Day 1
Linking theory and data: how to do it?

FRB
FONDATION
POUR LA RECHERCHE
SUR LA BIODIVERSITÉ

CESAB
CENTRE FOR THE SYNTHESIS AND ANALYSIS
OF BIODIVERSITY

TheoMoDive

It depends on your objective

It depends on your objective

1. **Test/evaluate the theory**

- Test the model predictions: *do they match with observations?*
- Test the model assumptions: *are they well supported by empirical evidence?*

It depends on your objective

1. **Test/evaluate the theory**

- Test the model predictions: *do they match with observations?*
- Test the model assumptions: *are they well supported by empirical evidence?*

2. **Use the theory**

- Adopt a general framework: *put your study in a broader context*
- Use mathematical equations: to *estimate parameters*

It depends on your objective

1. **Test/evaluate the theory**

- Test the model predictions: *do they match with observations?*
- Test the model assumptions: *are they well supported by empirical evidence?*

2. **Use the theory**

- Adopt a general framework: *put your study in a broader context*
- Use mathematical equations: *to estimate parameters*

1. Adopt a general framework

Provide a context to integrate evidence from observations, experiments, models

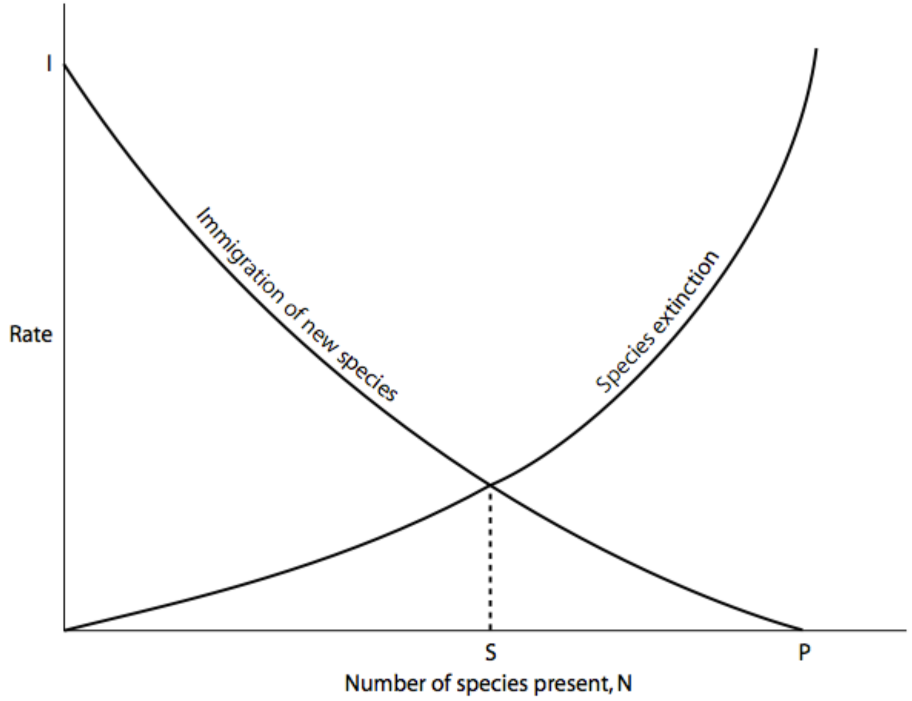
Help focus empirical research on a specific process or relationship

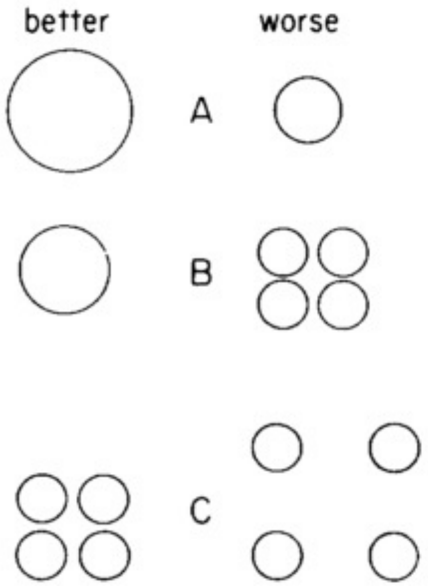
1. Adopt a general framework

Provide a context to integrate evidence from observations, experiments, models

Help focus empirical research on a specific process or relationship

- E.g., Theory of island biogeography (MacArthur & Wilson, 1967)





Design of protected areas (Diamond 1975)

1. Adopt a general framework

Provide a context to integrate evidence from observations, experiments, models

Help focus empirical research on a specific process or relationship

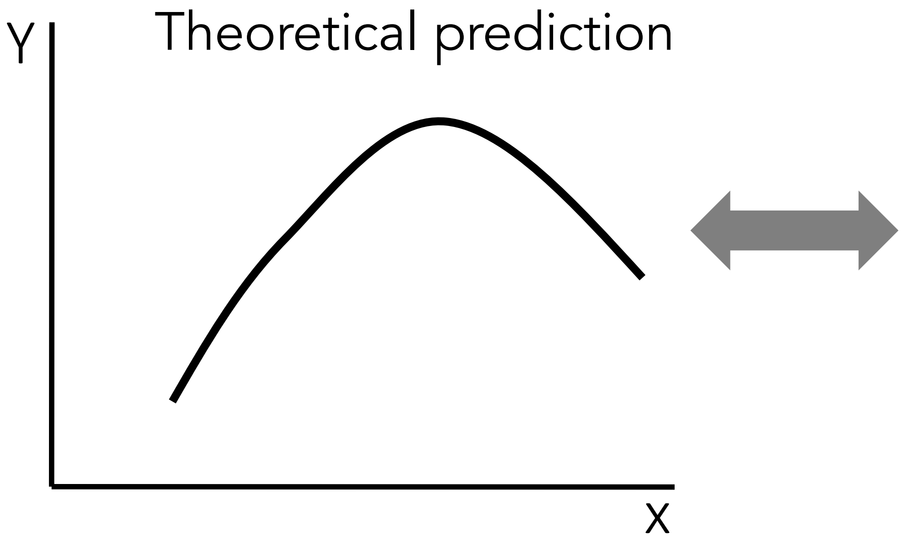
Tips:

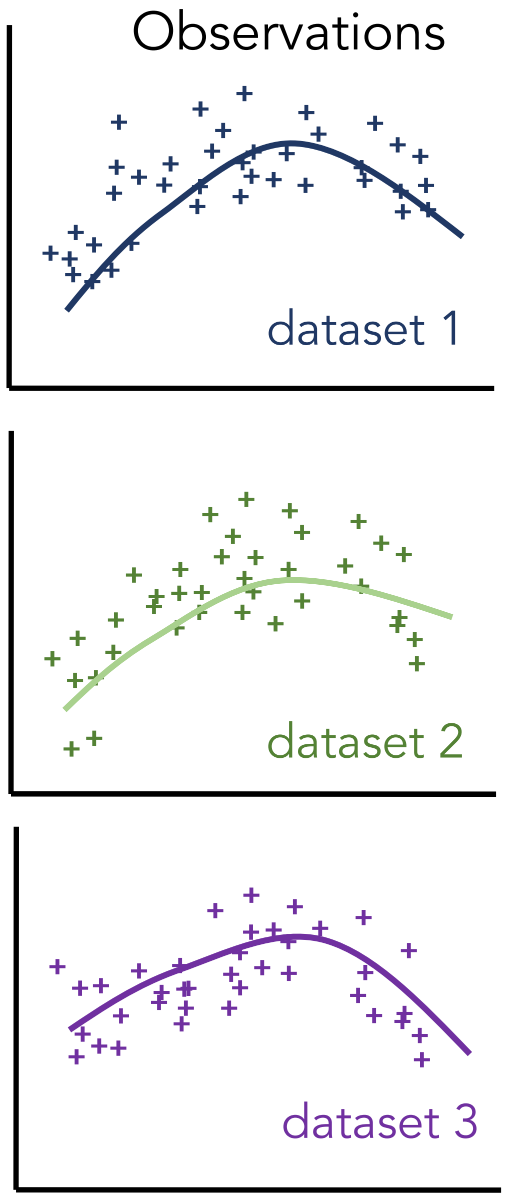
- Become familiar with the framework
- Look at existing empirical work that has adpoted it
- Which aspects have been well explored?
- Where are the knowledge gaps?

