

Thesis proposal

Michael D. Catchen^{1,2}

¹ McGill University ² Québec Centre for Biodiversity Sciences

Correspondance to:

Michael D. Catchen — michael.catchen@mail.mcgill.ca

This work is released by its authors under a CC-BY 4.0 license



Last revision: *November 23, 2021*

The proposal for my thesis, *Simulation models for predictive ecology*

1 Introduction

2 Within the last several hundred years, human activity has induced rapid changes in Earth's atmosphere,
3 oceans, and surface. Greenhouse gas emissions have caused an increase the temperature of both Earth's
4 terrain and oceans, and both agricultural and urban development has rapidly reshaped the Earth's land
5 cover. These the bulk of this change has occurred within the last several hundred years, a geological
6 instant, inducing a sudden shift in conditions to Earth's climate and biosphere. As a result *ecological*
7 *forecasting*—modeling how ecosystems and their services will change in the future—and then using these
8 forecasts to make decisions to mitigate the negative consequences of this change on ecosystems, their
9 functioning, and the services they provide to humans has emerged as an imperative for ecology and
10 environmental science (Dietze 2017). However, robust prediction of ecological processes is, to say the
11 least, quite difficult (Beckage *et al.* 2011; Petchey *et al.* 2015). This difficultly is compounded by a few
12 factors, the first being that sampling ecosystems is not easy. Ecological data is often biased, noisy, and
13 sparse in both space and time. The current paucity of ecological data has resulted in much interest in
14 developing global systems for *ecosystem monitoring* (Makiola *et al.* 2020), which would systematize the
15 collection of biodiversity data in manner that makes detecting and predicting change more possible than
16 at the moment (Urban *et al.* 2021).

17 The second major challenge in ecological forecasting is that the underlying dynamics of most ecological
18 processes are unknown and instead must be inferred from this (sparse) data. Much of the history of
19 quantitatively modeling ecosystems have been done in the language of dynamical systems, describing how
20 the value of an observable state of the system, represented by a vector of numbers $[x_1, x_2, \dots, x_n]^T = \vec{x}$
21 changes as over time, yielding models in the form of differential equations in continuous-time settings,
22 $\frac{dx}{dt} = f(x)$, or difference equations in discrete-time settings, $x_t = f(x_{t-1})$, where $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is an
23 arbitrary function describing how the system changes on a moment-to-moment basis (e.g. in the context of
24 communities, f could be Lotka-Volterra, Holling-Type-III or DeAngelis-Beddington functional response).
25 The form of this functional response in real systems, and whether it is meaningfully non-zero for a given
26 species interaction, is effectively unknown, and some forms are inherently more “forecastable” than
27 others (Beckage *et al.* 2011; Chen *et al.* 2019; Pennekamp *et al.* 2019). The initial success of these forms of
28 models can be traced back to the larger program of ontological reductionism, which became the default
29 approach to modeling in the sciences after its early success in physics, which, by the time ecology was

30 becoming a quantitative science (sometime in the 20th century, depending on who you ask), became the
31 foundation for mathematical models in ecology.

32 However, we run into many problems when aiming to apply this type of model to empirical ecological
33 data. Ecosystems are perhaps the quintessential example of system that cannot be understood by iterative
34 reduction of its components into constituent parts—ecological phenomena are emergent: the product of
35 different mechanisms operating at different spatial, temporal, and organizational scales (Levin 1992).

36 Further this analytical approach to modeling explicitly ignores known realities: ecological dynamics not
37 deterministic and many analytic models in ecology assume long-run equilibrium. Finally, perhaps the
38 biggest challenge in using these models to describe ecological processes is ecosystems consist of more
39 dimensions than the tools of analytic models are suited for. As the number of variables in an analytic
40 model increases, so does the ability of the scientist to discern clear relationships between them given a
41 fixed amount of data, the so-called “curse” of dimensionality.

42 But these problems are not solely unique to ecology. The term *ecological forecasting* implicitly creates an
43 analogy with weather forecasting. Although it has become a trite joke to complain about the weather
44 forecast being wrong, over the last 50 years the field of numerical weather prediction (NWP) has
45 dramatically improved our ability to predict weather across the board (Bauer *et al.* 2015). The success of
46 NWP, and the Earth observations systems that support it (Hill *et al.* 2004), should serve as a template for
47 development of a system for monitoring Earth’s biodiversity. Much like ecology, NWP is faced with
48 high-dimensional systems that are governed by different mechanisms at different scales. The success of
49 NWP is that, rather than, say, attempt to forecast the weather in Quebec by applying Navier-Stokes to
50 entire province, to instead use simulation models which describe known mechanisms at different scales,
51 and use the availability of increasing computational power to directly simulate many batches of dynamics
52 which directly incorporate stochasticity and uncertainty in parameter estimates via random number
53 generation.

