A data-intensive assessment of the species-abundance distribution.

Elita Baldridge

Open science

- Code:
  - github.com/embaldridge
  - github.com/weecology
- Data: figshare.com
- Twitter: @elitabaldridge

figshare
credit for all your research

Open Source Initiative

Feel free to:

Copy, share, adapt, or re-mix;

Photograph, film, or broadcast;

Blog, live-blog, or post video of;

Provided that:

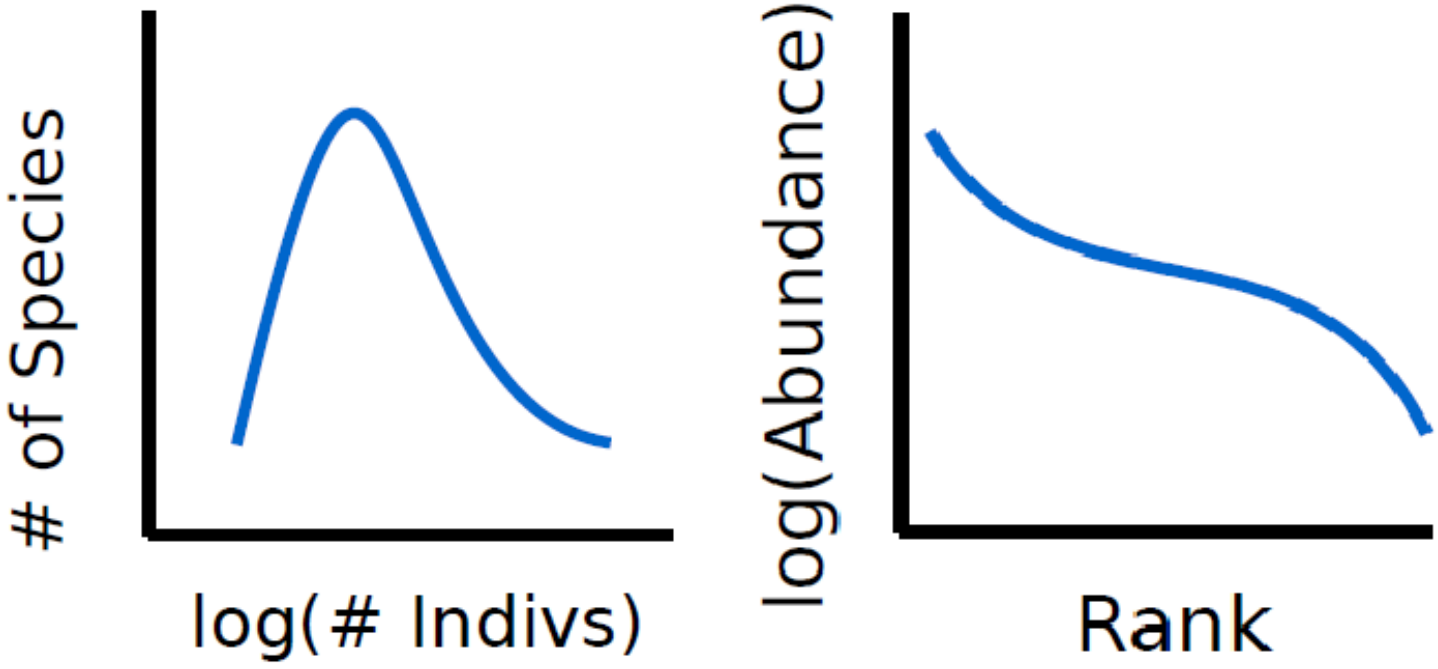
You attribute the work to its author and respect the rights and licenses associated with its components.

Commonness & rarity

”Who can explain why one species ranges widely, and is very numerous, and why another allied species has a narrow range and is rare? Yet these relations are of the highest important, for they determine the present welfare and, as I believe, the future success and modification of every inhabitant of this world.”
Darwin, 1859.

Species abundance distribution

- Describes the distribution of commonness & rarity of species.
- Exhibits a hollow curve distribution.



Macroecology

One approach to studying ecological patterns and processes.

- Data intensive.
- Large scales.
  - Spatial
  - Temporal
  - Taxonomic
- Search for generality.

Signal & Noise

Macroecology

Pattern

↓

Process

↓

Prediction

Macroecology

Challenges of macroecology

- Studies performed with a limited number of large datasets.
- Lack of identification of pattern generating mechanisms.

Macroecology

Best practice recommendations

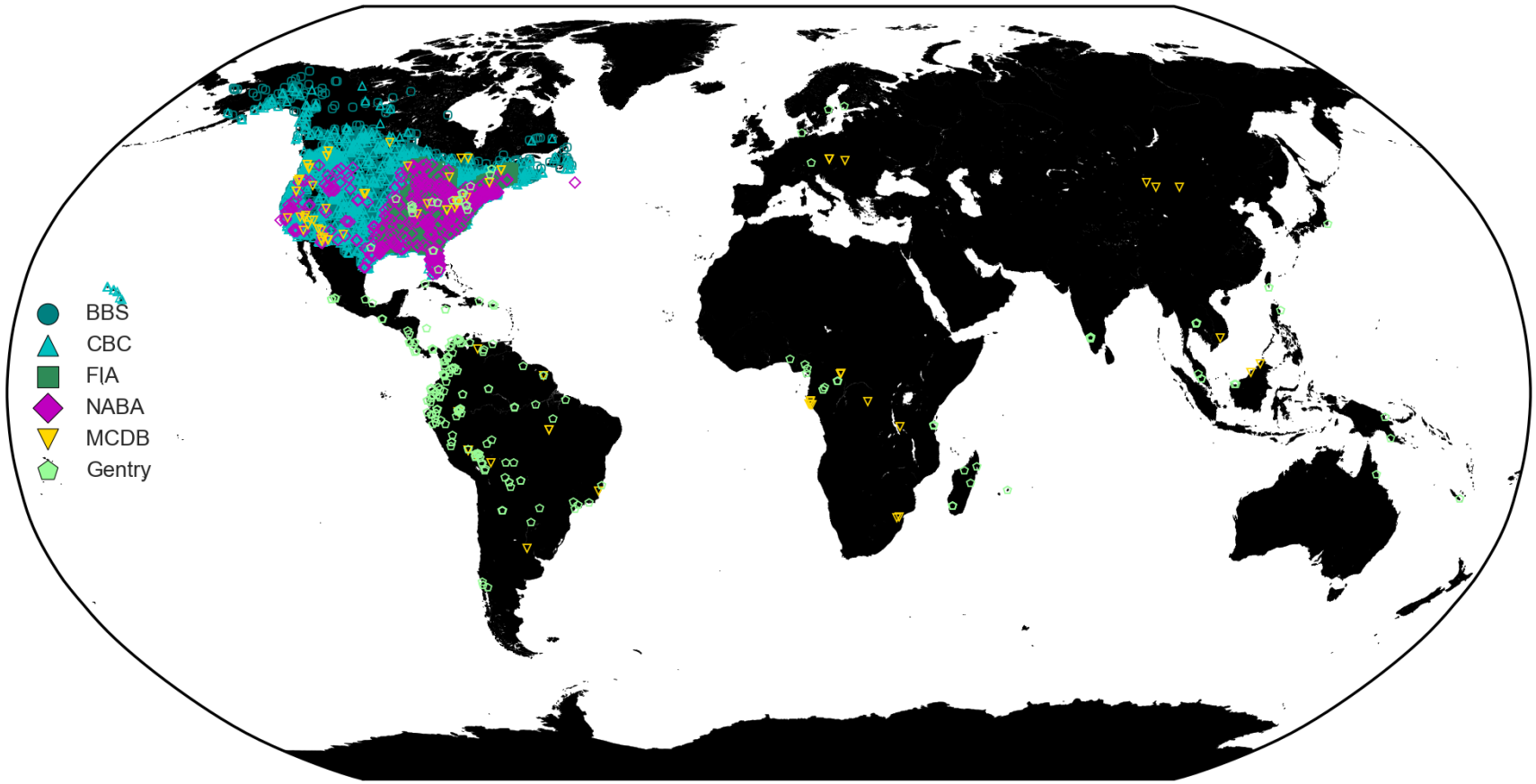
- Test patterns with multiple taxonomic groups/ecosystems.
- Simultaneous testing of competing models and model predictions with a consistent statistical approach.

The Rules of Ecoinformatics

Garbage in, garbage out.

- All data are good, not all data are appropriate.
- Fit the data to the question.

Data



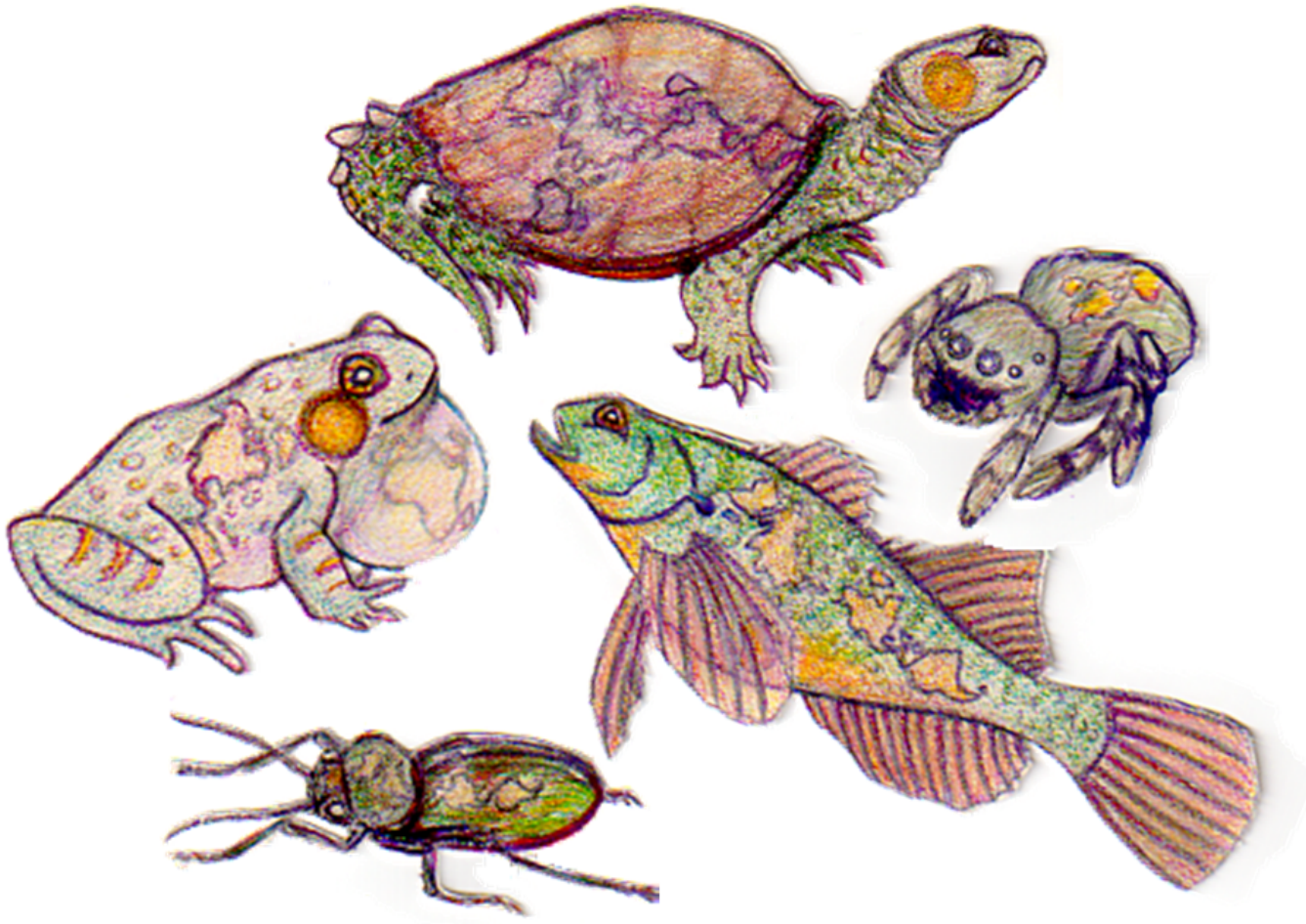
Data

Major macroecological datasets

- Largely terrestrial
- Largely North American
- Many publicly available, some not.

