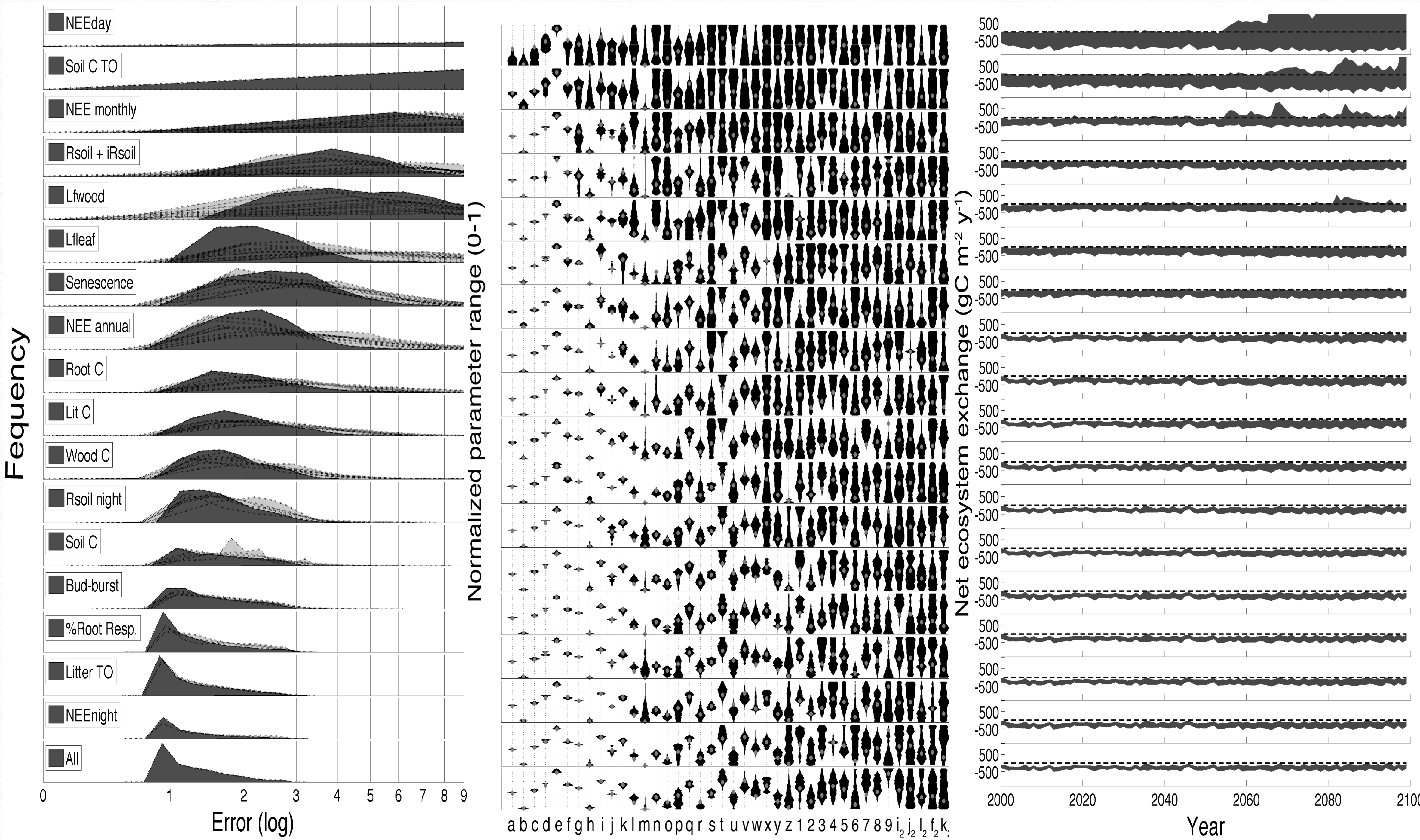
NPP: Inventory, leaf litter, root turnover, CWD

Soil Resp

Sapflux: T

Gas Exchange: GPP, Ra

Rate my data

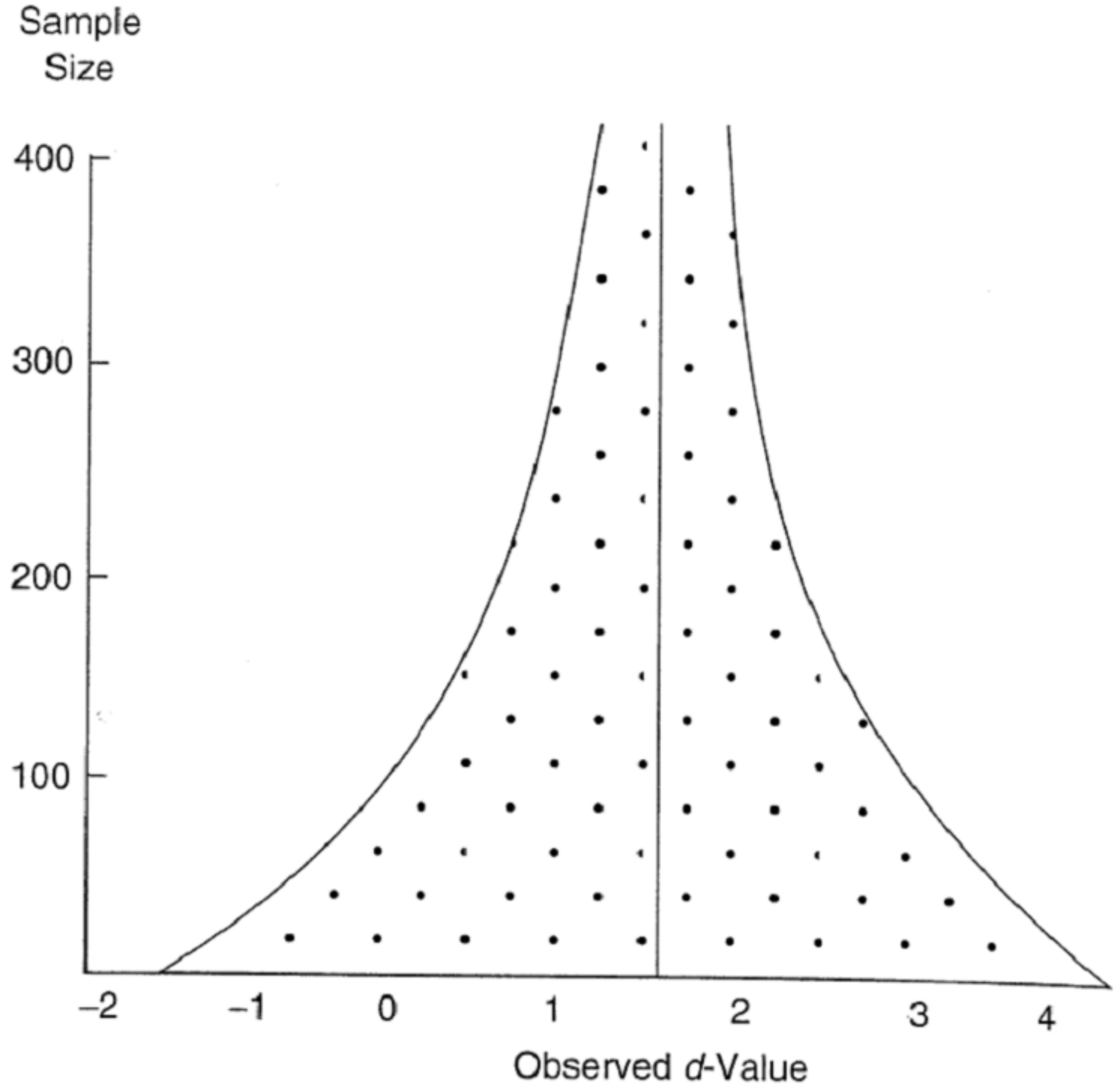


Keenan et al. (2013) Ecological Applications

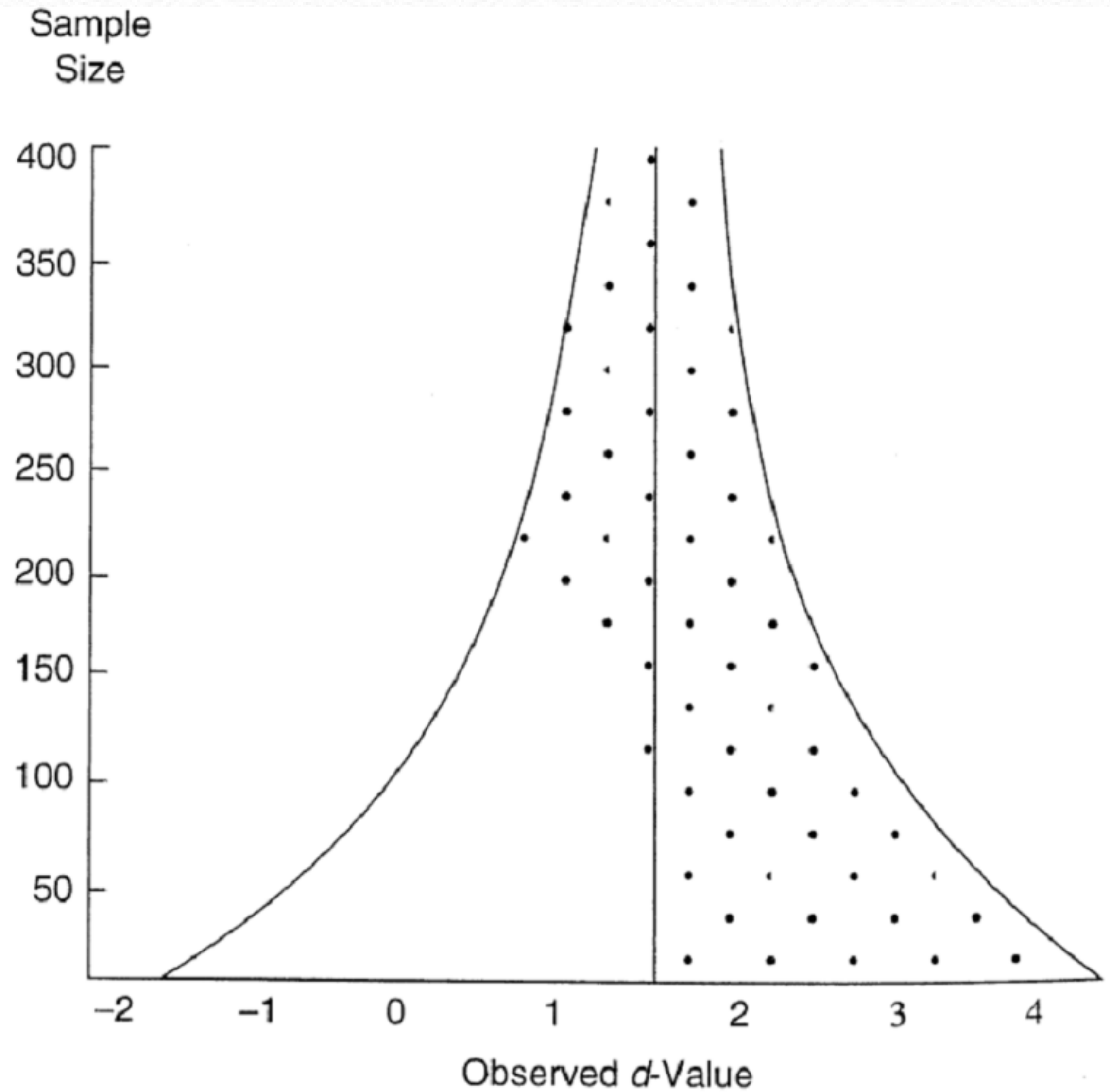
Meta-Analysis

- Combine information, usually in the form of summary statistics, from independent studies
- **effect size:** difference between the means, correlation coefficients, and regression slopes. Alt: model parameters
- **reporting bias problem:** neutral or negative results are less likely to be published than positive results