2. Test the predictions

2.1. Test the predictions, qualitatively

- Data visualization
- Satistical analysis

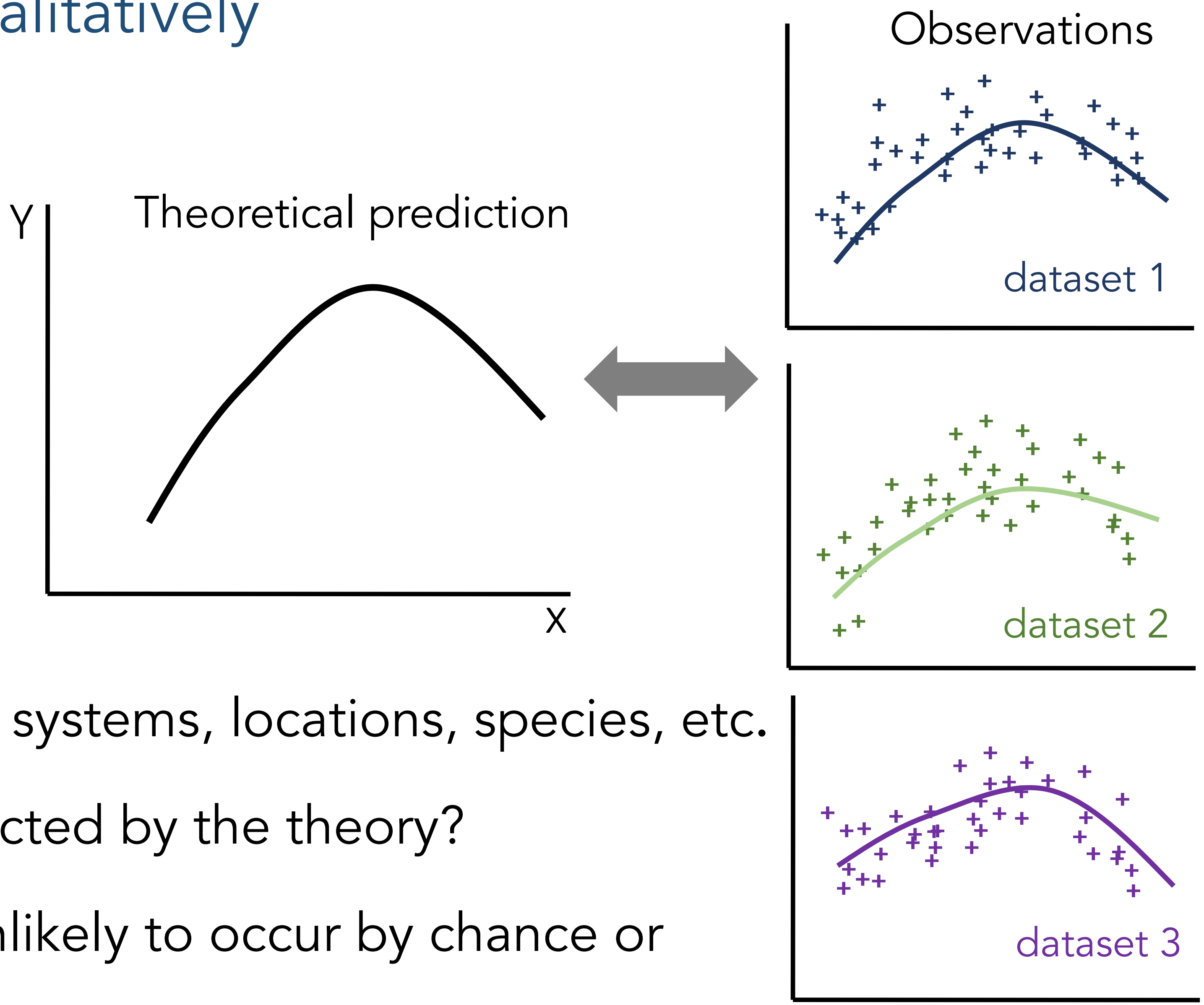




- Repeated observations across systems, locations, species, etc.

2.1. Test the predictions, qualitatively

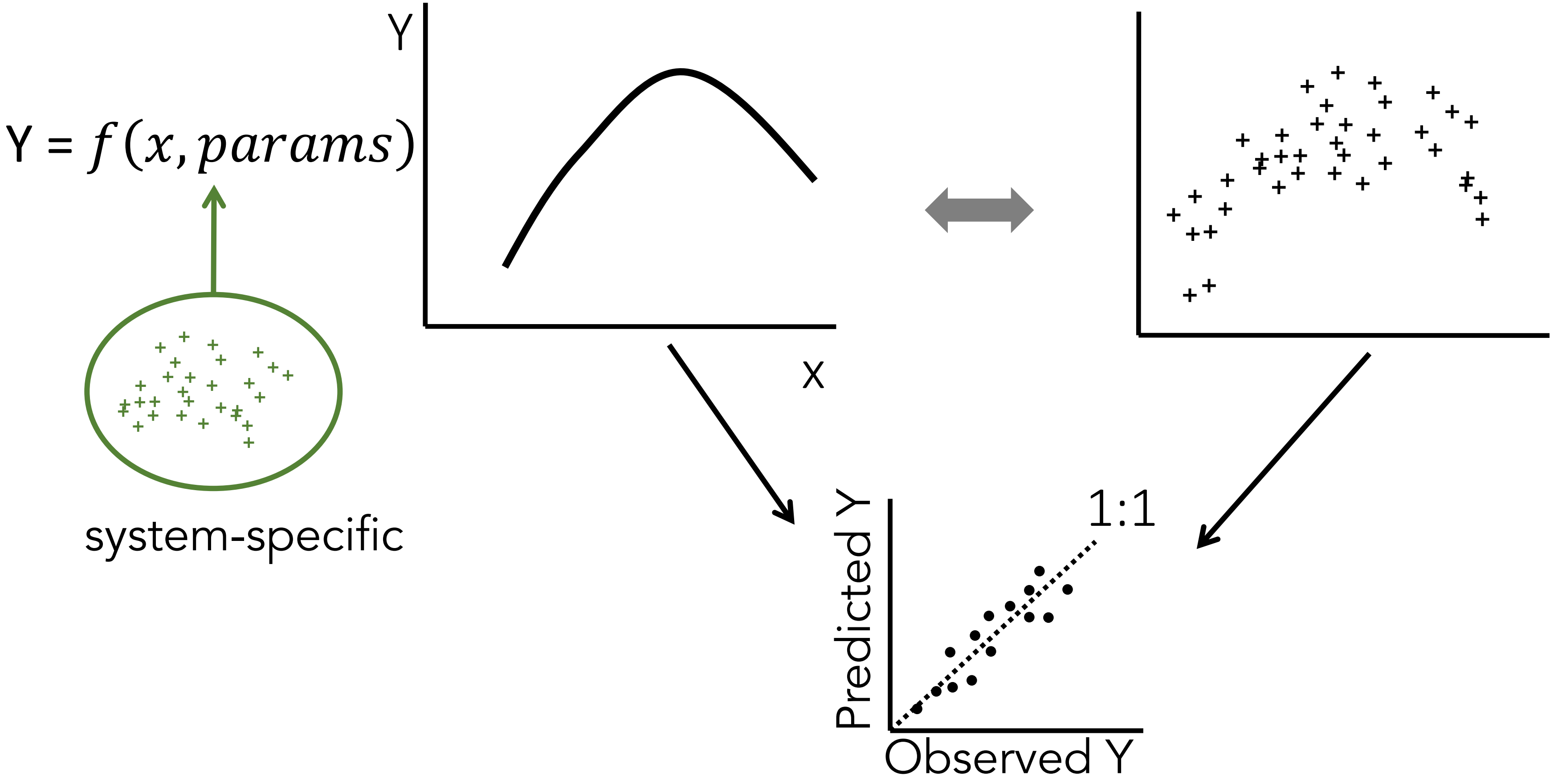
- Data visualization
- Satistical analysis



- Repeated observations across systems, locations, species, etc.
- When does x affect y, as predicted by the theory?
- Powerful if the prediction is unlikely to occur by chance or by alternative mechanisms