Lots of data in the literature.

Data

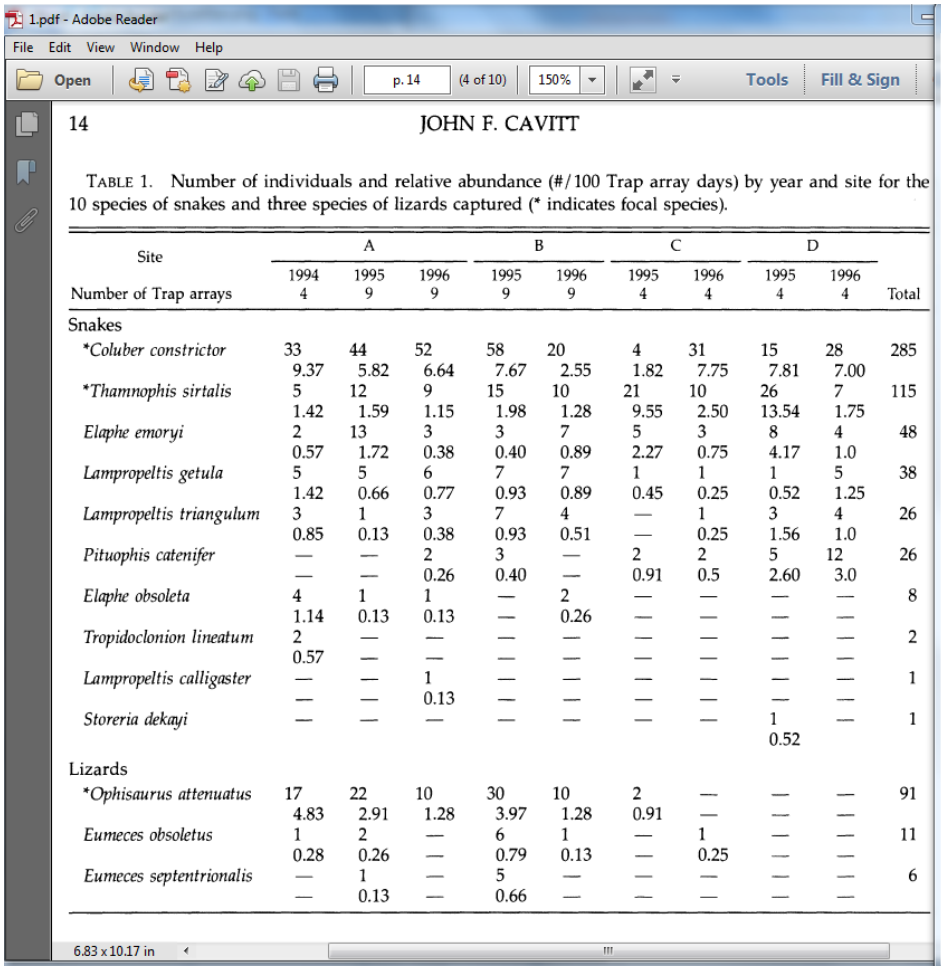


Data

Inclusion criteria:

- Quantitative abundances
- Animals
- Complete sampling
- Must not be heavily summarized or processed
- High degree of taxonomic resolution
- Observational

Data wrangling



14 JOHN F. CAVITT

TABLE 1. Number of individuals and relative abundance (#/100 Trap array days) by year and site for the 10 species of snakes and three species of lizards captured (* indicates focal species).

Site	A			B		C		D		
Number of Trap arrays	1994 4	1995 9	1996 9	1995 9	1996 9	1995 4	1996 4	1995 4	1996 4	Total
Snakes										
Coluber constrictor	33 9.37	44 5.82	52 6.64	58 7.67	20 2.55	4 1.82	31 7.75	15 7.81	28 7.00	285
Thamnophis sirtalis	5 1.42	12 1.59	9 1.15	15 1.98	10 1.28	21 9.55	10 2.50	26 13.54	7 1.75	115
Elaphe emoryi	2 0.57	13 1.72	3 0.38	3 0.40	7 0.89	5 2.27	3 0.75	8 4.17	4 1.0	48
Lampropeltis getula	5 1.42	5 0.66	6 0.77	7 0.93	7 0.89	1 0.45	1 0.25	1 0.52	5 1.25	38
Lampropeltis triangulum	3 0.85	1 0.13	3 0.38	7 0.93	4 0.51	— —	1 0.25	3 1.56	4 1.0	26
Pituophis catenifer	— —	— —	2 0.26	3 0.40	— —	2 0.91	2 0.5	5 2.60	12 3.0	26
Elaphe obsoleta	4 1.14	1 0.13	1 0.13	— —	2 0.26	— —	— —	— —	— —	8
Tropidoclonion lineatum	2 0.57	— —	— —	— —	— —	— —	— —	— —	— —	2
Lampropeltis calligaster	— —	— —	1 0.13	— —	— —	— —	— —	— —	— —	1
Storeria dekayi	—	—	—	—	—	—	—	1 0.52	—	1
Lizards										
Ophisaurus attenuatus	17 4.83	22 2.91	10 1.28	30 3.97	10 1.28	2 0.91	— —	— —	— —	91
Eumeces obsoletus	1 0.28	2 0.26	— —	6 0.79	1 0.13	— —	1 0.25	— —	— —	11
Eumeces septentrionalis	— —	1 0.13	— —	5 0.66	— —	— —	— —	— —	— —	6

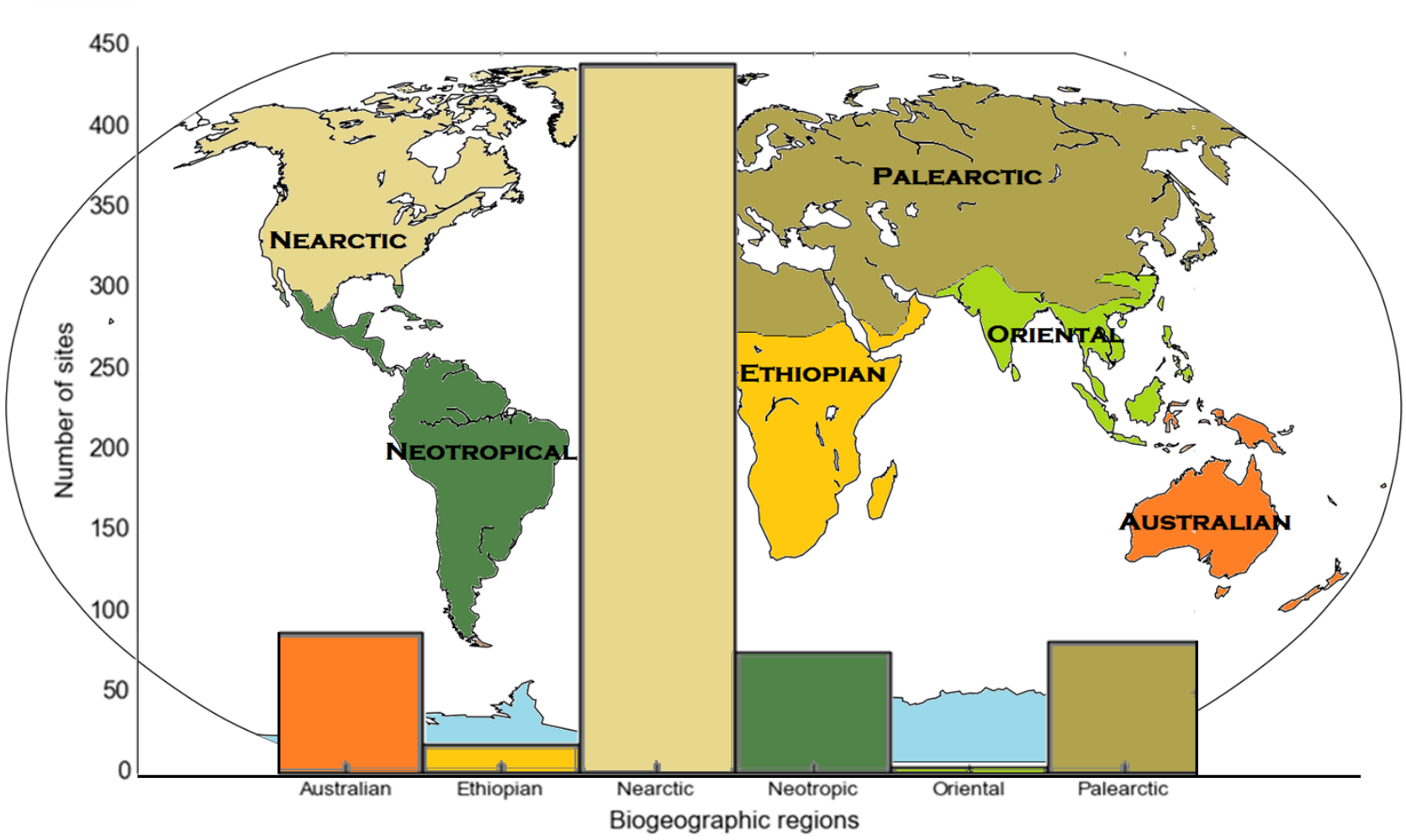
Cavitt 2000.