Funnel Plot

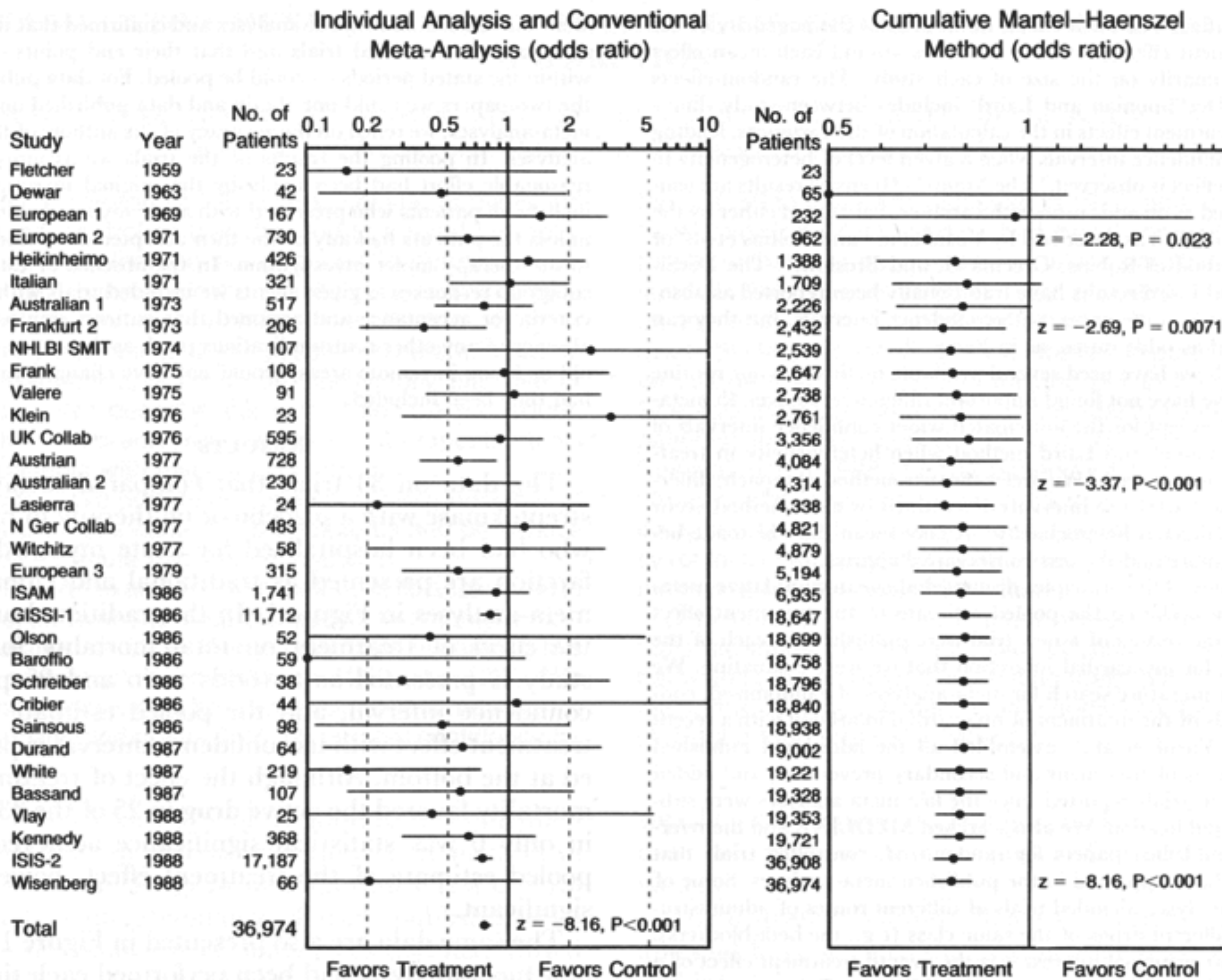


Sample Size
400
300
200
100
−2
−1
0
1
2
3
4
Observed d-Value



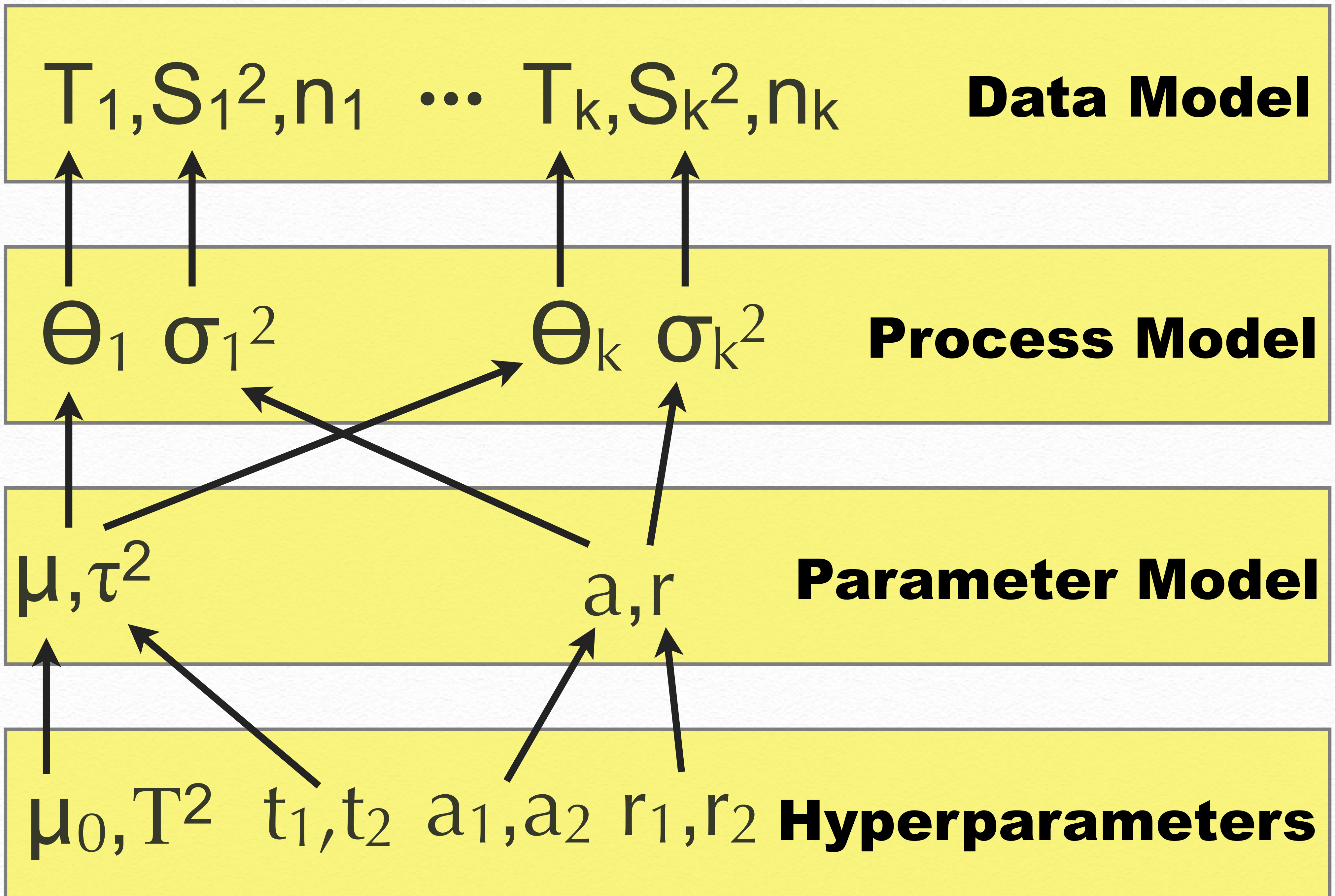
Sample Size
400
350
300
250
200
150
100
50
−2
−1
0
1
2
3
4
Observed d-Value

Cumulative MA



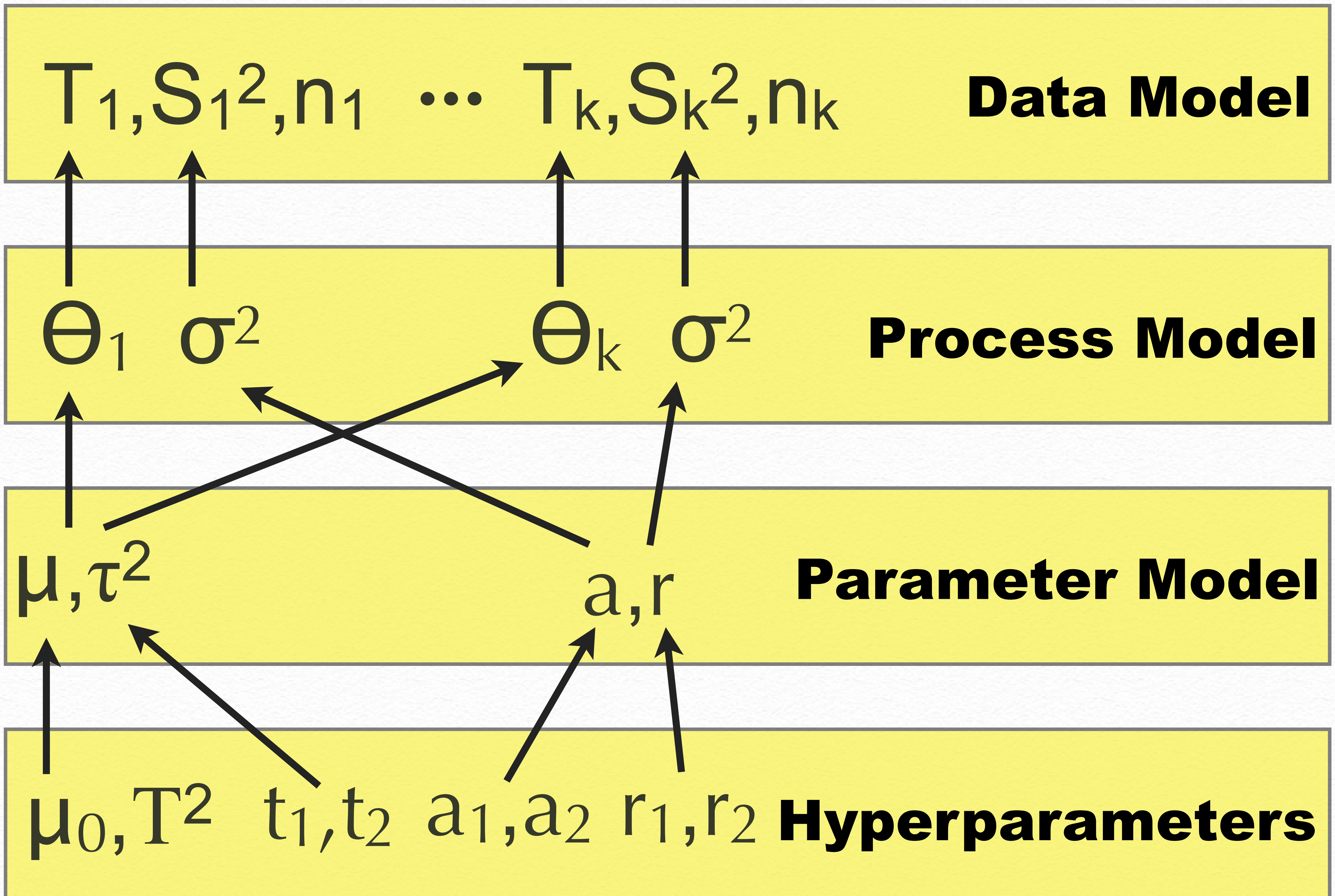
$$T_i \sim N(\theta_i, \sigma_i^2/n_i)$$