54 But forecasting is only half the story—if indeed “[ecologists] have hitherto only interpreted the world in
55 various ways; the point is to change it,” then once we have a forecast about how an ecosystem will change
56 in the future, what if this forecast predicts a critical ecosystem service will deteriorate? We are still left
57 with the question, what do we in the time being to mitigate the potentially negative consequences a
58 forecast predicts? In this framing, mitigating the consequences of anthropogenic change on ecosystems
59 becomes an optimization problem: given a forecast of the future state of the system, and some “goal” state

60 for the future, the problem is then to optimize our intervention into the system to maximize the
61 probability the system approaches our “goal” state. This dissertation aims to this framework for ecosystem
62 monitoring and forecasting (fig. 1, left), and each chapter address some aspect of this pipeline to data from
63 a monitoring network to forecasts to mitigation strategy (fig. 1, right).

64 [Figure 1 about here.]

65 The primary research challenges this thesis addresses: how do we design ecological samples to? How do
66 we build the software infrastructure to assimilate data from a variety of sources? How do we propagate
67 uncertainty from data to forecasts? The flow of chapters follows the flow in fig. 1 (left), from data
68 collection via a monitoring network, to forecasting an essential biodiversity variable (EBV), to optimizing
69 mitigation strategy based on constraints. In chapter one, we discuss how simulation can aid in the design
70 of ecological samples and monitoring network design. In chapter two we use data to forecast the
71 uncoupling of a plant-pollinator network. In chapter three, we apply simulation methods in landscape
72 ecology to optimize corridor placement with respect maximize the time-until-extinction of a
73 metapopulation. The fourth and final chapter is the software (*MetacommunityDynamics.jl*) which enables
74 the rest of the dissertation.

75 **Chapter One: Optimizing spatial sampling of species interactions**

76 **Objective**

77 This chapter uses simulation models to investigate the relationship between species relative abundance,
78 sampling effort, and probability of observing an interaction between species in order to aid in the design of
79 samples of ecological interactions, and to provide a null expectation of false-negative probability for a
80 dataset of a given size. Further it then proposes a method for optimizing the spatial sampling locations to
81 maximize the probability of detecting an interaction between two species given a fixed number of total of
82 observations, and the distributions of each species. This addresses the optimization of monitoring
83 network part of the flow from data to mitigation at the top of fig. 1, left. As explored in the previous
84 chapter, there are false-negatives in interaction data. However, there is more than one way to observe a
85 false-negative when sampling interactions. fig. 2 shows a taxonomy of false-negatives in occurrence,
86 co-occurrence, and interaction data.

87

[Figure 2 about here.]

88 Methods

89 The first result is to compute a null expectation of the probability of an interaction false-negative as a
 90 function number of total observations of individuals of *any species*. This is done by simulating the process
 91 of observation, where the probability of observing a given species is its relative abundance. We use a
 92 log-normal distribution of relative abundance (Hubbell 2001) and simulating the process of observation on
 93 food-webs generated using the niche model (Williams & Martinez 2000) with connectance parameterized
 94 by the flexible-links model (MacDonald *et al.* 2020). An example of this relation for networks with varying
 95 species richness is shown in fig. 3.

96 We then go on to testing some assumptions of this neutral model with empirical data. Primarily that we
 97 analytically show that our neutral model, if anything, underestimates the probability of false-negatives if
 98 there are positive associations between species co-occurrence, and we show these positive associations
 99 exist in two sets of spatially replicated samples of interaction networks (Thompson & Townsend 2000;
 100 Hadfield *et al.* 2014), fig. 4—further I'm planning to add the field data from the previous chapter into this
 101 analysis once available. Finally this chapter proposes a simulated annealing method to optimize the a set
 102 of n points in space to maximize the probability of detecting an interaction between two species a and b
 103 with *known* distributions D_a, D_b .

104 Results

105 The first major result is using the simulation of the observation process described above to generate
 106 expectations of interaction false-negative rate (FNR) as a function of total number of observations, with
 107 the goal being for this estimate to be used as correction for detection error when fitting an interaction
 108 prediction model. This relationship varies with the total richness of the metaweb fig. 3.

109

[Figure 3 about here.]

110 The second major result is that we analytically show that the this simulated observation model, by
 111 assuming that there is no association between observing two species given that they interact, actually

112 under predicts the realized false-negative interaction rate. We then demonstrate that this positive
113 association association exists in two empirical systems fig. 4.