Data

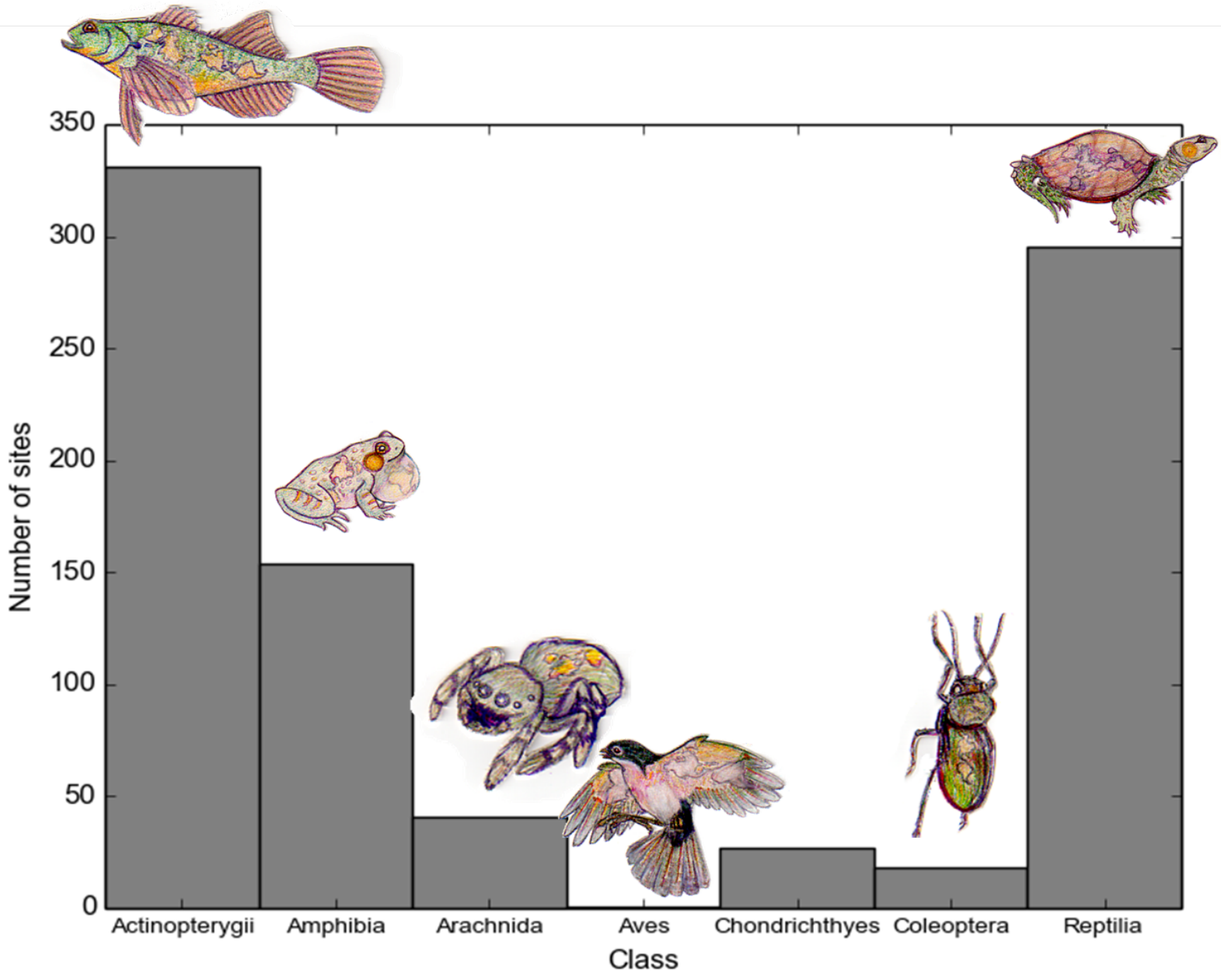
Variables collected
Class
Family
Genus
Species (Specific epithet)
Relative abundance
Abundance
Collection Year, starting
Collection Year, ending
Site Name
Biogeographic region
Site notes

Table : List of variables collected.

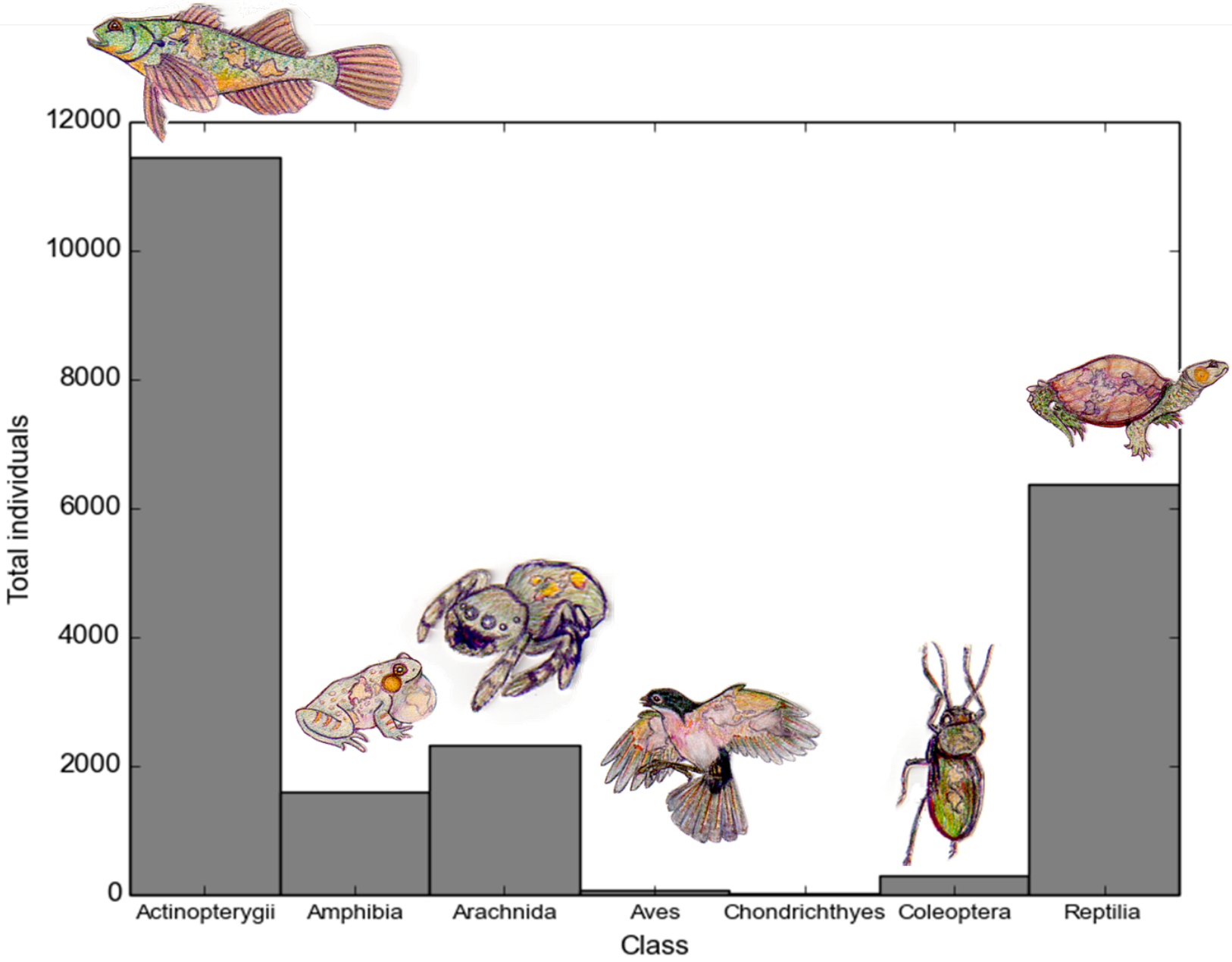
Data


NEARCTIC
PALEARCTIC
ORIENTAL
ETHIOPIAN
NEOTROPICAL
AUSTRALIAN
Number of sites
0
50
100
150
200
250
300
350
400
450
Australian
Ethiopian
Nearctic
Neotropic
Oriental
Palearctic
Biogeographic regions

Data


350
300
250
200
150
100
50
0
Number of sites
Actinopterygii
Amphibia
Arachnida
Aves
Chondrichthyes
Coleoptera
Reptilia
Class

Data


12000
10000
8000
6000
4000
2000
0
Total individuals
Actinopterygii
Amphibia
Arachnida
Aves
Chondrichthyes
Coleoptera
Reptilia
Class

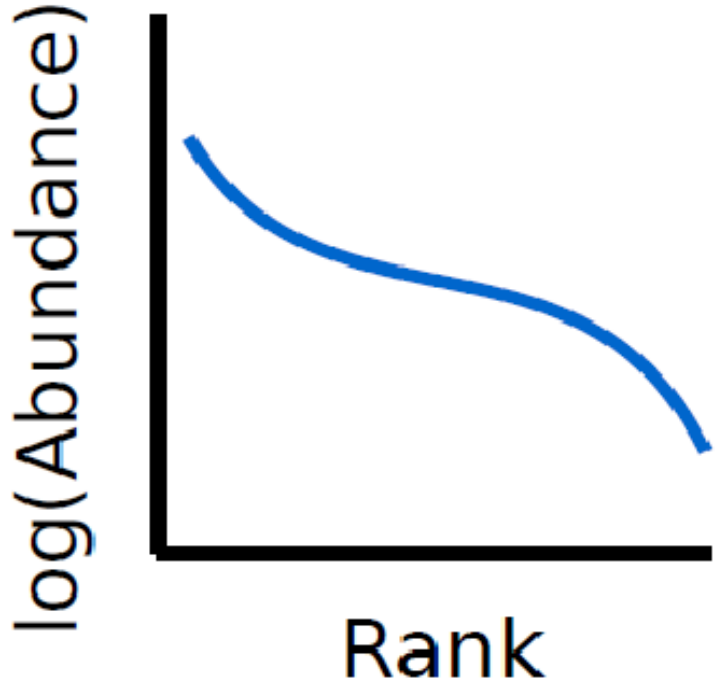
Data Availability

Public & open access through figshare.
EcoData Retriever importable.

```
(http://figshare.com)
(http://www.ecodataretriever.org)
sad_data =
ecoretriever::fetch('MiscAbundanceDB')
```

Commonness & rarity

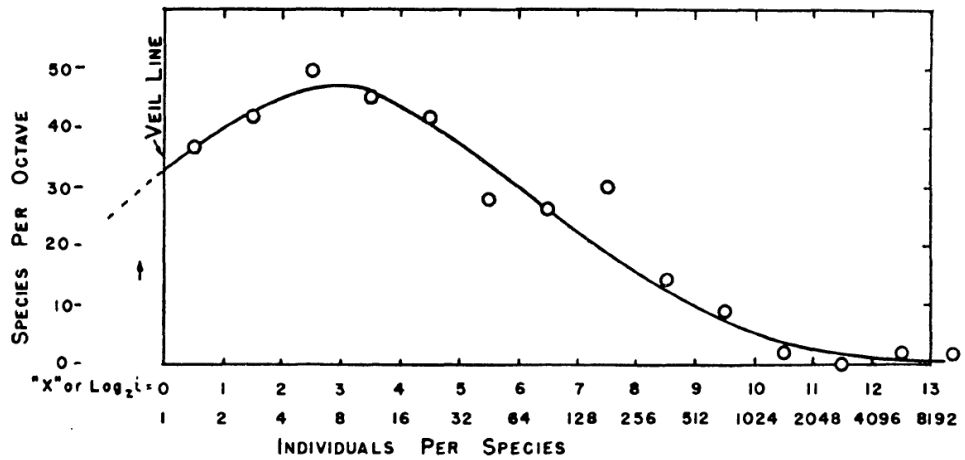
The species abundance distribution:



Many models of the species abundance distribution (SAD).