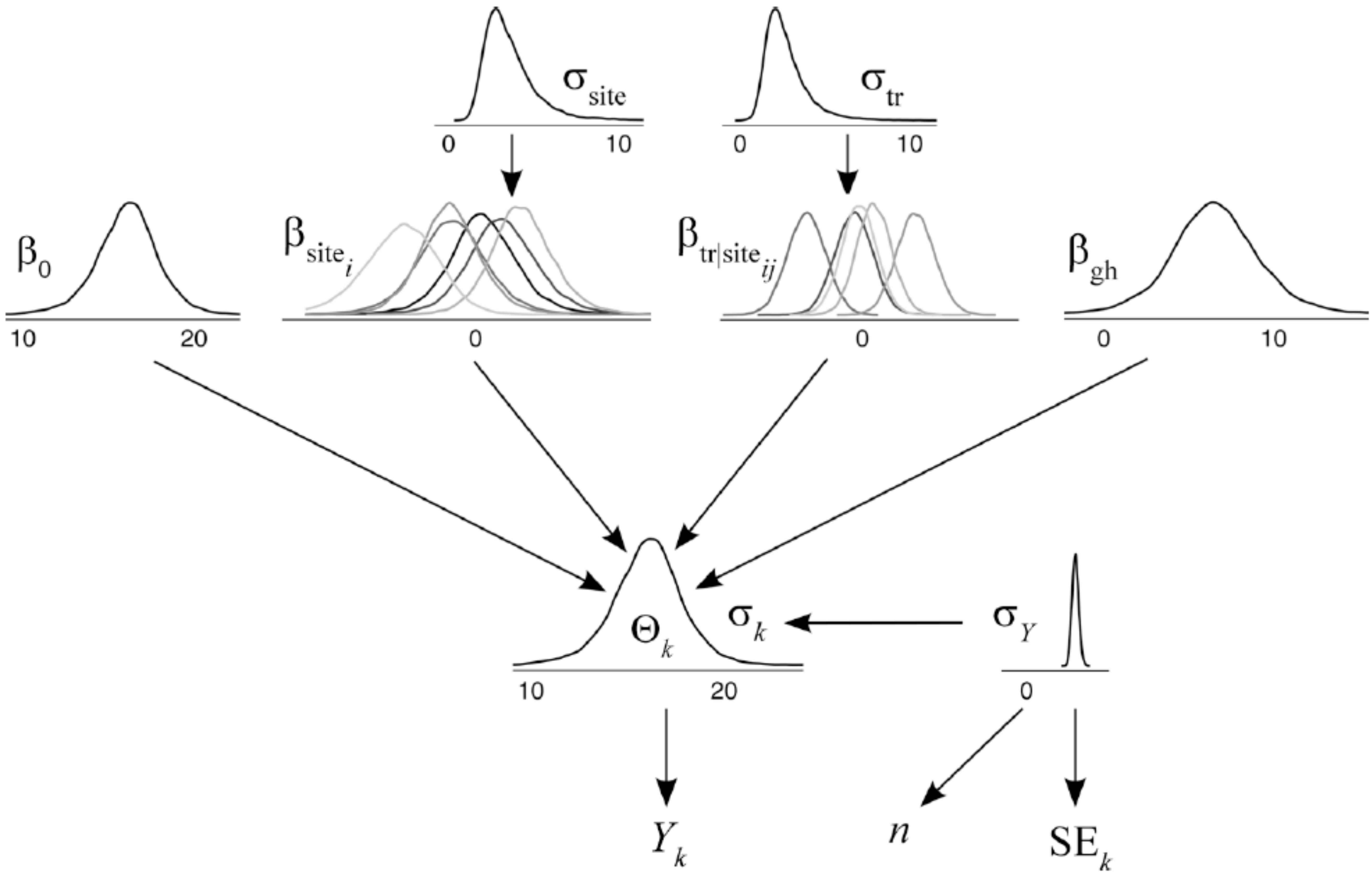
$$\frac{1}{S_i^2} \sim Gamma\left(\frac{n_i}{2}, \frac{n_i \sigma_i^2}{2}\right)$$



$$T_i \sim N(\theta_i, \sigma_i^2/n_i)$$

$$\frac{1}{S_i^2} \sim Gamma\left(\frac{n_i}{2}, \frac{n_i\sigma_i^2}{2}\right)$$





σ_site
0
10
σ_tr
0
10
β_0
10
20
β_site_i
0
β_tr|site_ij
0
β_gh
0
10
Θ_k
σ_k
10
20
σ_Y
0
Y_k
n
SE_k

Combining data:

Practice, pitfalls, and opportunities

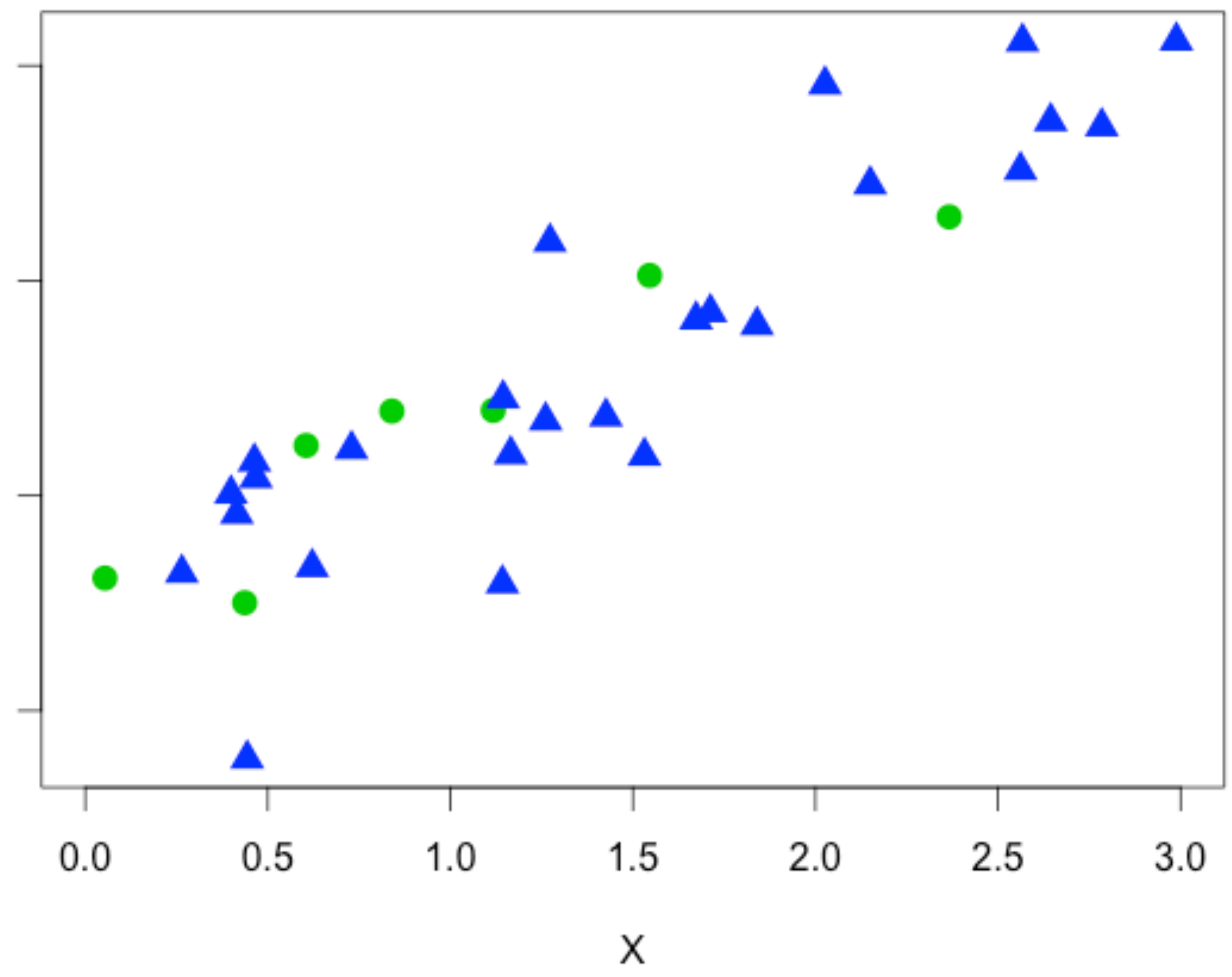
$$\mu = f(x|\theta)$$

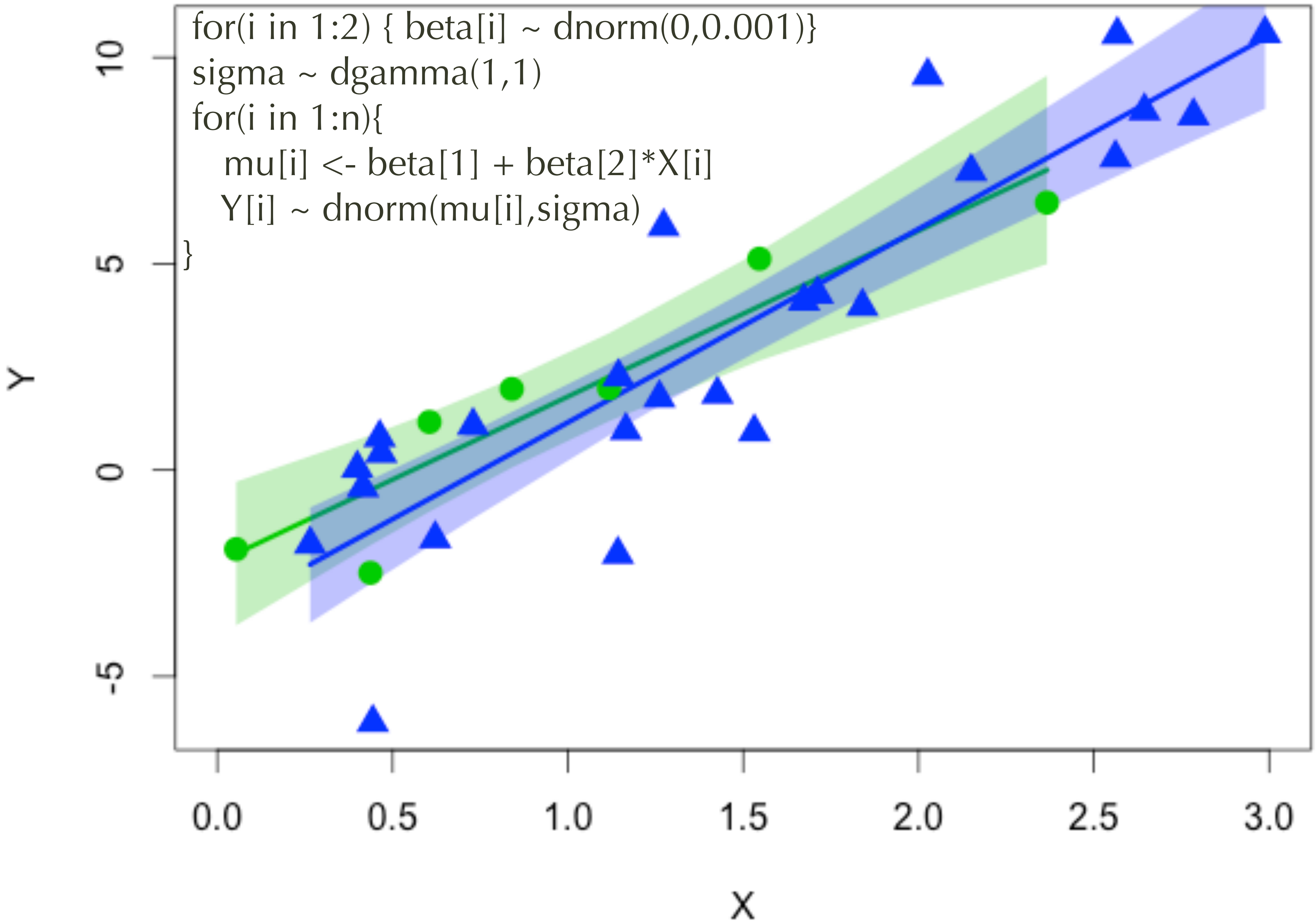
$$\overrightarrow{Y_1} \sim g_1(\mu|\phi_1)$$