2.2. Test the predictions, quantitatively

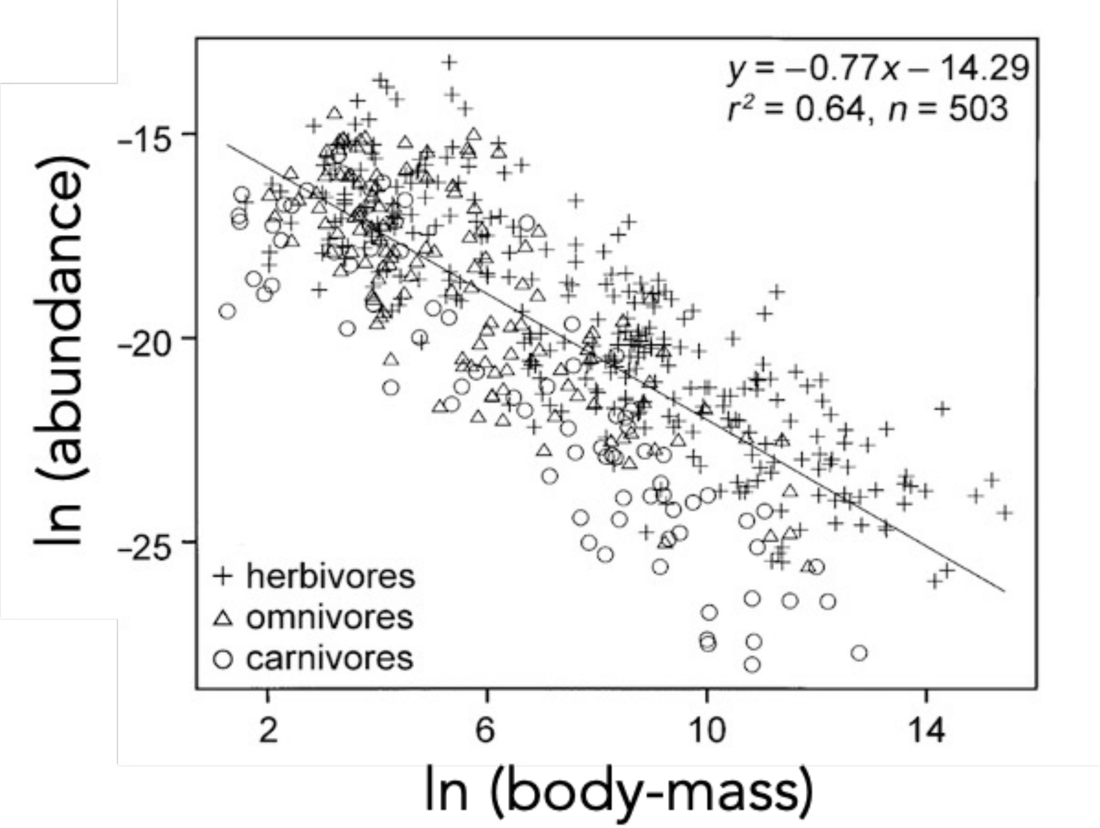
- Aim: determine if a theoretical prediction applies to a specific system
- Model parameterization: input parameters are known for the system



2.2. Test the predictions, quantitatively

Example: how do disturbance frequency impact size-abundance relationships?

Model + experiment

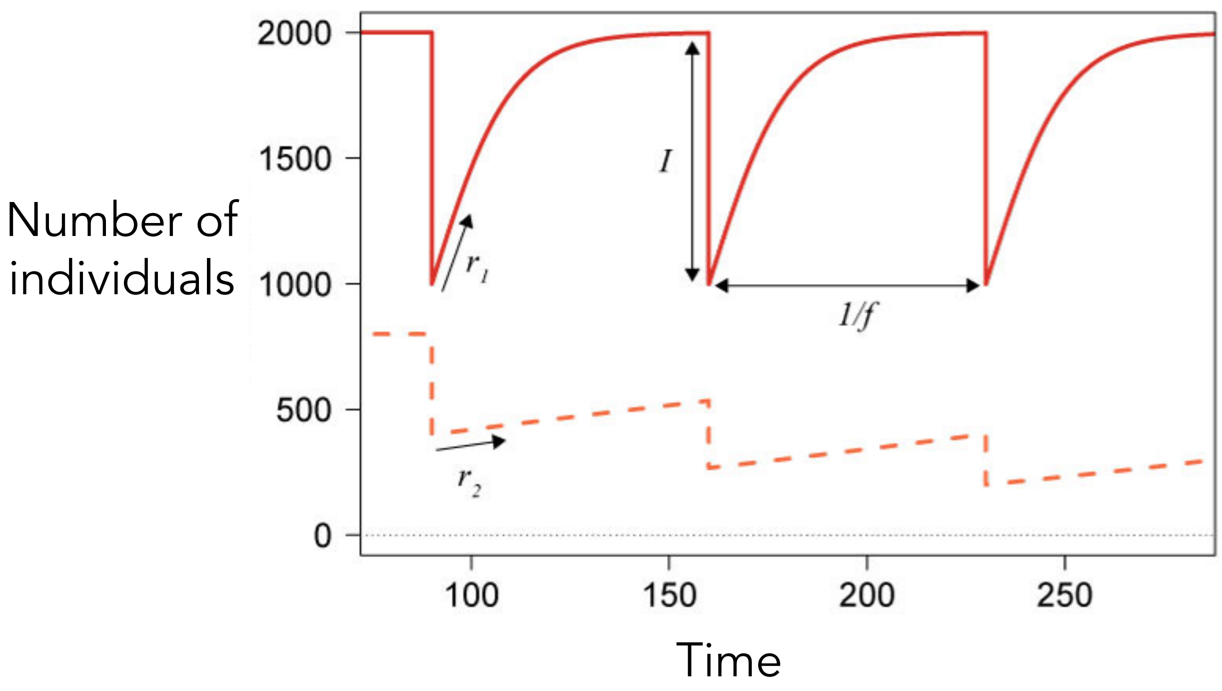


2.2. Test the predictions, quantitatively

The model: multiple populations with logistic growth

Parameters:

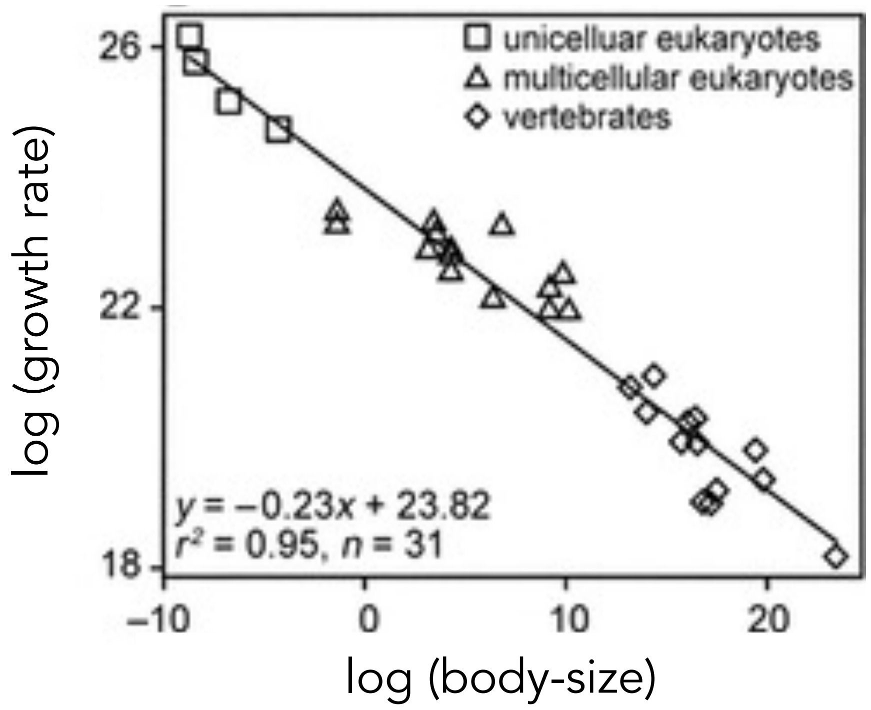
- Population growth rates, *r*
- Carrying capacities, *K*
- Disturbance frequency, *f*
- Disturbance intensity, I