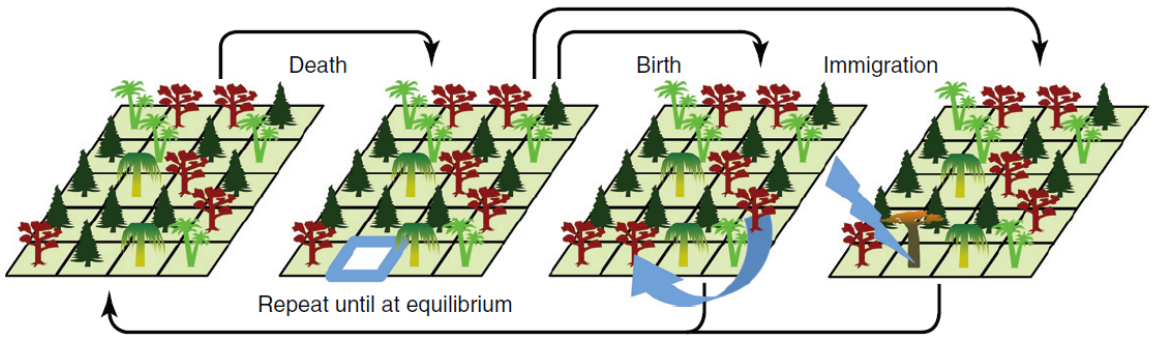
SAD models

Statistical description



Preston 1962a.

Process-based



Rosindell et al. 2011.

SAD Comparisons

Most comparisons of the different models:

- Use only a small subset of available models (typically two).
- Focus on a single ecosystem or taxonomic group
- Fail to use the most appropriate statistical methods.

SAD Comparisons

Selected five models from four classes for comparison.

Model class	Form of the distribution
Purely statistical	Logseries, Poisson lognormal
Branching process	Zipf
Population dynamics	Negative binomial
Niche partitioning	Geometric

Table : After B.J. McGill et al. 2007.

SAD Comparisons

Analysis:

- Model fitting with maximum likelihood estimation.

$$l_x(\theta) = \log h(x) + \log f_\theta(x)$$

SAD Comparisons

Analysis:

- Likelihood based model selection to compare the fits of the different models.
    - Assess the fit of the model without corrections for parameter number or similarity to other models.
    - How well does the model describe the data?

SAD Comparisons

Analysis:

- Model comparison with corrected Aikaike Information Criterion (AICc) weights.
  - Assess the fit of the model correcting for parameter number and small sample size.
  - How well does the model describe the data relative to the number of parameters?

SAD Comparisons

Computational tools:

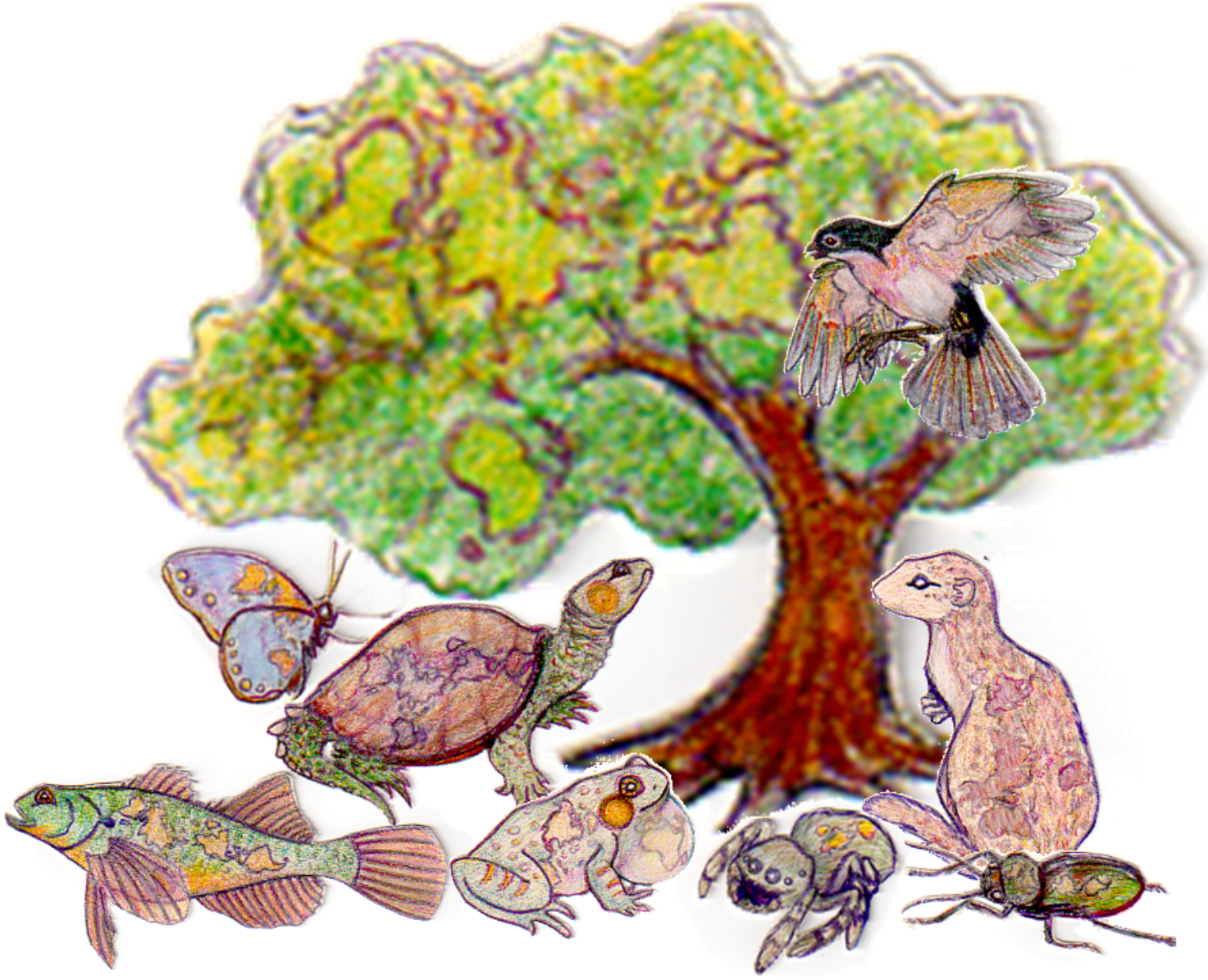
- Model fitting, log-likelihood, and AICc calculations performed with macroecotools Python package.
  (https://github.com/weecology/macroecotools)
- All of the analysis code and the majority of the data is publicly available.
  (https://github.com/weecology/sad-comparison)

SAD Comparisons

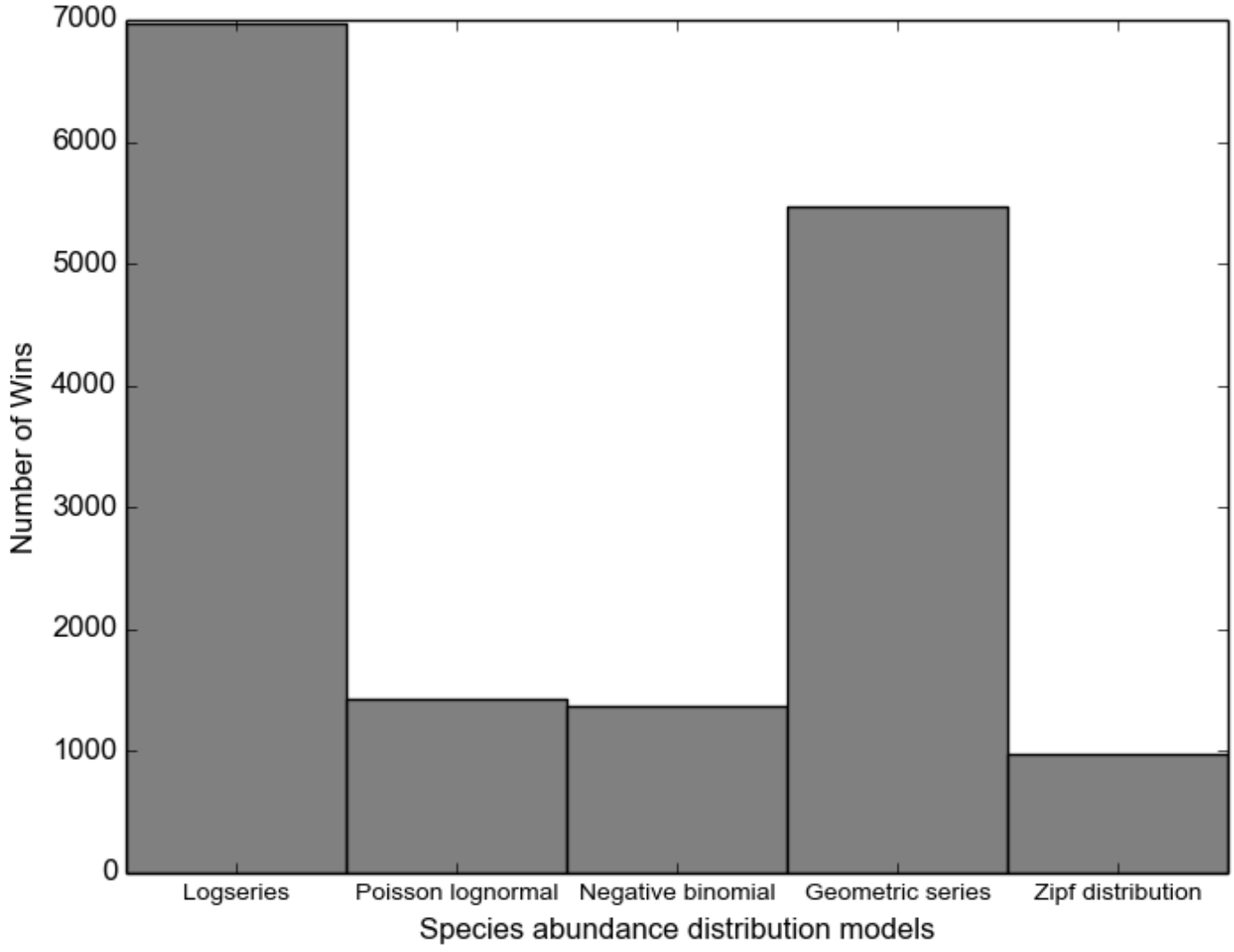
Dataset	Dataset code	Availability	Sites
Gentry's Forest Transects	Gentry	Public	10355
Breeding Bird Survey	BBS	Public	2769
Christmas Bird Count	CBC	Private	1999
Forest Inventory Analysis	FIA	Public	220
N. American Butterfly Count	NABA	Private	400
Actinopterygii, compiled	Actinopterygii	Public	161
Reptilia, compiled	Reptilia	Public	138
Mammal Community Database	MCDB	Public	103
Amphibia, compiled	Amphibia	Public	43
Arachnida, compiled	Arachnida	Public	25
Coleoptera, compiled	Coleoptera	Public	5

Table : Datasets used for species-abundance distribution comparisons. Datasets marked as Private obtained through data requests to the providers with Memorandums of Understanding.