$$\overrightarrow{Y_2} \sim g_2(\mu|\phi_2)$$

$$\vdots$$

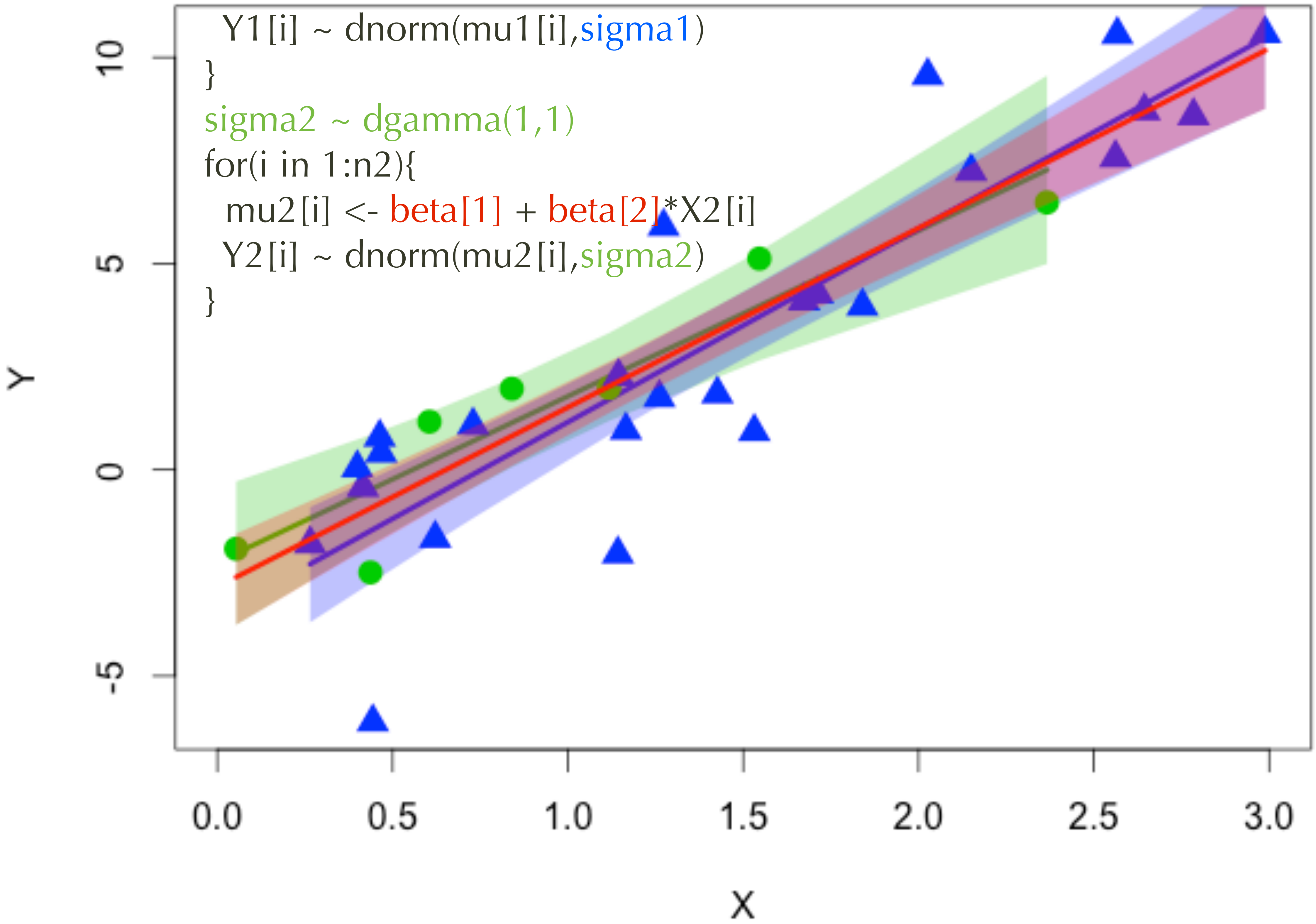
$$\overrightarrow{Y_k} \sim g_k(\mu|\phi_k)$$



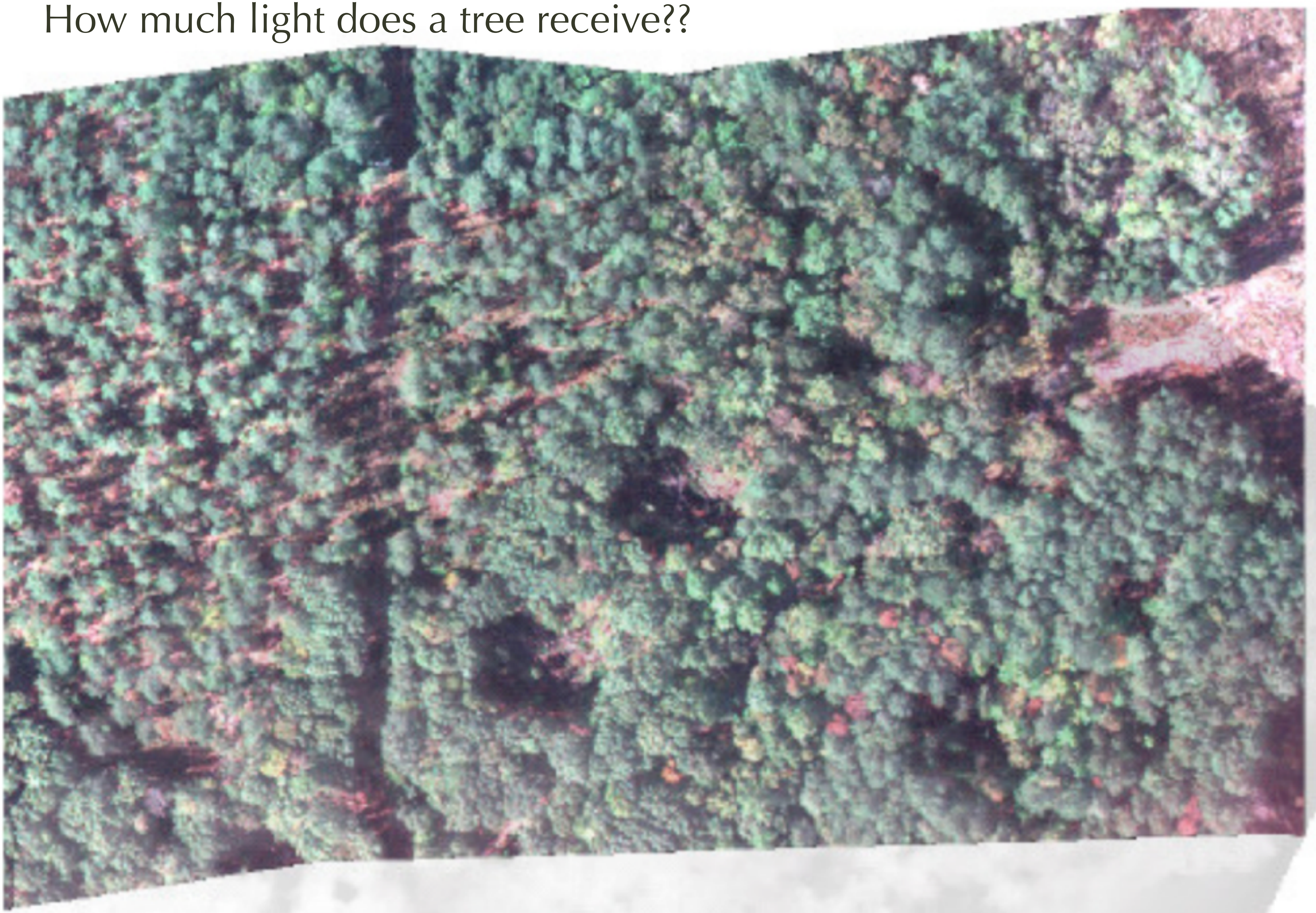


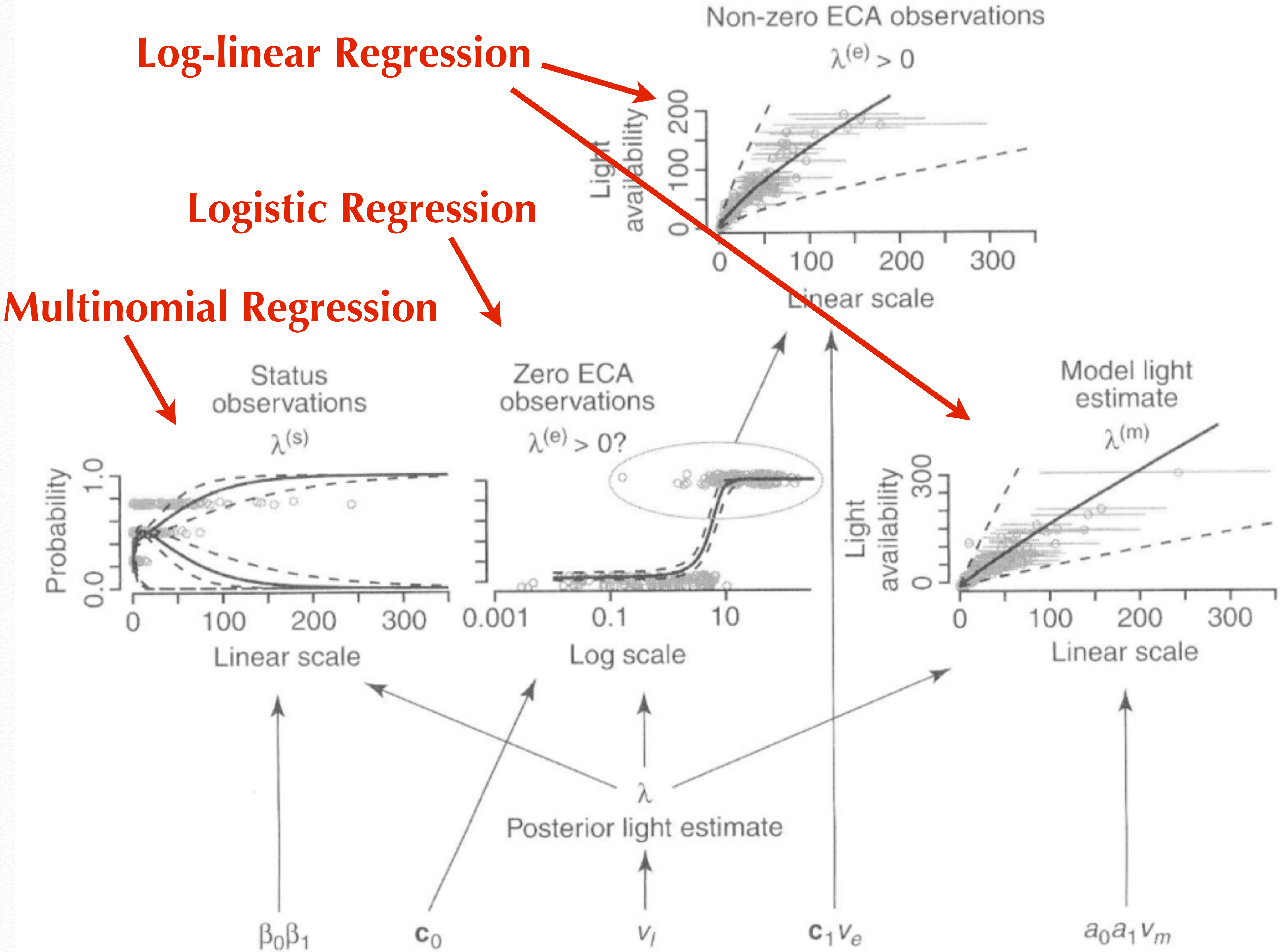

```
for(i in 1:2) { beta[i] ~ dnorm(0,0.001)}
sigma ~ dgamma(1,1)
for(i in 1:n){
   mu[i] <- beta[1] + beta[2]*X[i]
   Y[i] ~ dnorm(mu[i],sigma)
}
```

```
for(i in 1:2) { beta[i] ~ dnorm(0,0.001)}
sigma1 ~ dgamma(1,1)
for(i in 1:n1){
  mu1[i] <- beta[1] + beta[2]*X1[i]
  Y1[i] ~ dnorm(mu1[i],sigma1)
}
sigma2 ~ dgamma(1,1)
for(i in 1:n2){
  mu2[i] <- beta[1] + beta[2]*X2[i]
  Y2[i] ~ dnorm(mu2[i],sigma2)
}
```



How much light does a tree receive??