2.2. Test the predictions, quantitatively

Use of **allometric scaling** to link population growth rate *r* and body-size M

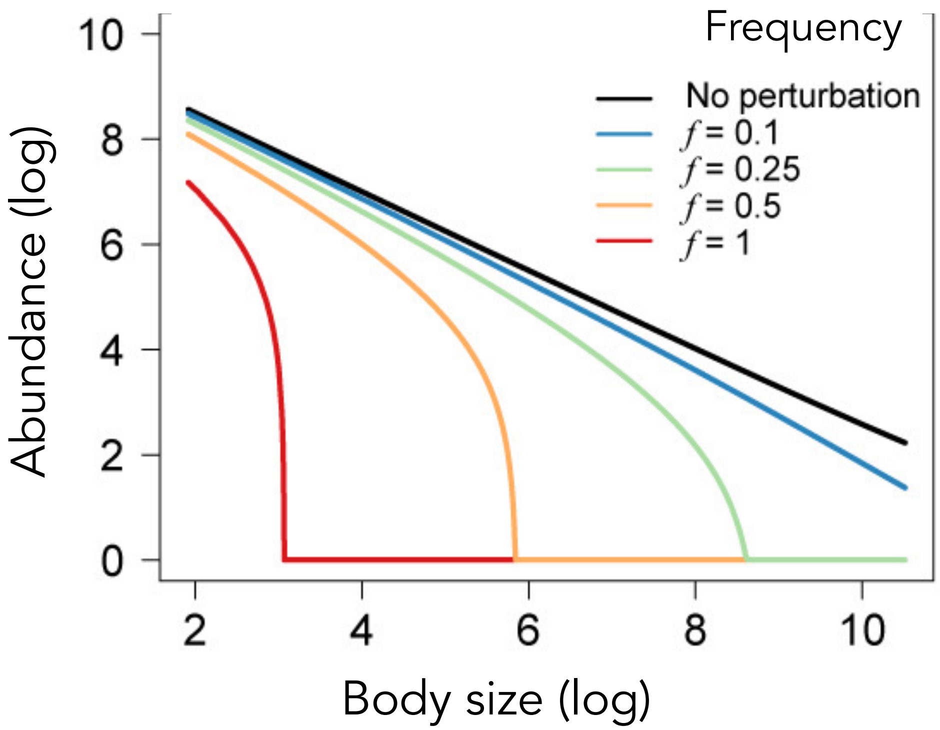
Theoretical framework: Metabolic Theory of Ecology (Brown *et al*. 2004)



2.2. Test the predictions, quantitatively

How do disturbance frequency impact size-abundance relationships?

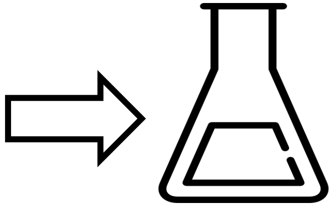
Model predictions



Jacquet *al.* (2020)

2.2. Test the predictions, quantitatively

Experiment



100 mL microcosm

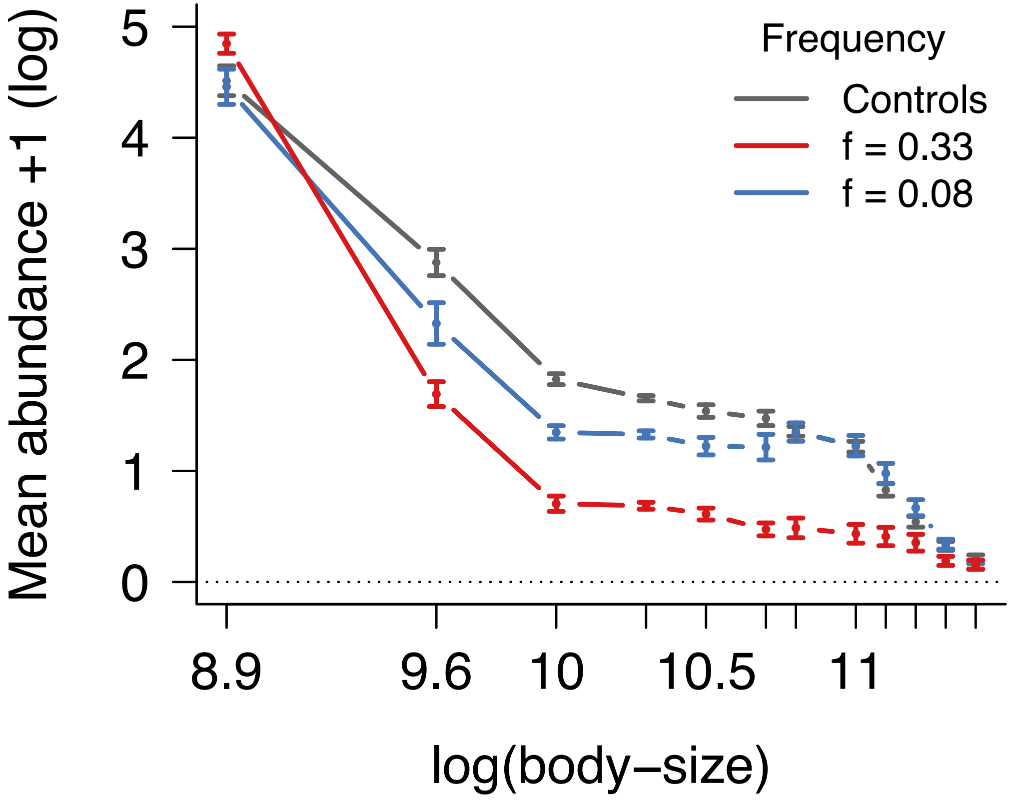
Disturbance = boiling a fraction of the microcosms
4 frequencies:
Every 3, 6, 9 or 12 days

- 12 protist species with body-size between 10 µm and 1 mm
- 6 replicate per treatment + 8 controls
- Daily measurements during 21 days (density, size)

2.2. Test the predictions, quantitatively

How do disturbance frequency impact size-abundance relationships?

Experimental results



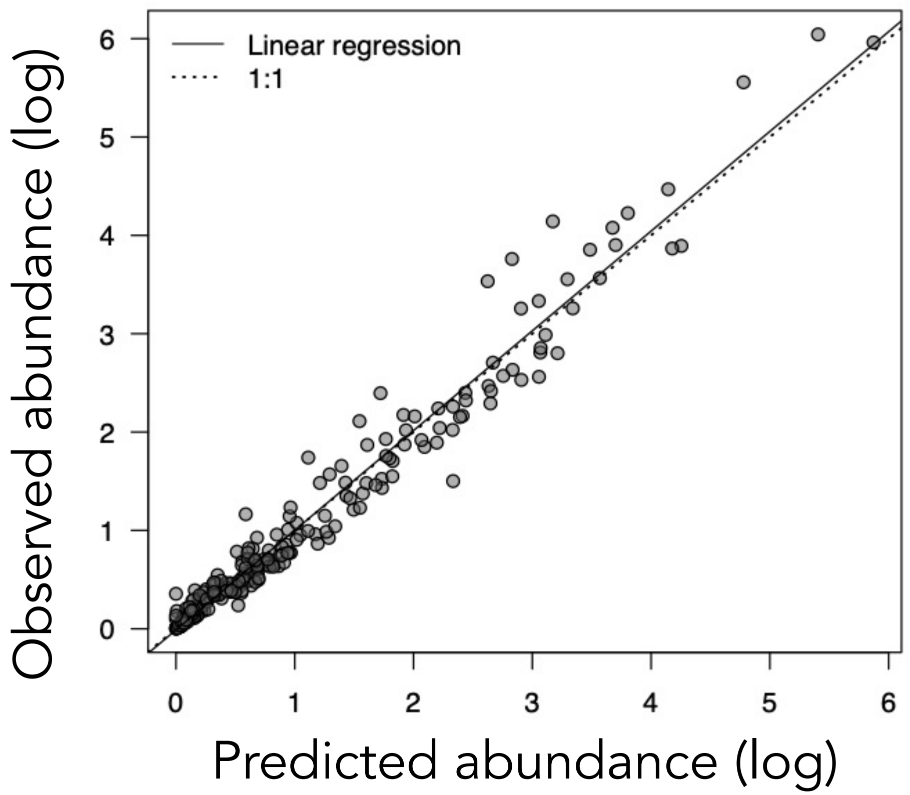
Jacquet *al.* (2020)

2.2. Test the predictions, quantitatively

How much is the model right?