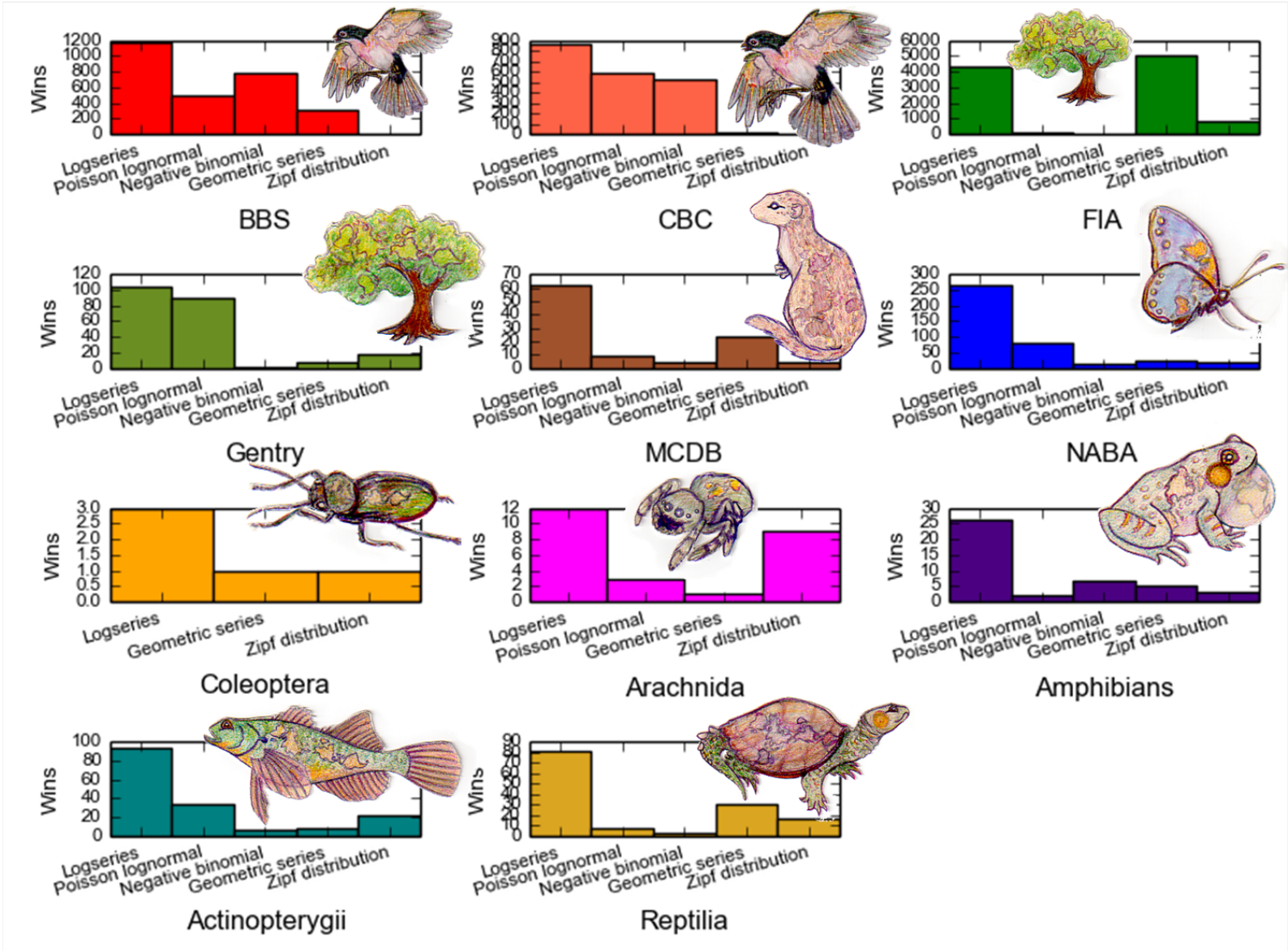
Data



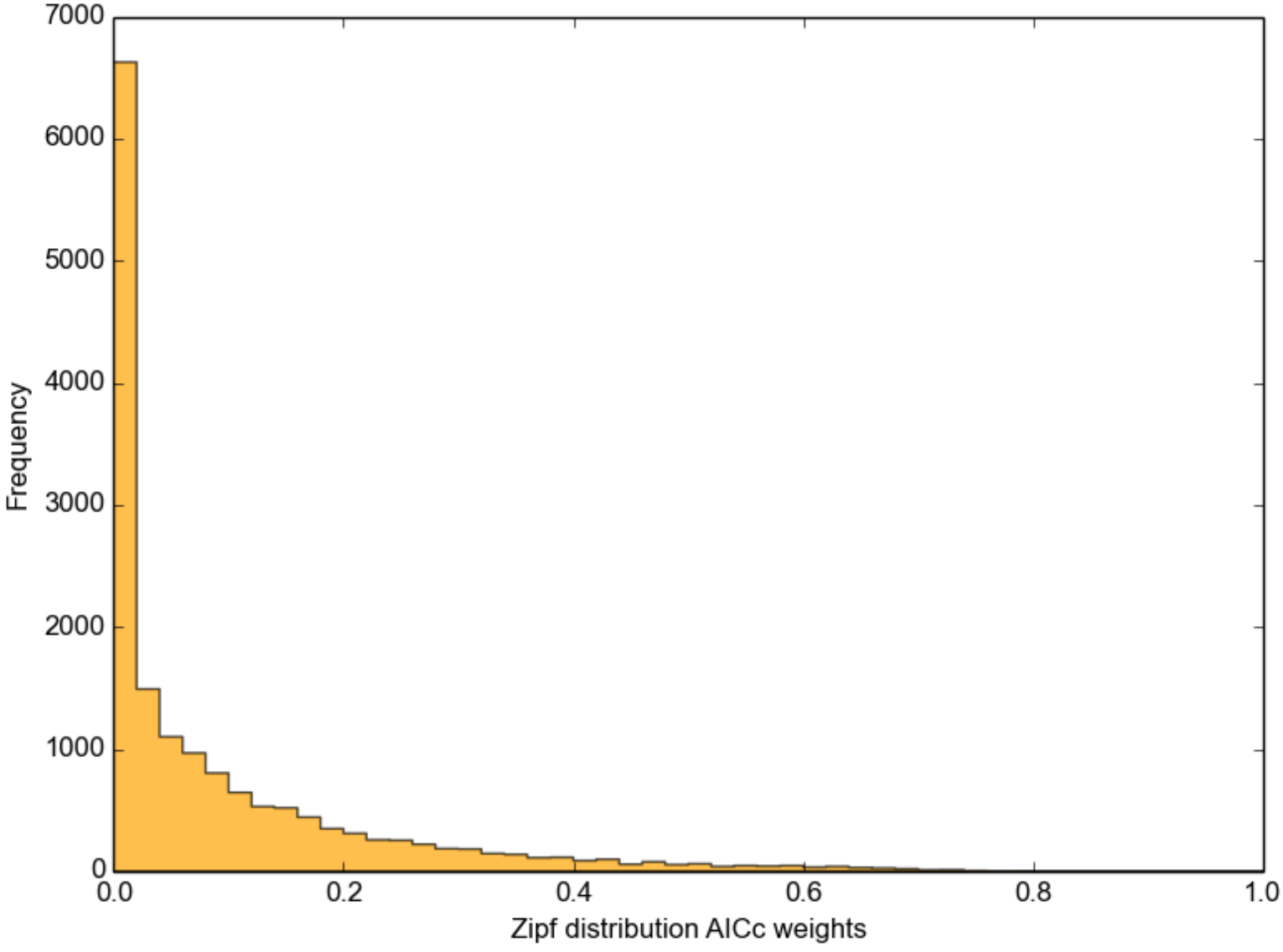
SAD Comparisons



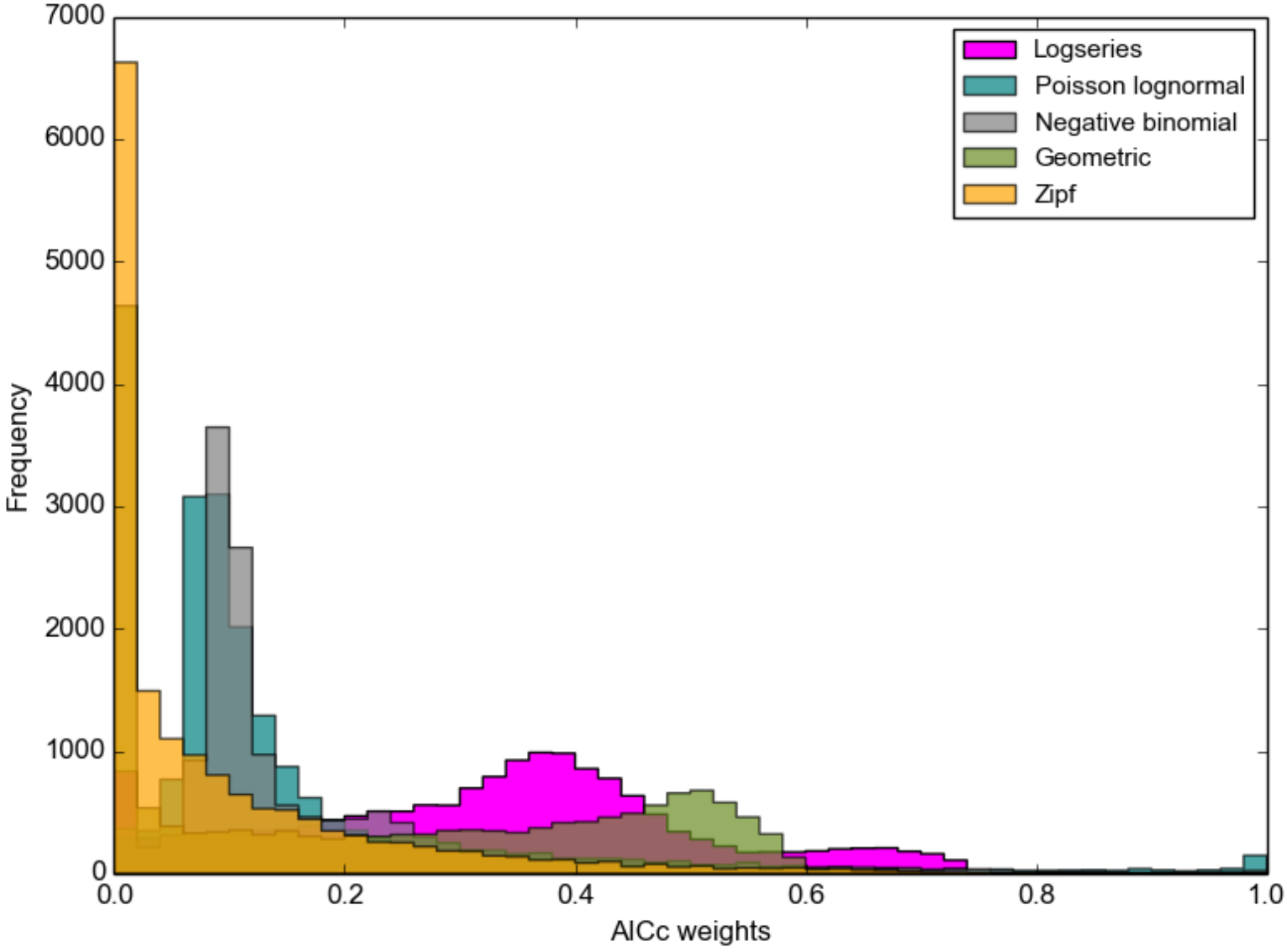
SAD Comparisons



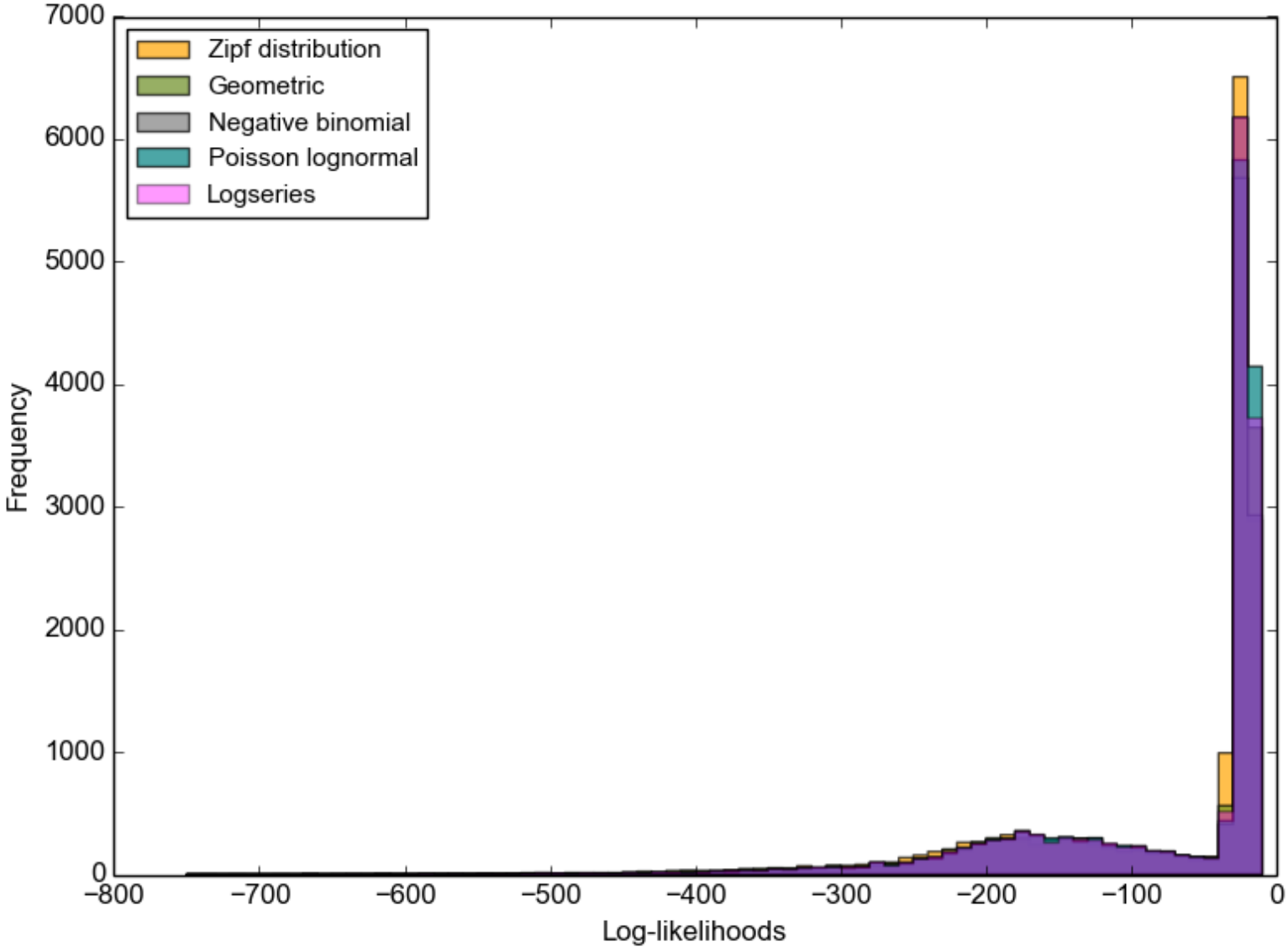
SAD Comparisons