Log-linear Regression
Logistic Regression
Multinomial Regression
Non-zero ECA observations
λ(e) > 0
Light availability
0 100 200
0 100 200 300
Linear scale
Status observations
λ(s)
Probability
0.0 1.0
0 100 200 300
Linear scale
Zero ECA observations
λ(e) > 0?
0.001 0.1 10
Log scale
Model light estimate
λ(m)
Light availability
0 300
0 100 200 300
Linear scale
λ
Posterior light estimate
β0β1
c0
vl
c1ve
a0a1vm

Unequal sizes

- Common when combining manual & automated data
- Subset of sources dominates likelihood
- Ad hoc sol'n in the literature
  - Likelihood reweighting **Invalid**
  - Data subsampling
  - Data averaging

Loss of Info, Arbitrary

Avoid Double Dipping

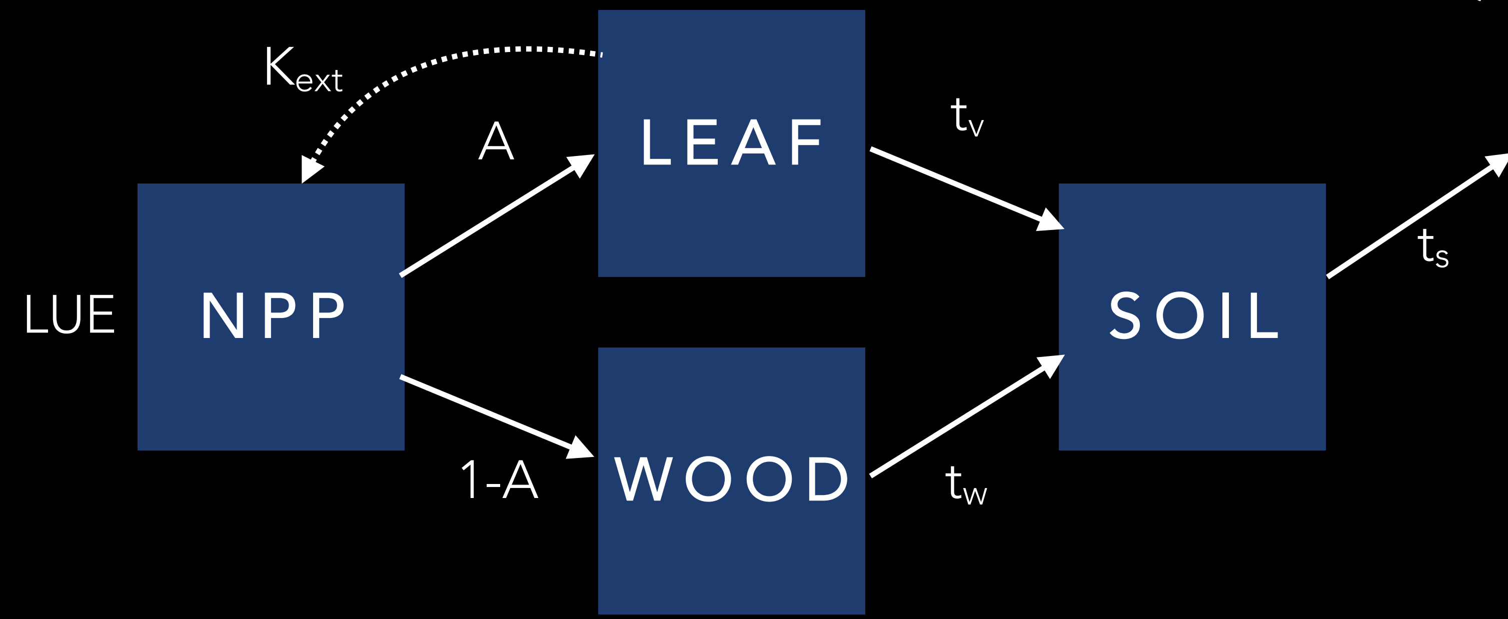
Treat Uncertainty Appropriately!

- Autocorrelation
  - time (inc. repeated measures)
  - space
- Random effects
  - n=1 sensor? Bias
- Systematic error Does not average out!

PSEUDODATA EXPERIMENTS

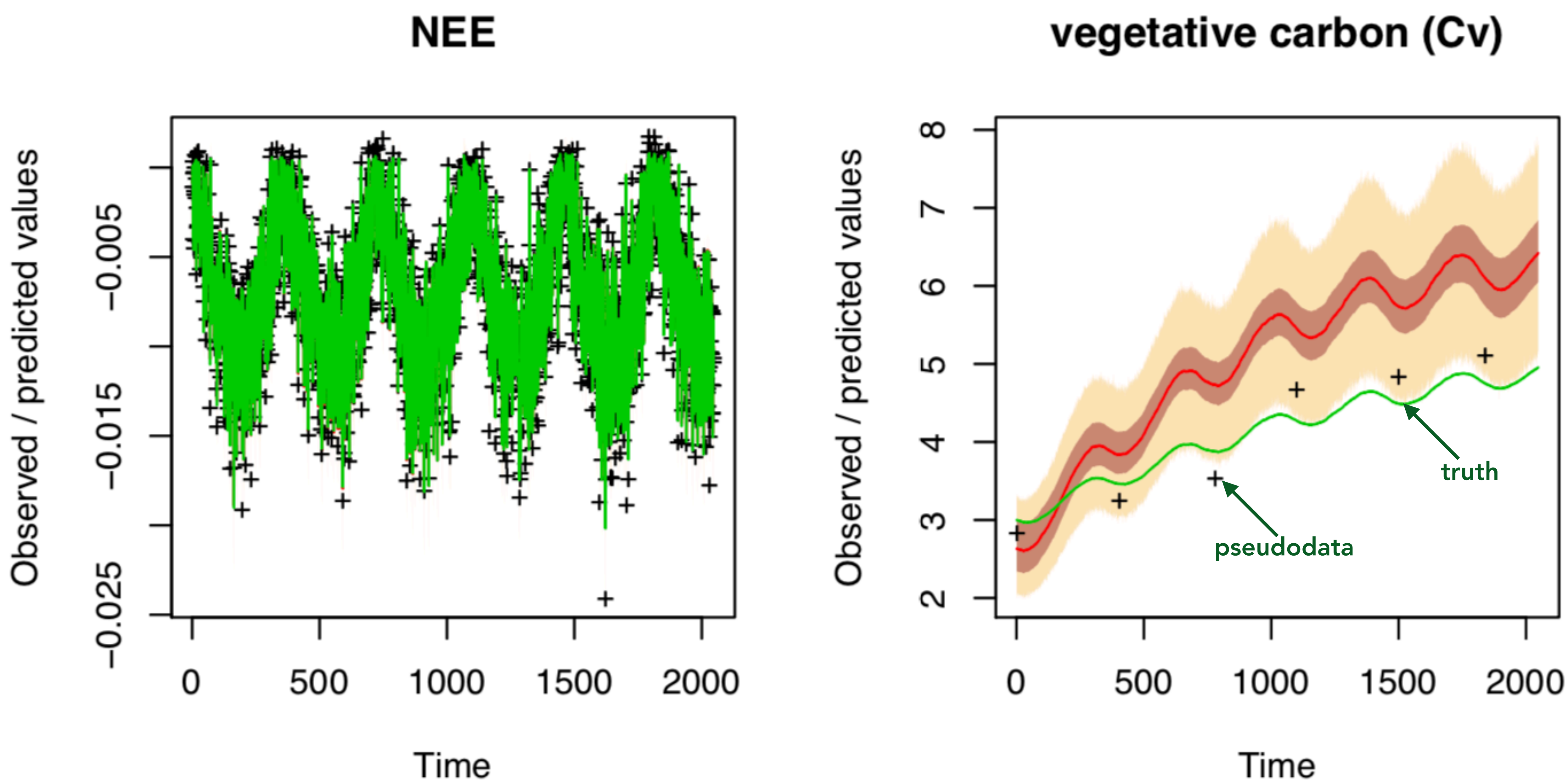
work w/ David Cameron (in prep)

VSEM



	Fit	**Param**
• Perfect model, balanced data	✓	✓
• Perfect model, unbalanced data	✓	✓
• Model error, balanced data	✓	✗
• Model error, unbalanced data	✗	✗