Parameterization:

- relationship between growth rate *r* and body-size *M* for the specific system
- K = equilibria in controls

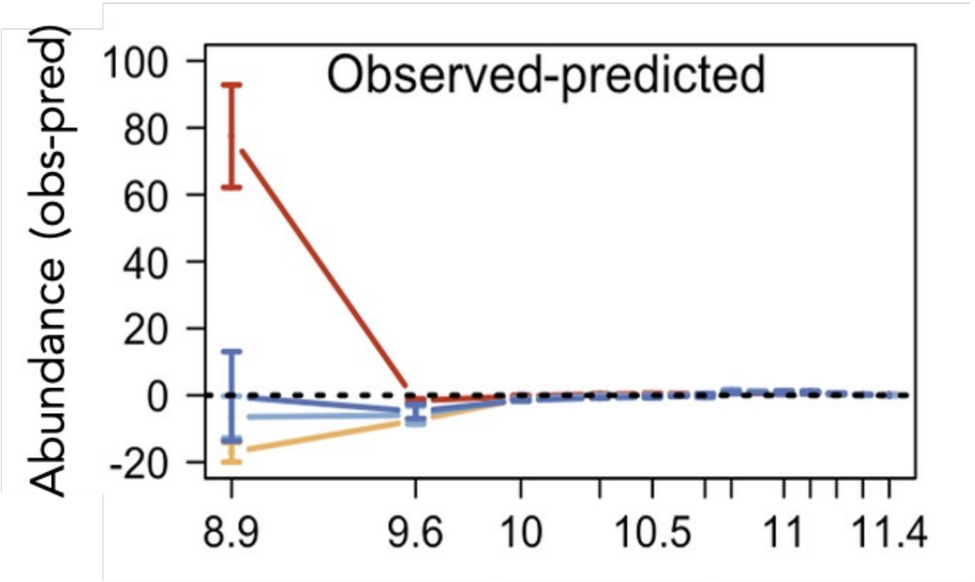


$y = 1.01x - 0.01$ ($R2 = 0.96$, $P < 0.001$)

All size classes and disturbance treatments together

2.2. Test the predictions, quantitatively

Where is the model wrong?



log body size

The model does not predict well the responses of small species

Interpretation: disturbances induced predation/competition release

→ put inter-specific interactions in the model

2. Test the predictions

Key questions:

- What evidence is needed to strongly support or refute the theory?
- What assumptions the model makes?
- Are experimental/natural conditions consistent with the model assumptions?
- How can the theory inform experimental design?

3. Test model assumptions

Example: Evolution is slower than ecology (Losos *et al.* 1997)

3. Test model assumptions

Example: Diversity destabilize ecological communities (May 1972, Pimm 1984)

Model hypothesis:

- species interact at random
- interaction strengths are picked from a normal distribution

3. Test model assumptions

Example: Diversity destabilize ecological communities (May 1972, Pimm 1984)

Model hypothesis:
- species interact at random
- interaction strengths are picked from a normal distribution

Interactions are **not** distributed randomly in real communities (De Ruiter et 1995)

→ **Interactions can stabilize ecological communities**

4. Use mathematical equations to estimate parameters

Model fitting

Searching for the parameters that optimize the match between model predictions and observed data → **criteria of best fit**

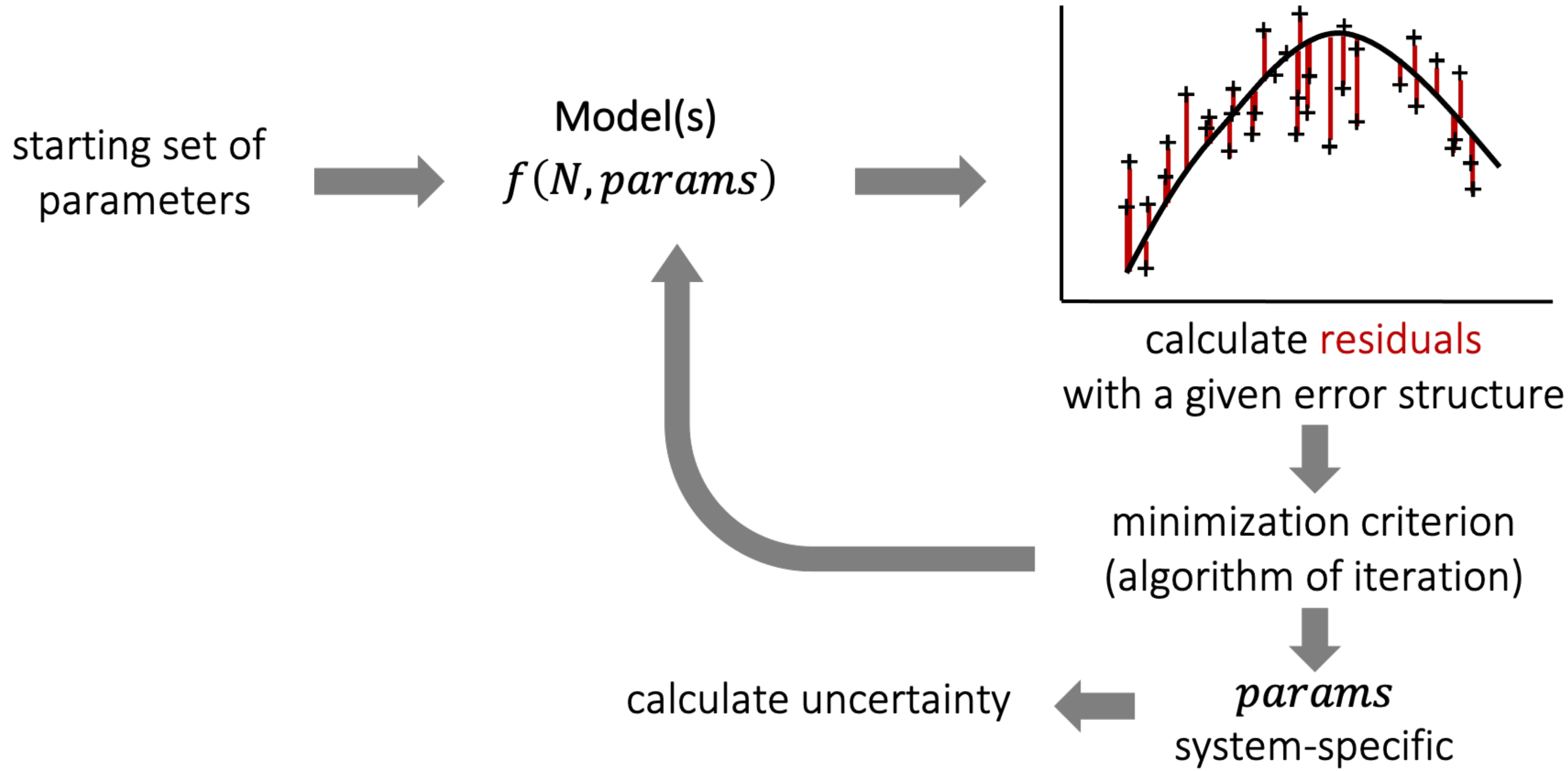
Requirements:

1. data (samples, observations)
2. selecting a model structure suitable for what you want to estimate
3. Choosing residual error structure → how predictions of the model will differ from observations when compared

4. Use mathematical equations to estimate parameters

Model fitting

Searching for the parameters that optimize the match between model predictions and observed data → **criteria of best fit**



4. Use mathematical equations to estimate parameters

Classical criteria of best fit to minimize

1. Sum of squared residuals
2. Negative log-likelihood (or maximizing the product of all likelihoods)
3. Bayesian method using prior probabilities (most likely parameters)

4.1 Minimization of the sum of squared residuals

The best model is the one with parameters that minimize the sum of squares

$$ssq = \sum_{i=1}^{n} \left(O_i - \hat{E}_i \right)^2$$

Observed value *i*

Expected value for a given observation *i*

4.2 Negative log-likelihood

The higher the probability that the model predicts the observed data, the higher the likelihood

Total likelihood

Probability density (likelihood) for each observation given the parameter values

$$L = \prod_{i=1}^{n} L\left(x_i|\theta\right)$$

$$-veLL = -\sum_{i=1}^{n} \log\left(L\left(x_i|\theta\right)\right)$$

Negative Log Likelihood

4.3 Bayesian statistics

Put knowledge in statistical models

4.3 Bayesian statistics

Frequentists
(maximum likelihood)

- Assume parameters are fixed
- Point estimate of the parameters
- No use of the knowledge

VS

Bayesians

- Assume that parameters are not fixed but have a fixed unknown probability distribution
- Uncertainty on the parameters
- Use of the knowledge

But same objective: estimating parameter θ with available data

4.3 Bayesian statistics

Baye's theorem: Let $\mathcal{M}$ be the model with p parameters: $\Theta = (\theta_1, \ldots, \theta_p)$.

Likelihood = probability of observing the data under a certain model

Prior = expertise or ecological information on the parameters

$$\mathcal{P}(\Theta|\text{data}) = \frac{\mathcal{P}(\text{data}|\Theta)\mathcal{P}(\Theta)}{\mathcal{P}(\text{data})}$$

Posterior = distribution of the parameters given the data (**what you know after having seen the data**)

$$\int \mathcal{P}(\text{data}|\Theta)\mathcal{P}(\Theta)d\Theta$$

Normalisation constant : difficult/impossible to calculate

→ simulation methods (MCMC)

4.3 Bayesian statistics: example

Think about the globe. **How much water of the surface is covered by water ?**

Imagine you toss the globe up in the air. When you catch it, you will record whether or not the surface under your right index finger is water or land.