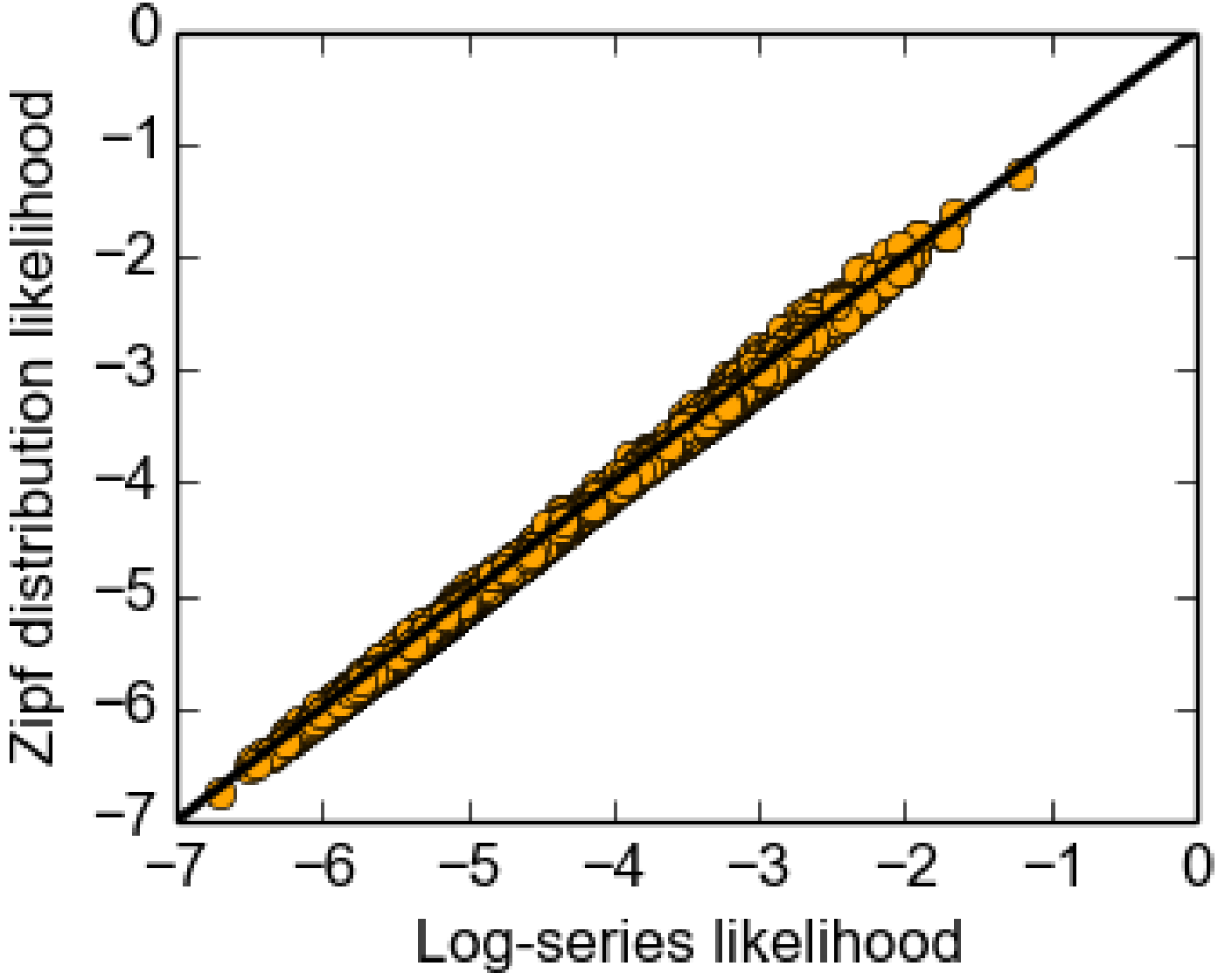
SAD Comparisons



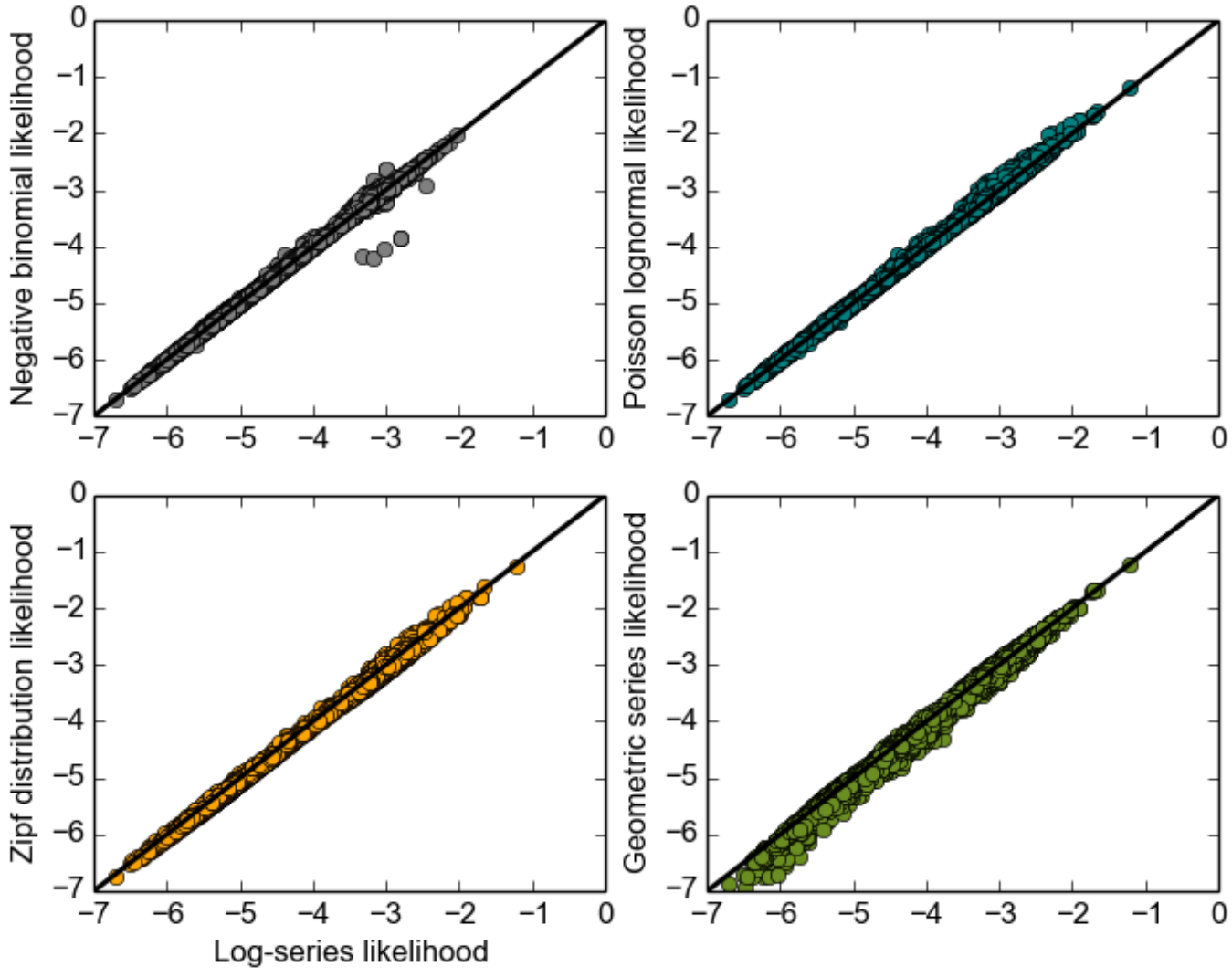
SAD Comparisons



SAD Comparisons



SAD Comparisons



SAD Comparisons

Existing models provide equivalently good absolute fits to empirical data.