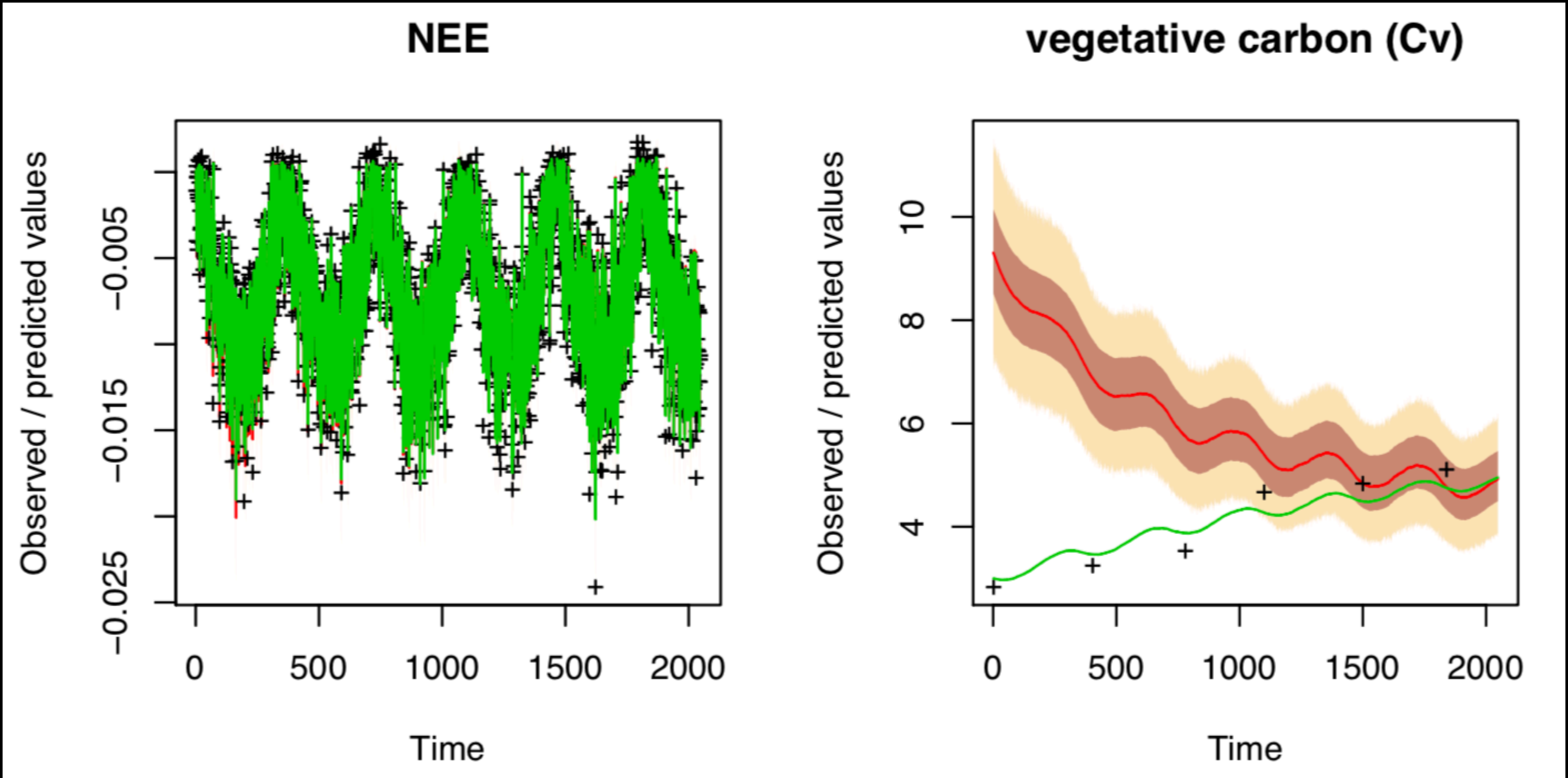
MODEL ERROR, UNBALANCED DATA



WHAT IF THERE ARE DATA BIASES

	Fit	Param
• PM, Balanced, multiplicative data bias	✓	✗
• PM, Unbalanced, MB data	✗	✗
• Model error, Unbalanced, MB data	✗	✗

MODEL ERROR UNBALANCED DATA W/ MULTIPLICATIVE BIAS

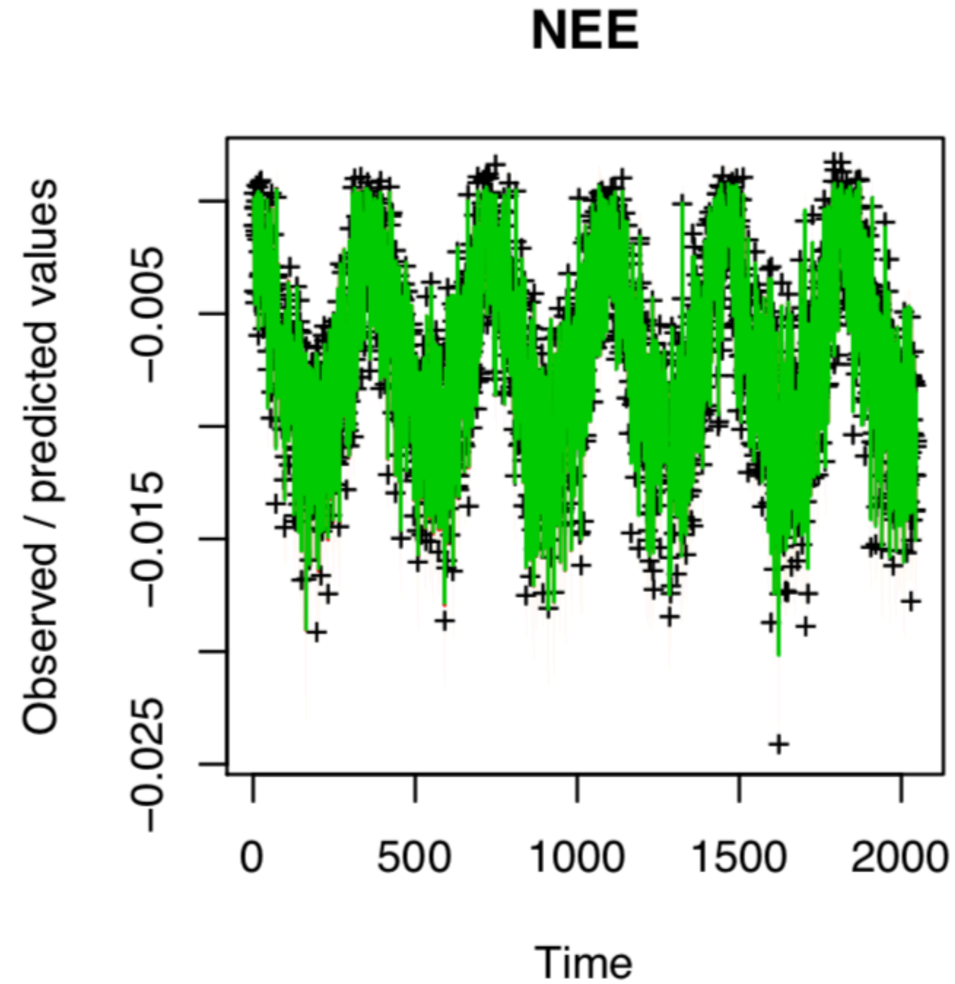


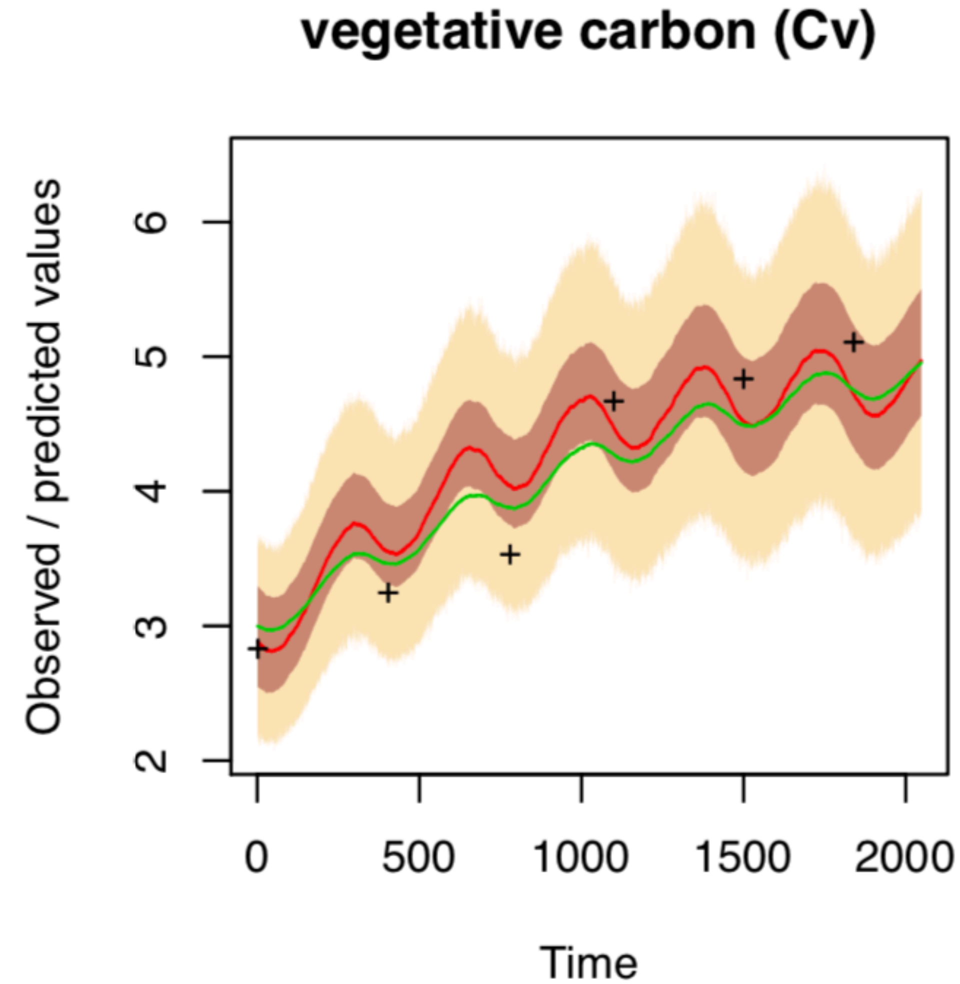
WHAT IF THERE ARE DATA BIASES

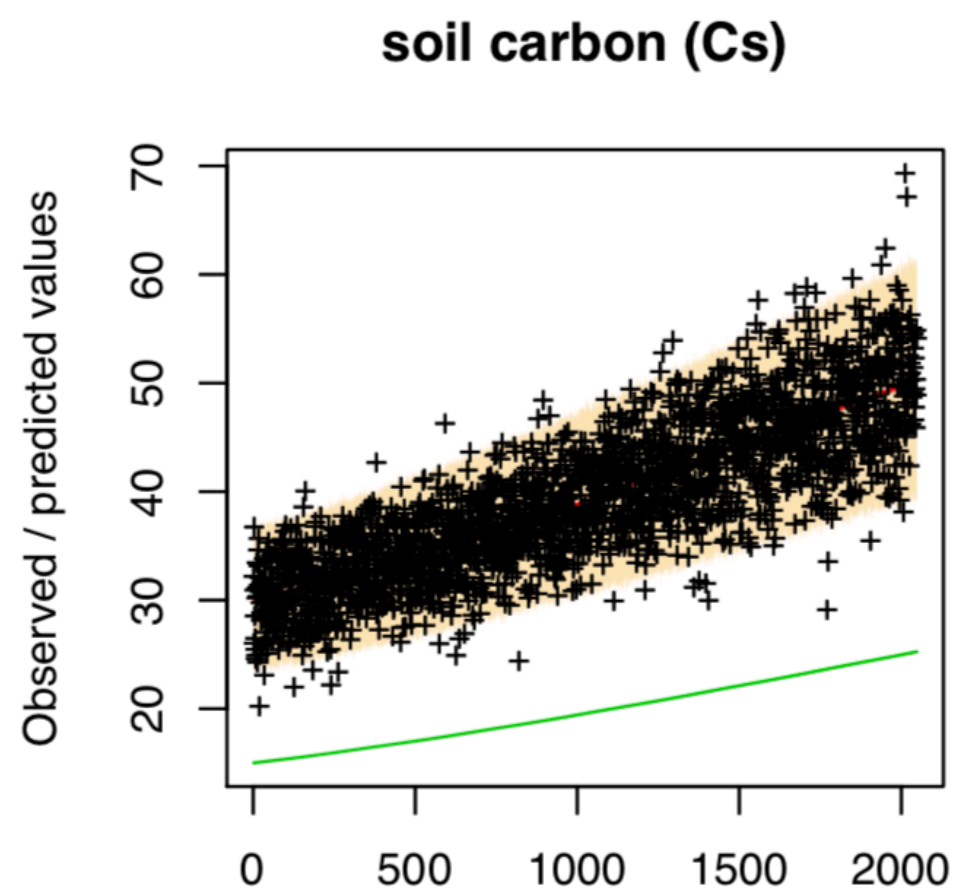
	Fit	Param
• PM, Balanced, multiplicative data bias	✓	✗
• PM, Unbalanced, MB data	✗	✗
• Model error, Unbalanced, MB data	✗	✗

CAN WE CORRECT BIAS W/ STATS?