You do that **6** times and get the following sample (W for water, L for land):

W L W W W L

We can formalize in this example:

W ~ Binomial (N,p) with N = W+L

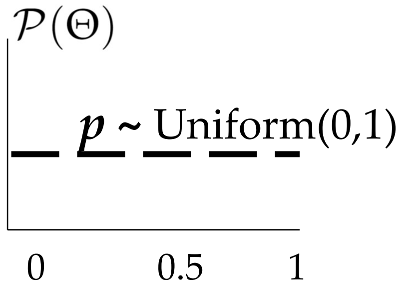
Likelihood $= \mathrm{P}(\text{data} \mid \Theta) = \mathrm{P}(\mathrm{W,L} \mid \boldsymbol{p}) = \binom{W+L}{W} p^{W}(1-p)^{L}$

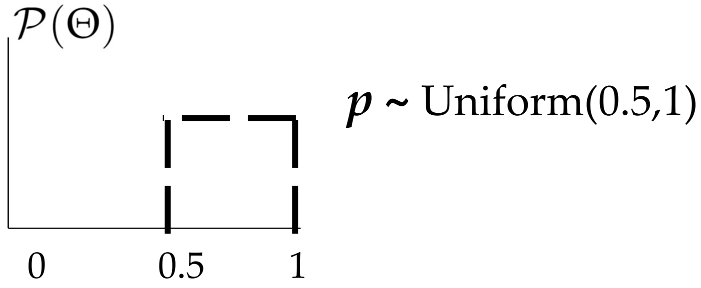
From McElreath (2020)

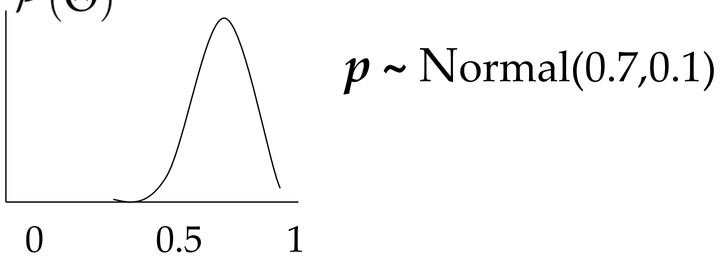
4.3 Bayesian statistics: priors

1. **Define a prior for *p***, the fraction of the globe covered by water

→ *a priori* knowledge





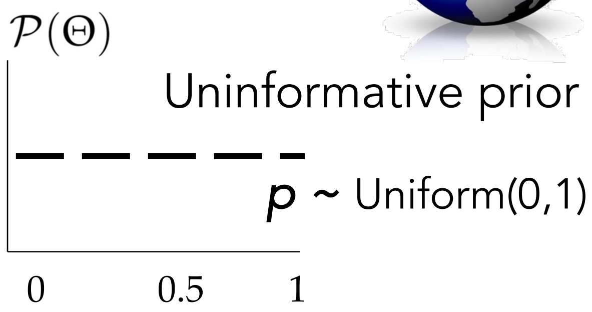


Uninformative prior

Prior with some knowledge: "I know that there is at least as much as land as water, if not more"

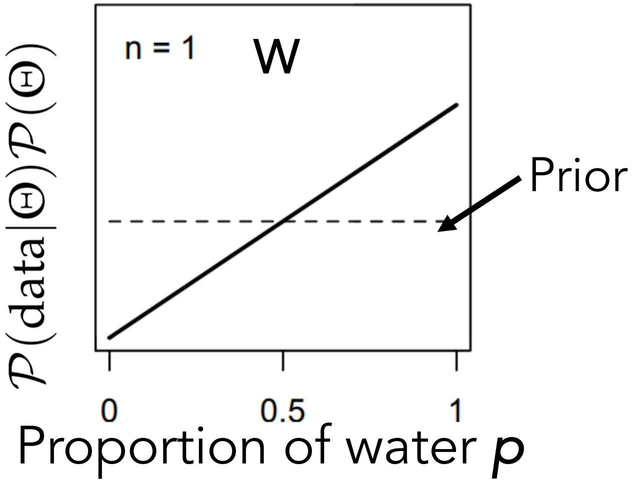
Similar as before, but more likely around 0.7

4.3 Bayesian statistics: bayesian update



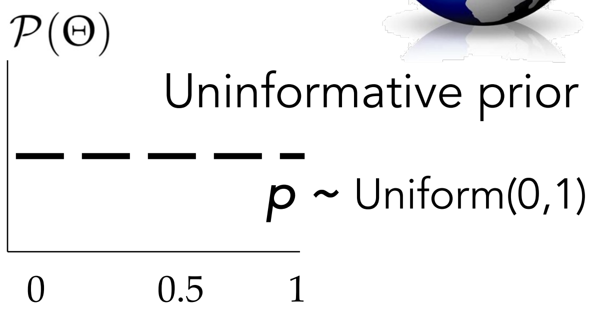
1. **Define a prior for *p*** (fraction of the globe covered by water)

2. **Update** with the first observation **W**:

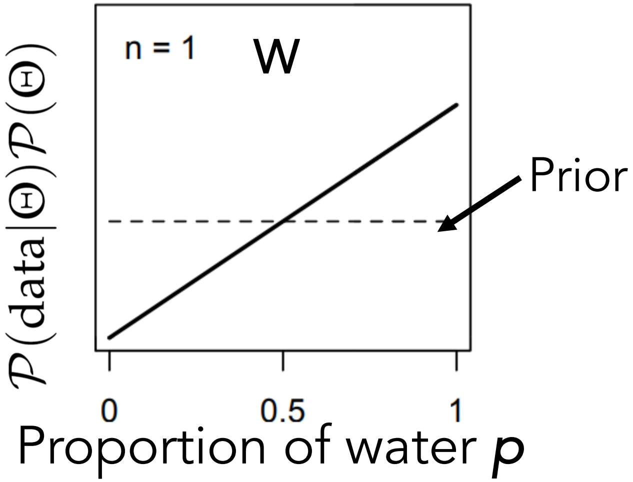


4.3 Bayesian statistics: bayesian update

1. **Define a prior for *p*** (fraction of the globe covered by water)

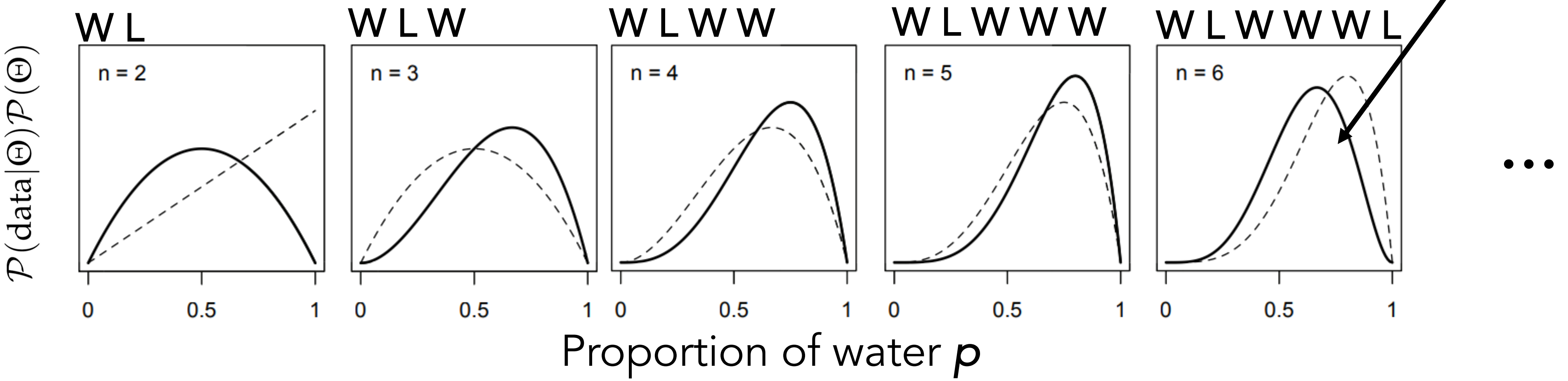


2. **Update** with the first observation **W**:



3. **Repeat** for all the observations: **we learn from the data:**

We approximate the posterior of p



4.3 Bayesian statistics: Approximate Bayesian Computing (ABC)

What if we don't know the likelihood?