- Models with fewer parameters perform better in AIC-based model selection.
- Logseries provides a good naive model for fitting SADs.
    - Produces equivalent likelihoods.
    - Has a single fitted parameter.
    - Easy to fit to empirical data.
    - Best overall model.

SAD Comparisons

Identifying pattern generating mechanisms:

- Compare predictions of different models using multiple macroecological patterns simultaneously.
- Examine scale dependence of pattern.

However, identification of mechanism may not be necessary for prediction.

Neutral Analysis

The unified neutral theory of biodiversity:

- Multiple formulations.
  - Species and individuals are ecologically and demographically equivalent.
  - Stochastic variation in birth, death, immigration, & speciation results in species abundance differences.

Neutral Analysis

Early tests of neutral theory based on comparing the fit of empirical species abundance distributions to the neutral prediction.

Later tests suggested species abundance comparisons were insufficient for a rigorous test of neutrality.

Neutral Analysis

Connolly et al. 2014 identified non-neutral species abundance distributions in marine communities.

- Compared model fits of a non-neutral distribution (Poisson lognormal) to a neutral distribution (negative binomial distribution).

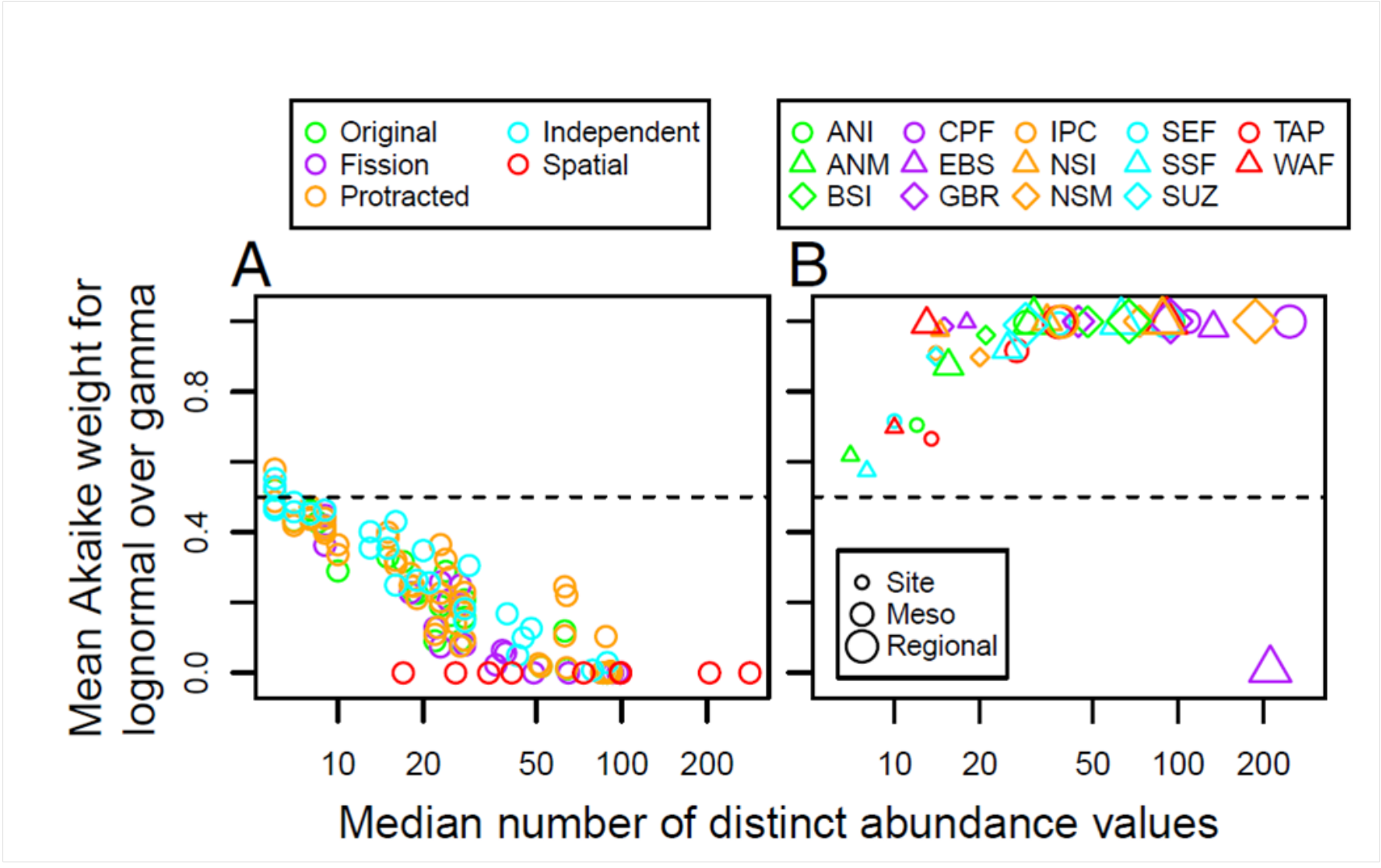
May be a robust method for identifying communities that exhibit non-neutrality.

Connolly et al. 2014.

Neutral Analysis

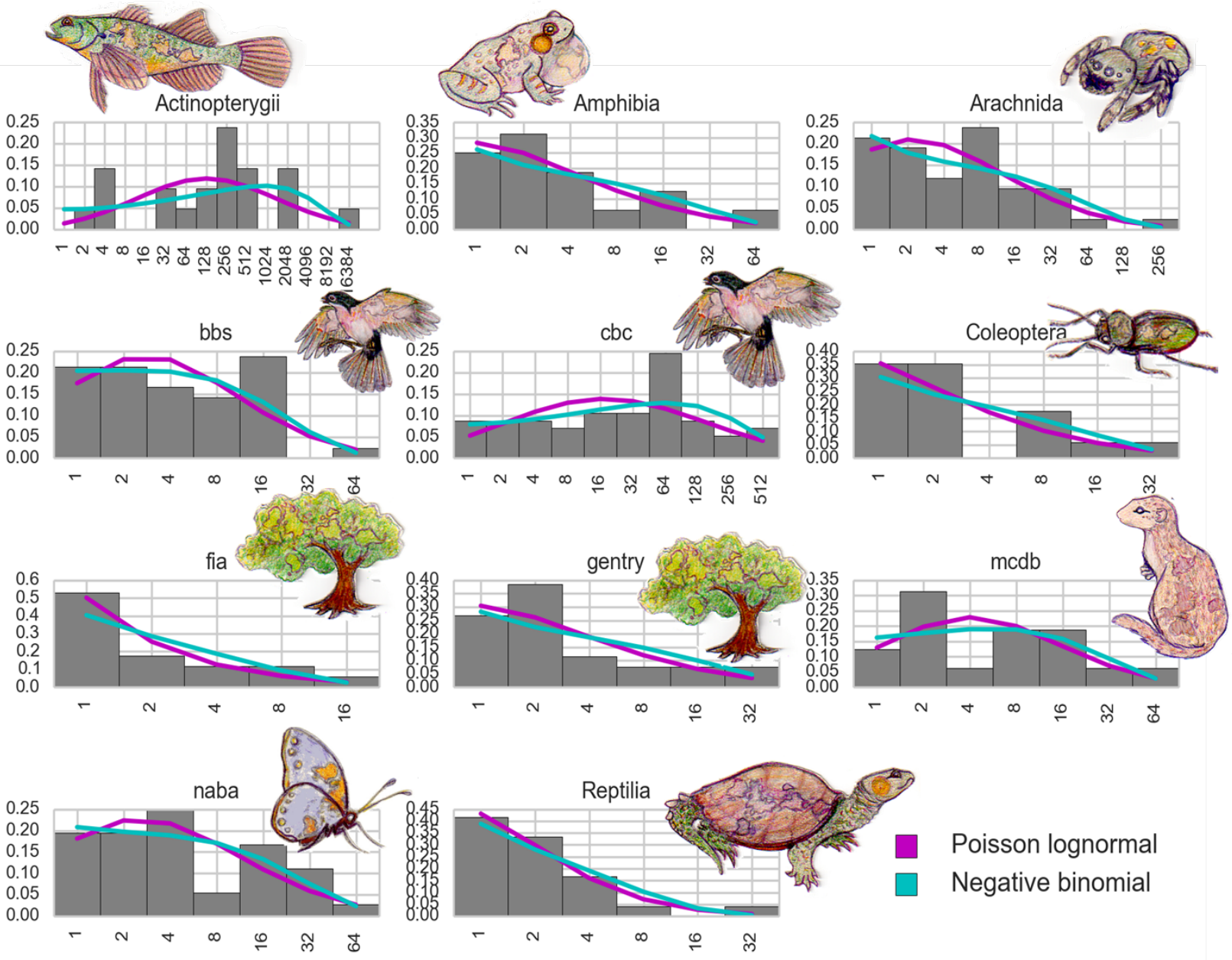
Used the same data and model fitting approach.
Compared a non-neutral model (Poisson lognormal) to a neutral model (negative binomial).