• ME, Unbalanced, linear model	✓	✗
• PM, Unbalanced, MB data, linear model	✓	✗
• ME, Unbalanced, MB data, linear model	✓	✗







MODEL W/ ERROR
UNBALANCED

DATA W/
MULTIPLICATIVE BIAS

LINEAR BIAS COR

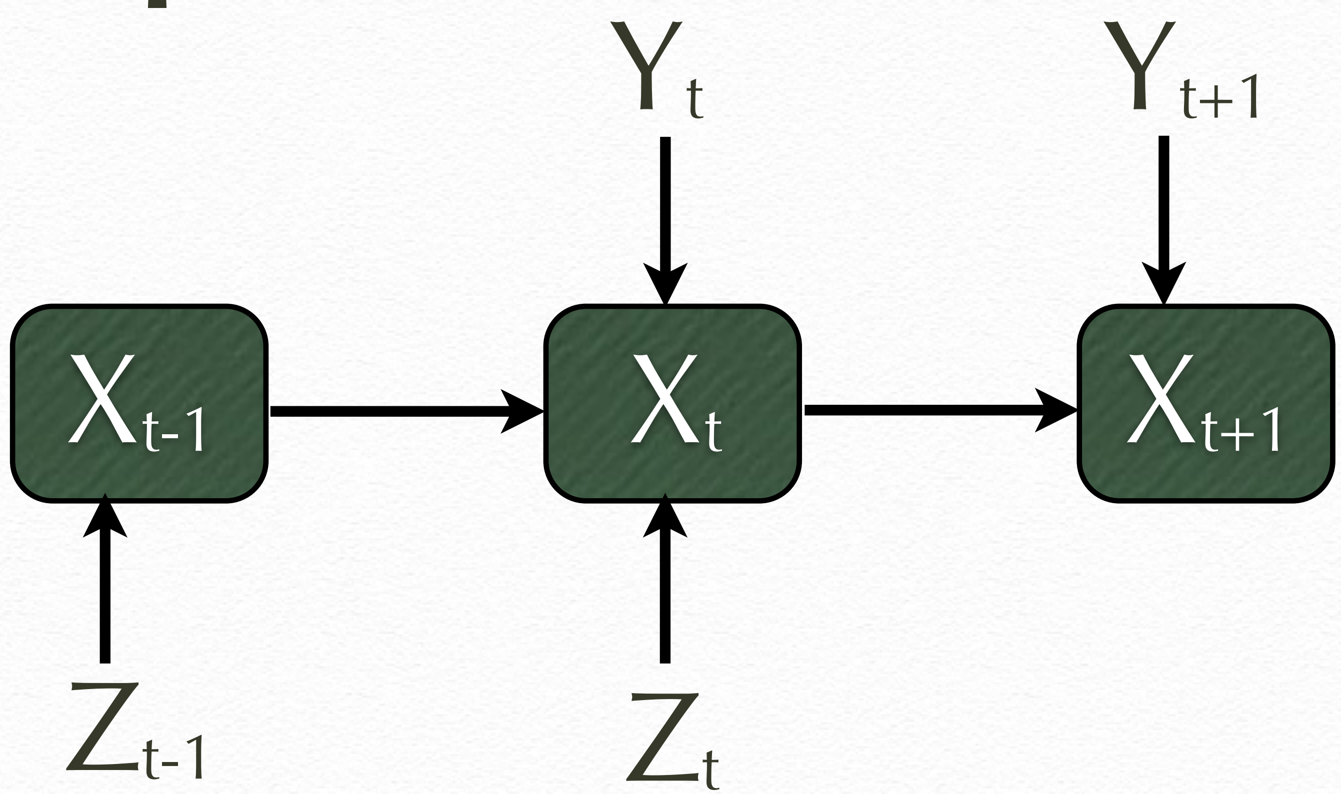
UNBALANCED DATA: TAKE HOMES

- More than Information content,
  issue is **errors & biases** in models and data

- Perfect models can fit unbalanced data

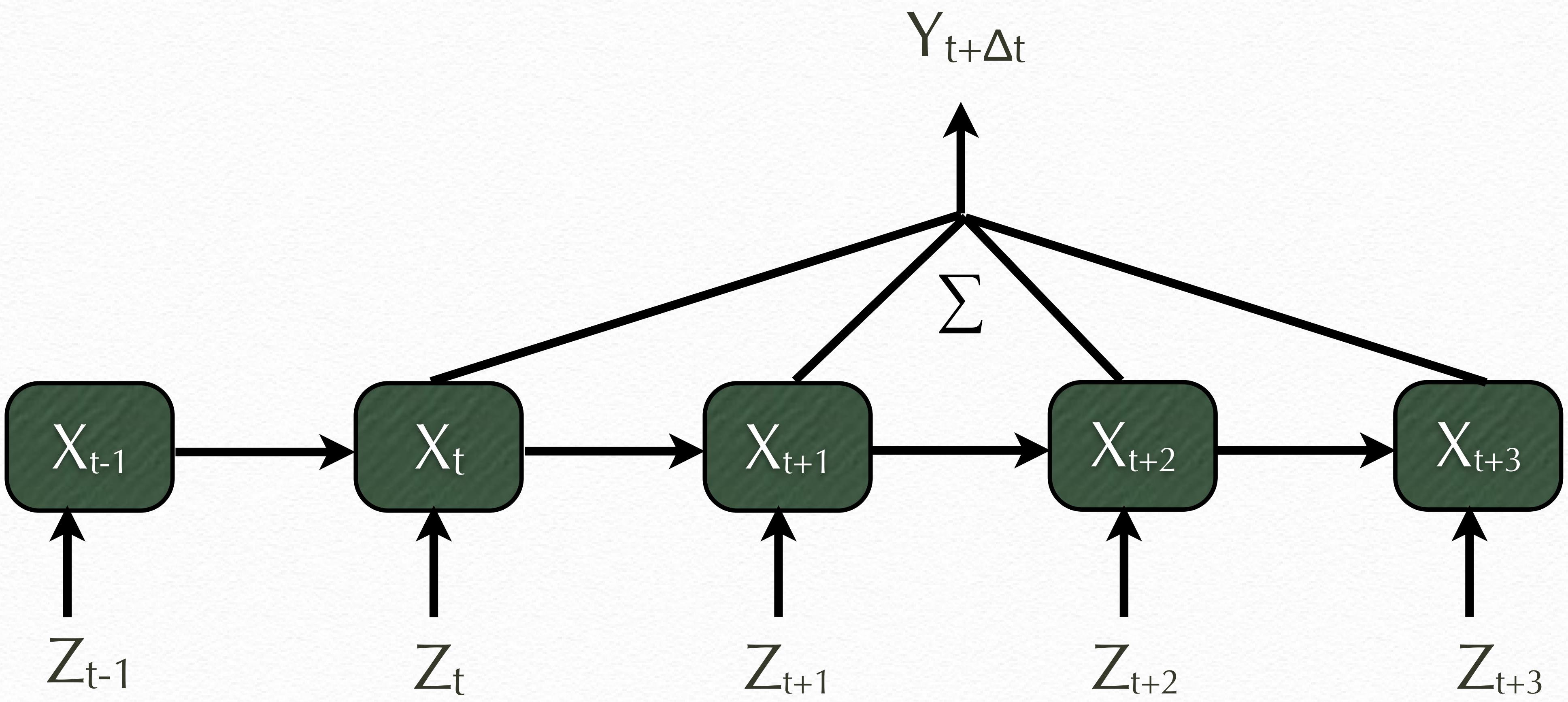
- Building bias correction into calibration can lead to acceptable performance, but not the 'true' parameters

Combining across space and time



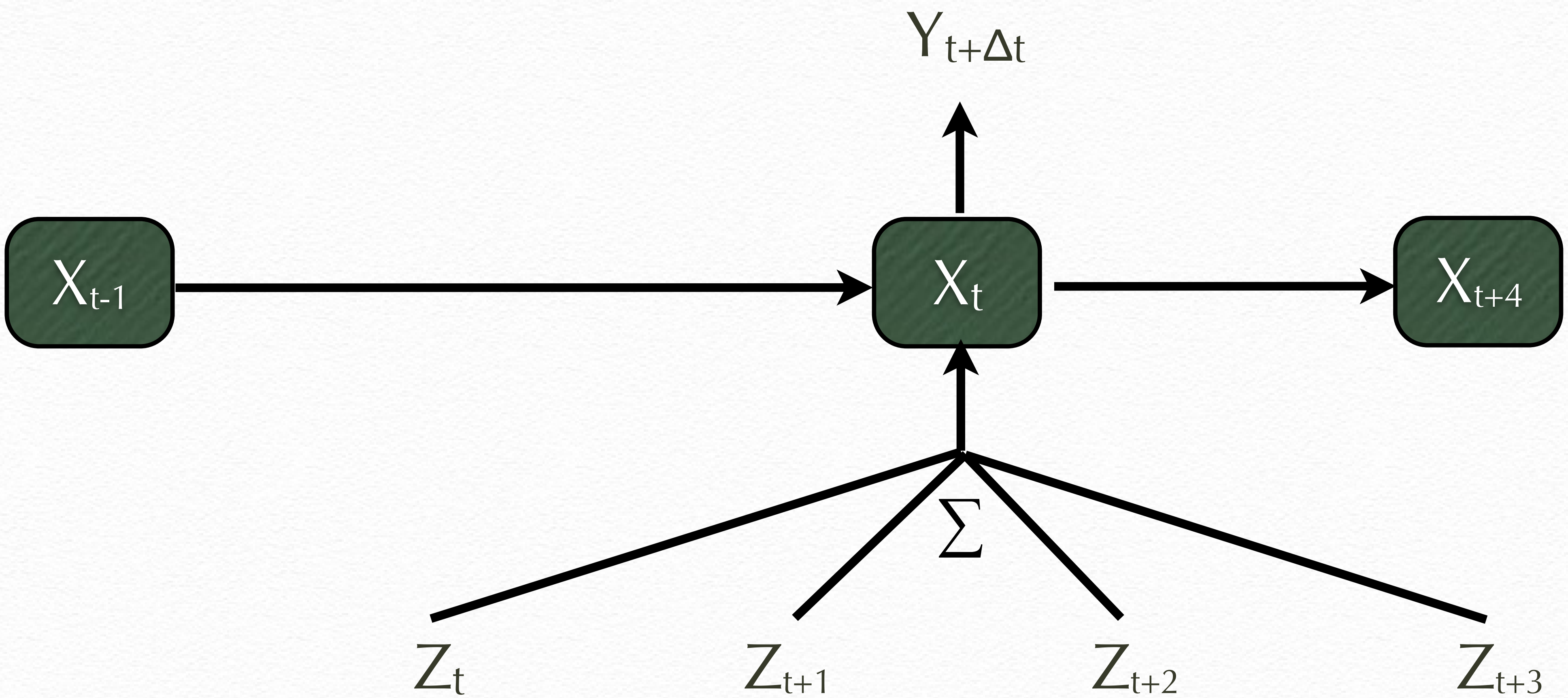
- Straightforward extension of state-space
- What if data are from different scales?
  - Scale the process model
  - Aggregate the data model

Integrated Observations



- Option 1: work at the finer resolution
  - Pro: full info of hi res
  - Con: computation, identifiability

Integrated Observations



- Option 2: work at the coarser resolution
  - Pro: computation
  - Con: loss of information

“Change of support”