We don't know the likelihood
but we want the posterior

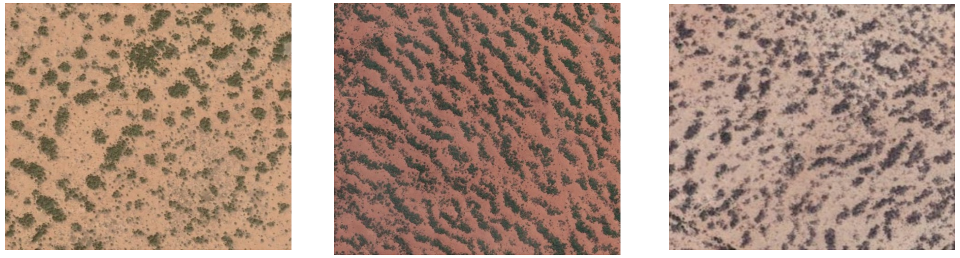
$$\mathcal{P}(\Theta|\text{data}) = \frac{\mathcal{P}(\text{data}|\Theta)\mathcal{P}(\Theta)}{\mathcal{P}(\text{data})}$$

Solution: approximate the posterior using simulations

Hypothesis: the data follow the model

4.3 Bayesian statistics: Approximate Bayesian Computing (ABC)

Example: stochastic spatial vegetation model

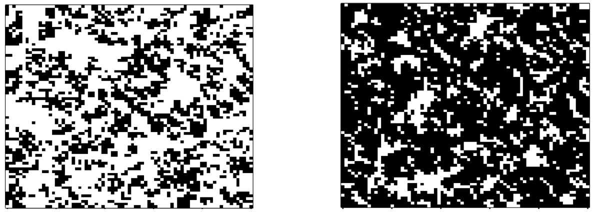


Observation of vegetation in arid ecosystem

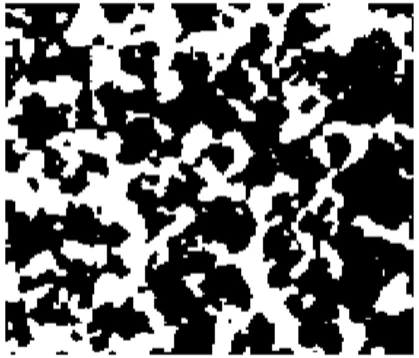
A model with 2 parameters (p, q):