Neutral Analysis

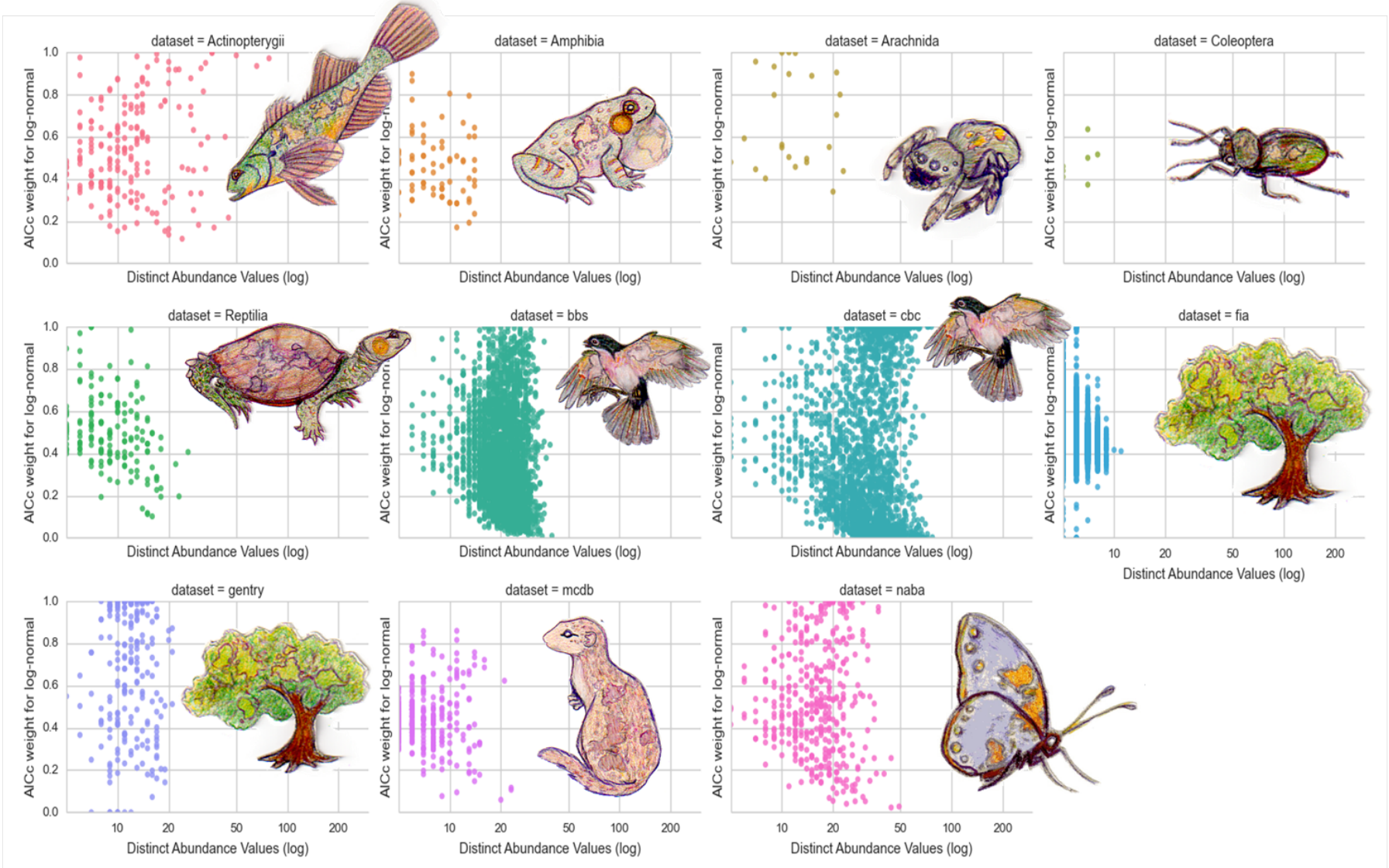


Connolly et al. 2014.

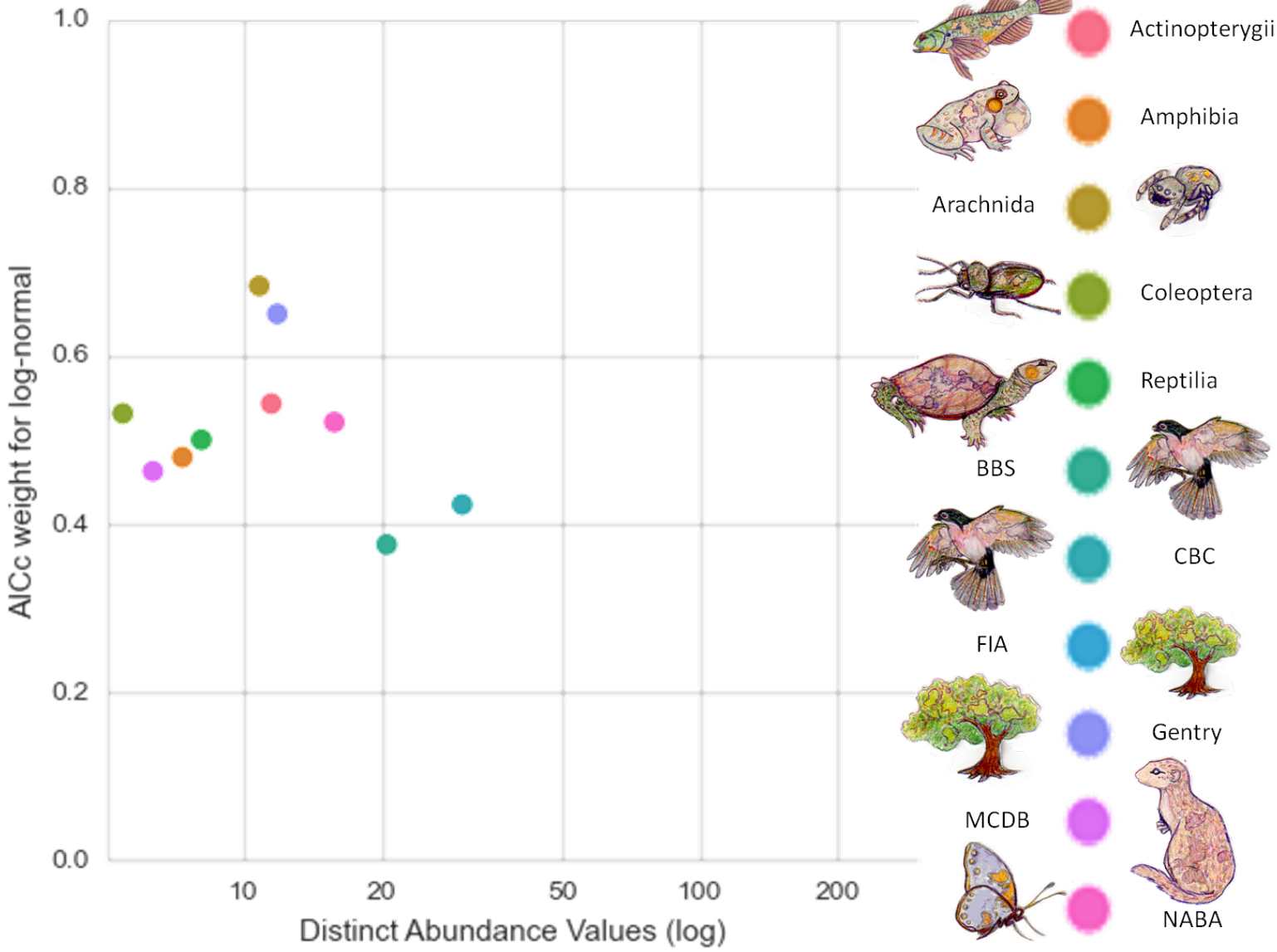
Neutral Analysis



Neutral Analysis



Neutral Analysis



Neutral analysis

Difficult to identify a clear winning model.

- Results consistent with our species abundance distribution model comparisons.
- Results different from Connolly et al. 2014.
    - Non-neutral model outperforms the neutral model in marine systems.
    - Our results suggest marine systems more generally approximated by non-neutral dynamics; terrestrial systems more variable between neutral and non-neutral dynamics.

Conclusions

Challenging to infer process from species abundance distributions alone.

- Multiple mechanisms proposed for each SAD formulation.
- Broad model categorization (i.e. neutral or non-neutral) may be more productive.
- May not be one single suite of processes that dominates.

Conclusions

Challenges in identifying mechanism among datasets.

- Biological vs. non-biological differences (spatial structuring, sampling intensity).
- Diverse data removes uncertainty about non-biological pattern generating mechanisms.
- Even with a great deal of data, identifying mechanism is still challenging.

Conclusions

Predictive macroecology

- Traditional approach is pattern to process to prediction.
- May be possible to generate robust ecological predictions from general patterns.
- Process and prediction may be two separate research goals.

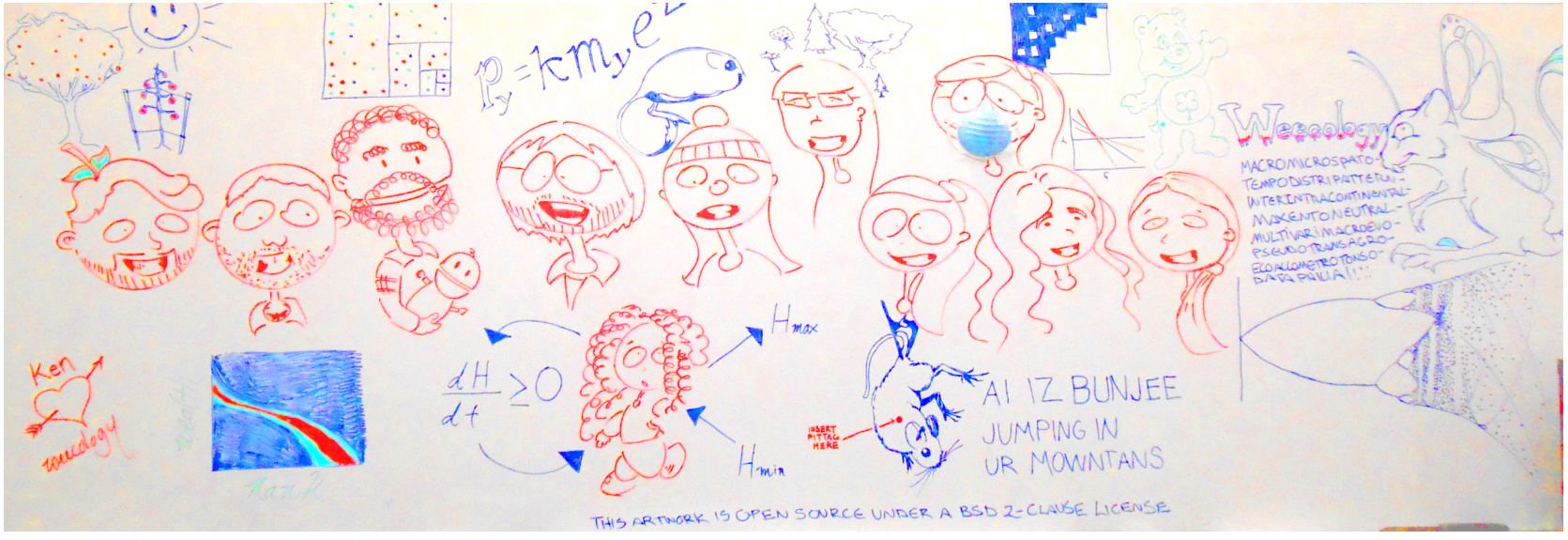
Acknowledgements

Funding sources:

- USU Department of Biology
- Intellectual Ventures, private funding to Morgan Ernest
- National Science Foundation CAREER Grant to Ethan White
- Gordon & Betty Moore Foundation's Data-Driven Discovery Initiative Grant to Ethan White.
- USU Graduate School Dissertation Fellowship

Acknowledgements

Weecologists past, present, & future



(especially Xiao Xiao & Ken Locey (creator of the whiteboard))

Acknowledgements

Dr. Thomas Price & USU Student Health Center.

Tea, heating pads, & the Flint Hills of Kansas.

A very supportive husband & family.

Publicly available data, & the citizen scientists that make that possible.

Accessibility

This dissertation brought to you by:

Disability accommodations

- Remote access & participation.
- Computational tools & tricks.
  - Version control (GitHub).
  - Publicly available data.
  - Programming skills (data manipulation & analysis).

Questions?