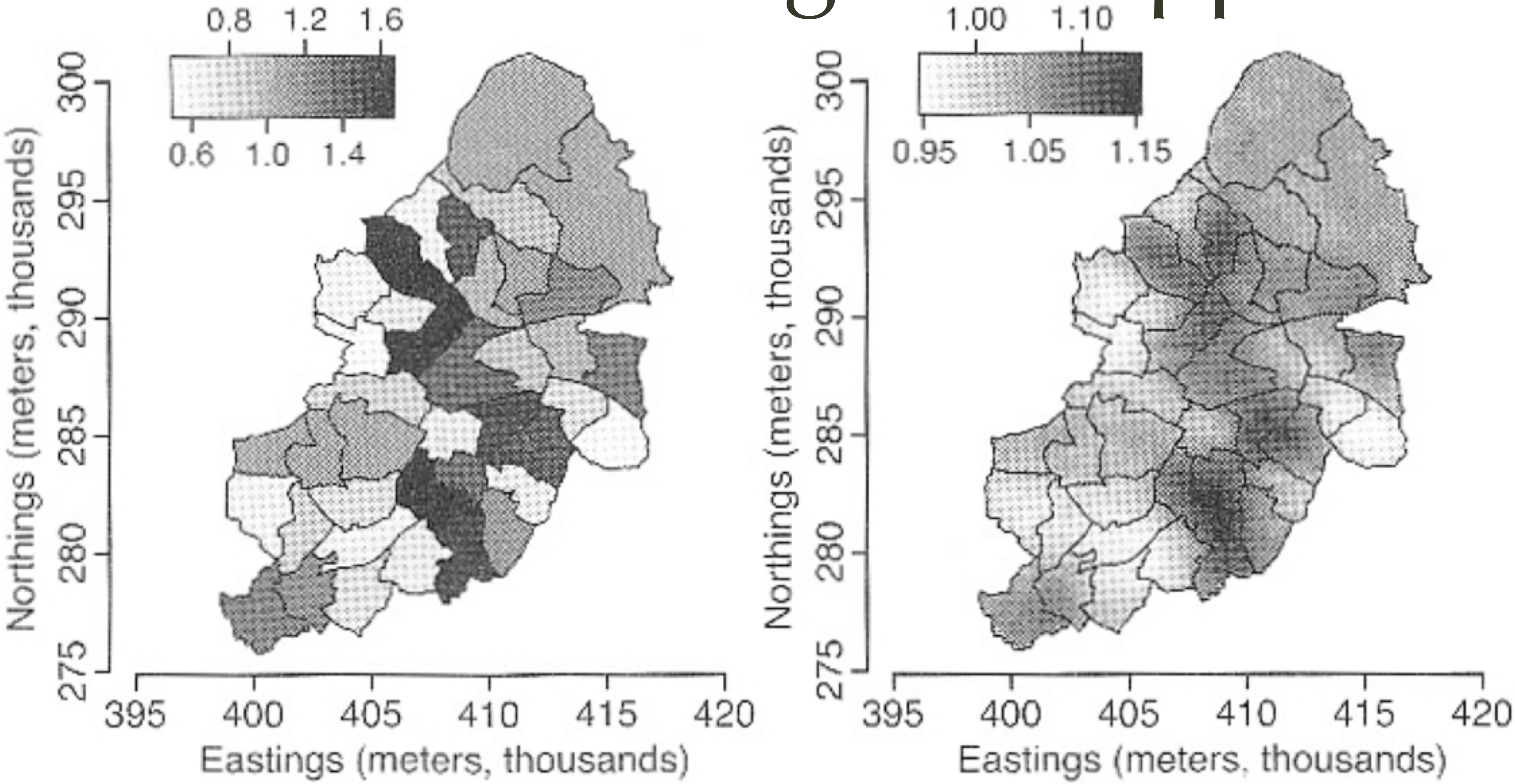
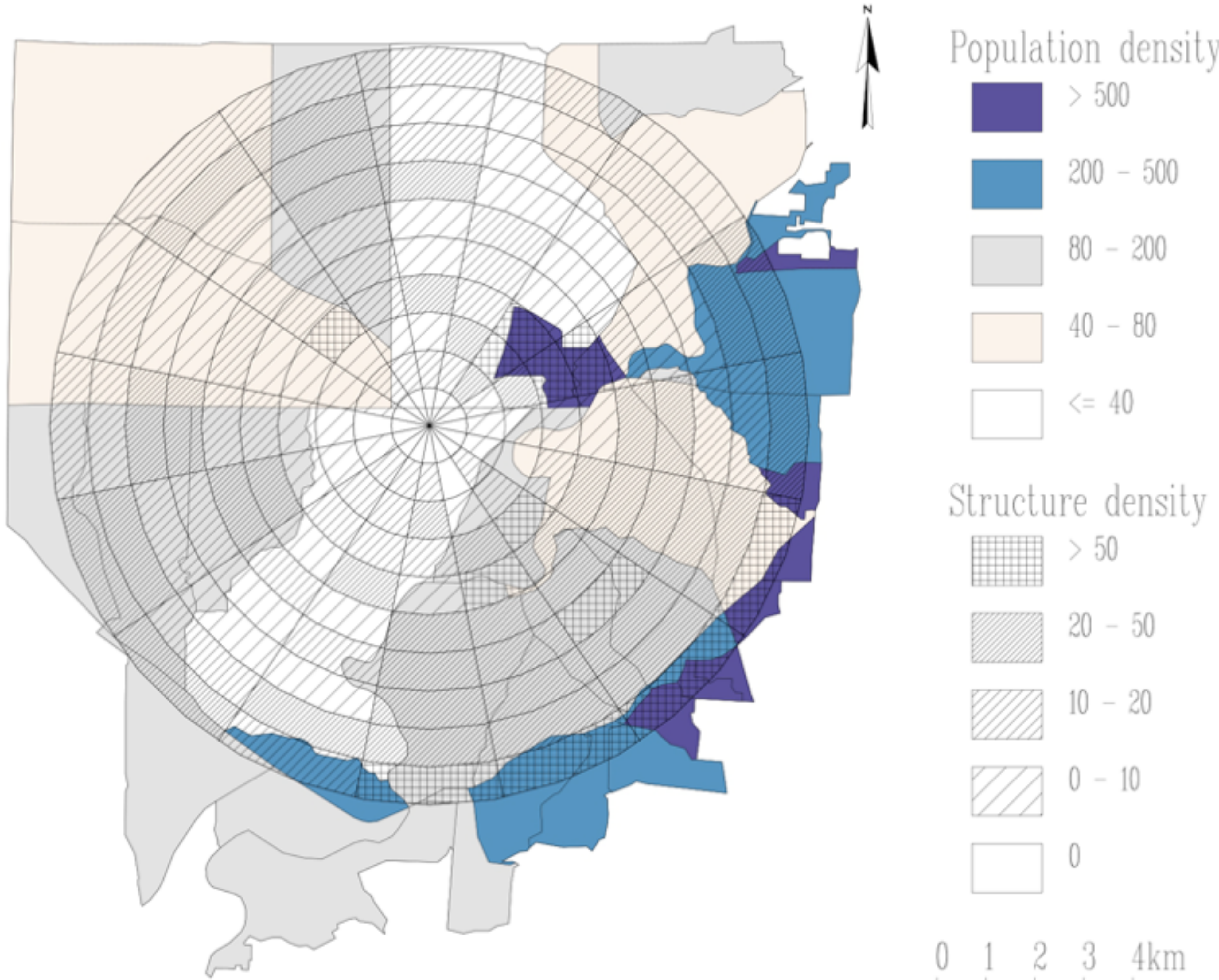


FIGURE 10.20. Standardized mortality ratios for thirty-nine wards in Birmingham, England, calculated as observed versus expected cases (*left*), and posterior median relative risk $y(s)$ (*right*). From Kelsall and Wakefield (2002).

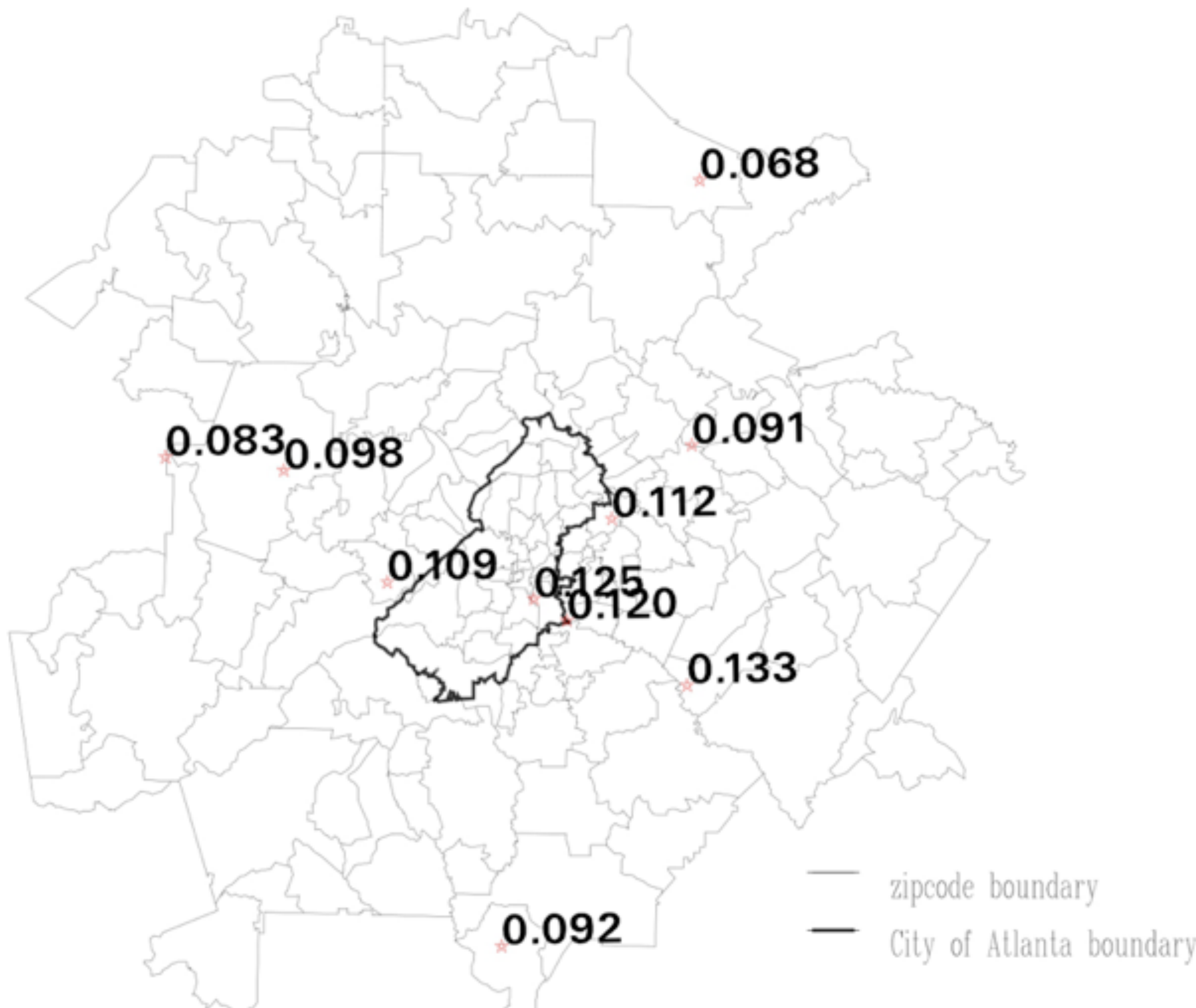
Block-Block Misalignment

Population by census tract; residential structures by "cell":



Bivariate misalignment

Ozone measurements at fixed sites; counts of pediatric asthma cases by zip code in Atlanta, GA:



SCALING "GOTCHAS"

- Naive regridding/interpolations
  - Artificially inflates sample size / information content
  - Ignores uncertainties
- Fit all at once
- Fit piecewise -> errors in variables