p: reproduction

q: local positive feedback



Black = vegetation
White = empty soil



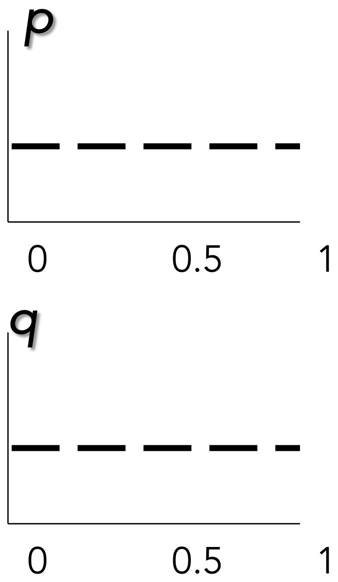
$p, q = ?$

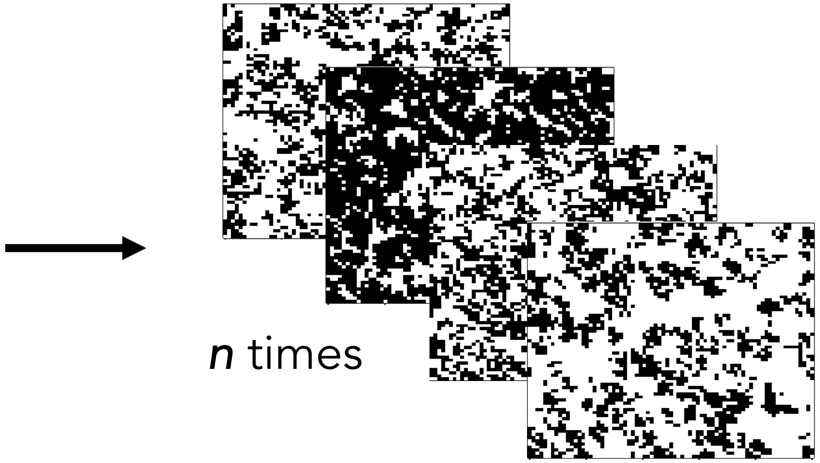
Hypothesis: the data follow the model

4.3 Bayesian statistics: Approximate Bayesian Computing (ABC)

1 Draw parameters from prior distributions

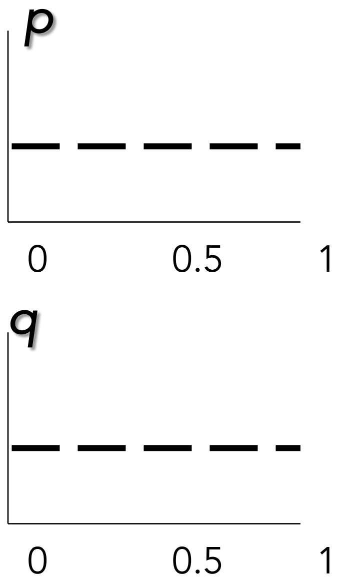
2 Simulate the model under the prior



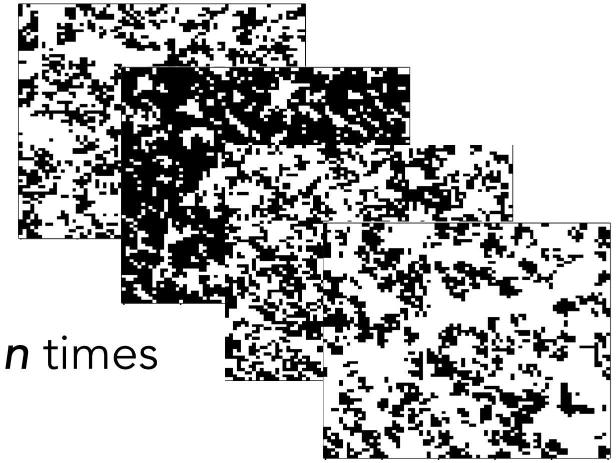


4.3 Bayesian statistics: Approximate Bayesian Computing (ABC)

1 Draw parameters from prior distributions

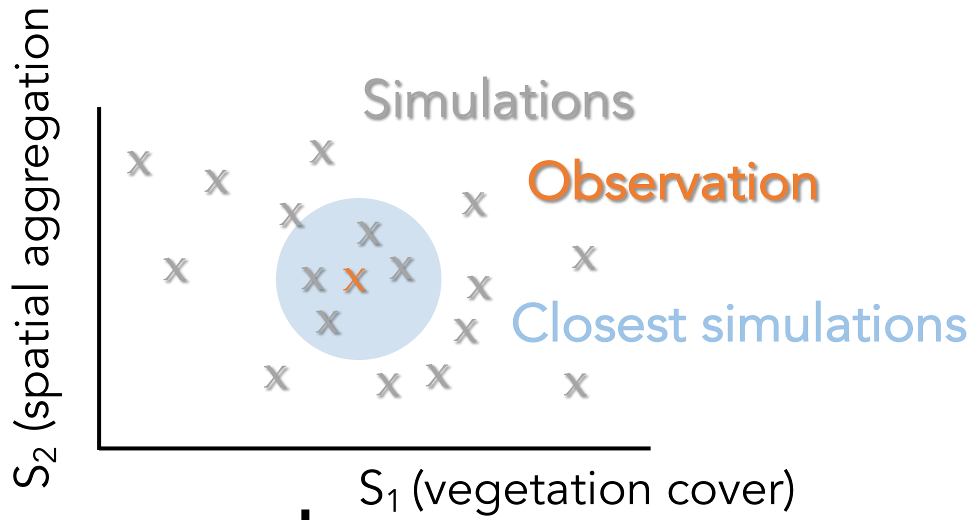


2 Simulate the model under the prior



3 Compute summary statistics (S_1, S_2) on both observation and simulations.

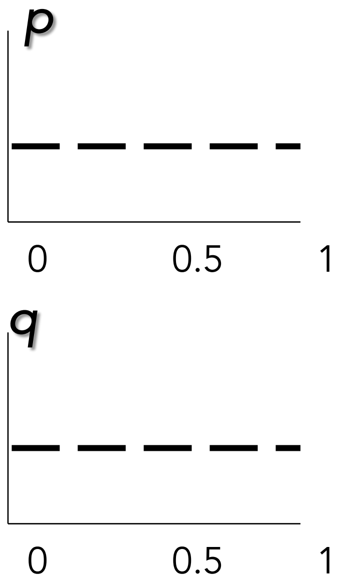
Project on the summary statistic space



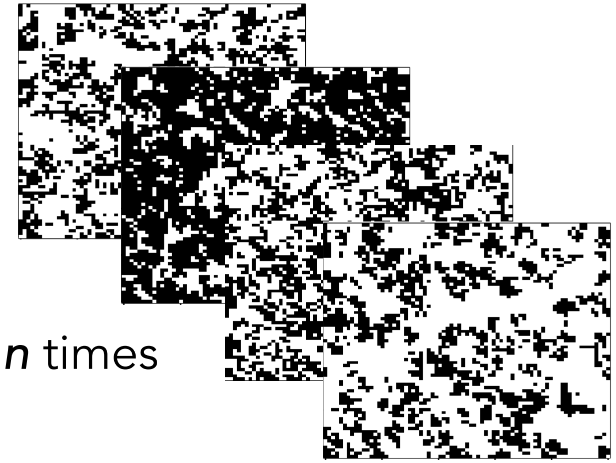
4 Select the k-closest simulations

4.3 Bayesian statistics: Approximate Bayesian Computing (ABC)

1 Draw parameters from prior distributions

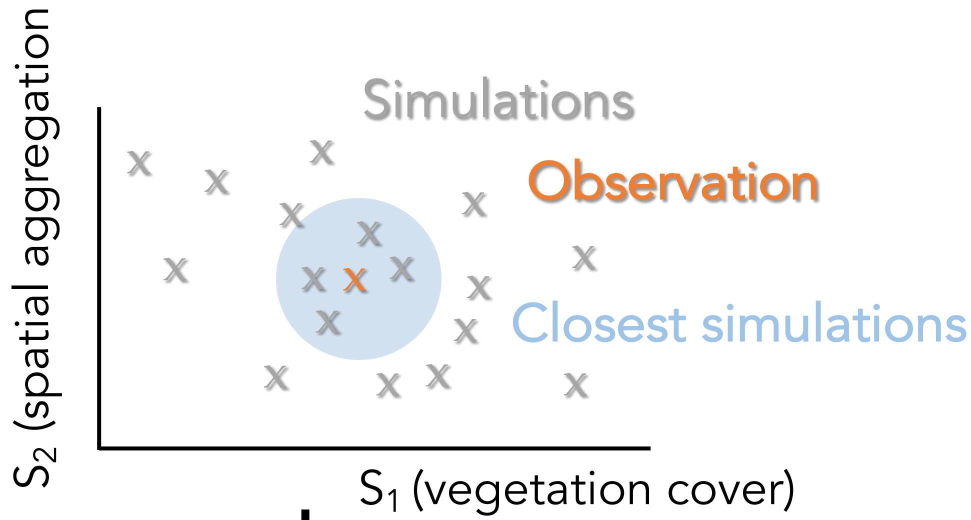


2 Simulate the model under the prior



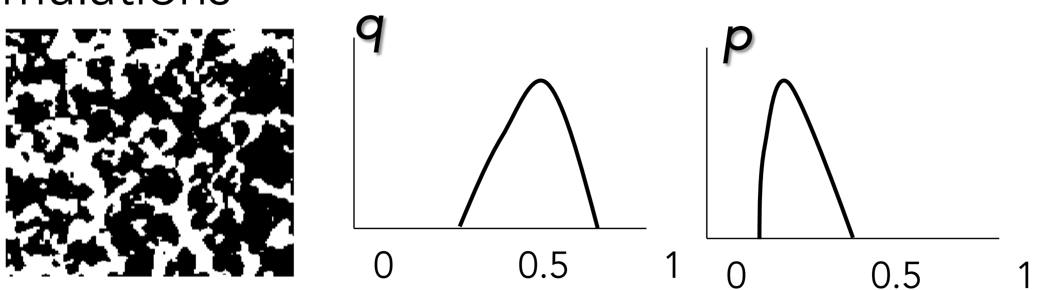
3 Compute summary statistics (S_1, S_2) on both observation and simulations.

Project on the summary statistic space



4 Select the k-closest simulations

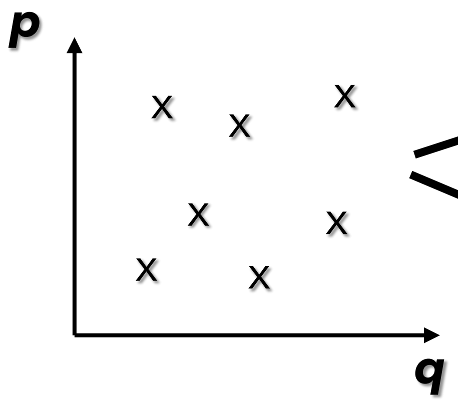
5 Approximate the posterior of (*p*, *q*) of the observation from the distribution of parameters of the accepted simulations



4.3 Bayesian statistics: Approximate Bayesian Computing (ABC)

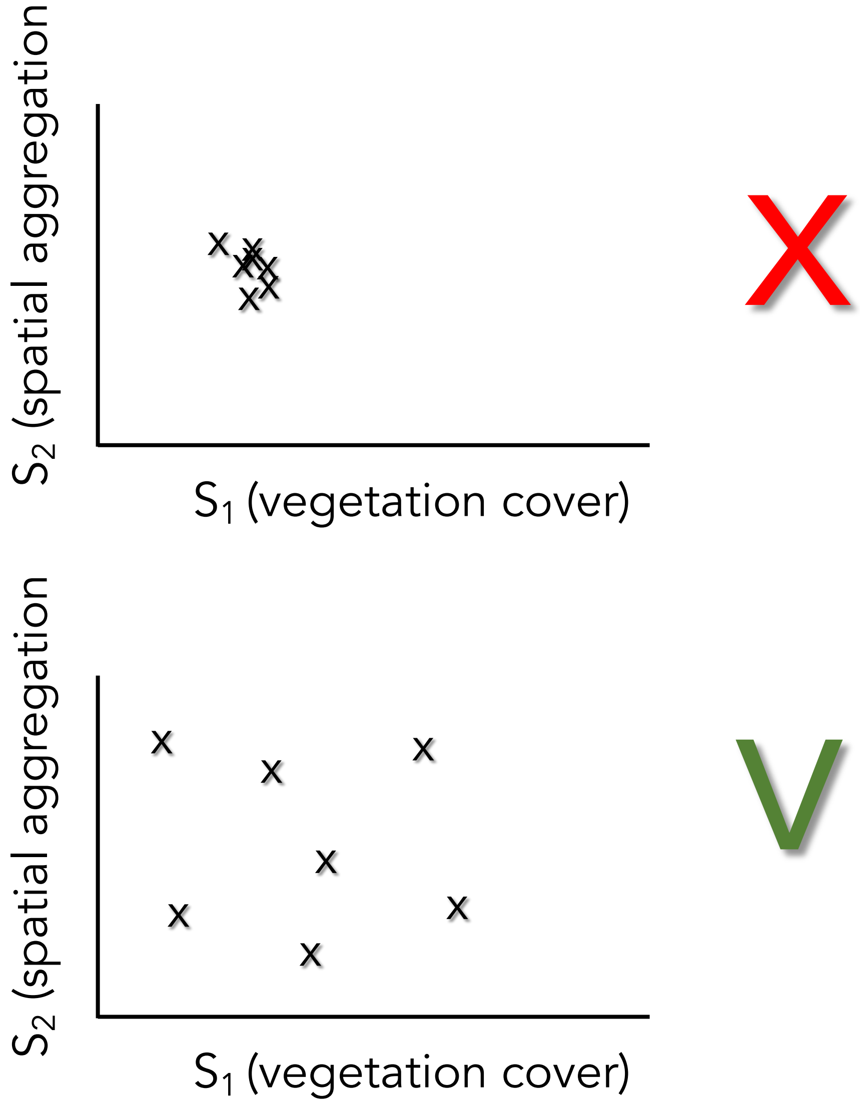
Conditions for ABC: model identifiability

Any combination of parameter gives
a unique combination of summary statistics



Joint priors

p = reproduction
q = local positive feedback



Linking theory and data

How to choose the right approach? It depends on…

- your objectives
- the structure of the models, number of parameters
- the nature of the data (time series, spatial data, interaction networks)

Linking theory and data

How to choose the right approach? It depends on…

- your objectives
- the structure of the models, number of parameters
- the nature of the data (time series, spatial data, interaction networks)

Practices on how to do it in the next days

Your turn on Friday!