114 [Figure 4 about here.]

115 **Progress**

116 This chapter is mostly complete. The only remaining work is the implementation of simulated annealing
117 optimization process. This will be done by using a proposal function which takes a set of coordinates in
118 space and proposes a new location for each point based on a distance-decaying kernel.

119 **Chapter Two: Forecasting the spatial uncoupling of a plant-pollinator 120 network**

121 Interactions between plants and pollinators form networks which together structure the “architecture of
122 biodiversity” (Bascompte & Jordano 2007). The functioning and stability of ecosystems emerge from these
123 interactions, but anthropogenic change threatens to unravel and “rewire” these interaction networks
124 (CaraDonna *et al.* 2017), jeopardizing the persistence of these systems. Plant-pollinator networks face two
125 possible forms of rewiring in response to anthropogenic environmental change: spatial and temporal.
126 Range shifts could cause interacting species to no longer overlap in space, and shifts in phenology could
127 cause interacting species to no longer occur at the same time of year. This chapter uses several years of
128 data on bumblebee-flower phenology and interactions across several field sites, each consisting of several
129 plots across an elevational gradient, combined with spatial records of species occurrence via GBIF to
130 forecast the uncoupling of the plant-pollinator metaweb of Colorado.

131 [Figure 5 about here.]

132 **Methods**

133 The data for this chapter is derived from multiple sources that can be split into four categories. (1) Field
134 data from three different field sites across Colorado, each with multiple plots across an elevational

135 gradient, for seven, seven, and three years respectively. This data was collected by Paul CaraDonna and
136 Jane Oglevie (from the Rocky Mountain Biological Laboratory; RMBL) and Julian Resasco (CU Boulder).
137 (2) GBIF spatial occurrence records of each of these species across Colorado, including a metaweb of
138 interactions across all of Colorado taken from GBIF. (3) Remotely sensed data consisting of current and
139 forecasting bioclimatic variables from CHELSA. (4) Phylogenies for both bee and flower species derived
140 from NCBI GenBank barcodes for mitochondrial COI (bumblebees) and chloroplast rbcL (flowers).

141 As the data we have is spatially sparse and likely to contain many interaction “false-negatives” (Strydom *et*
142 *al.* 2021b), we begin by predicting a metaweb of interactions across Colorado as they exist *in the present*.
143 We do this using a set of candidate interaction prediction models: relative abundance only, phylogenetic
144 embedding only (a la Strydom *et al.* (2021a)), niche embedding only (Gravel *et al.* 2019), and all pairwise
145 combinations of those constituent models. After validating and selecting the best performing model, we
146 then predict how these distributions of each of these species will change under the CMIP6 consensus
147 climate forecast (Karger *et al.* 2017), and then finally quantify the reduction in spatial between species for
148 which there is a predicted interaction.

149 **Results**

150 Here we show the in-progress results, which are the prerequisites for the analysis outlined above:
151 phylogenies for both plant and bee species (fig. 6) and species distribution models for all species (an
152 example shown in fig. 7).

153 [Figure 6 about here.]

154 [Figure 7 about here.]

155 **Progress**

156 At the moment, we have derived phylogenies (fig. 6) and SDMs (fig. 7) for all the species present in the
157 Colorado GBIF metaweb. I’ve also been exploring the data available from Julian Resasco. The primary
158 constraint on further progress is that we are waiting on the finalization of a data sharing agreement with
159 RMBL.

160 **Chapter Three: Optimizing corridor placement against ecological**
161 **dynamics**

162 **Objective**

163 As land-use change has caused many habitats to become fragmented and patchy, promoting landscape
164 connectivity has become of significant interest to mitigate the effects of this change on Earth's biodiversity.
165 However, the practical realities of conservation mean that there is a limitation on how much we can
166 modify landscapes in order to do this. So what is the best place to put a corridor given a constraint on how
167 much surface-area you can change in a landscape? This is the question this chapter seeks to answer.
168 Models for inferring corridor locations have been developed, but are limited in that they are not developed
169 around promoting some element of ecosystem function, but instead by trying to find the path of least
170 resistance in an existing landscape from a derived resistance surface (Peterman 2018). This chapter
171 proposes a general algorithm for choosing corridor placement to optimize a measurement of ecosystem
172 functioning derived from simulations run on each proposed landscape modification.

173 **Methods**

174 [Figure 8 about here.]

175 We propose various landscape modifications which alter the cover of a landscape, represented as a raster.
176 We then compute a new resistance surface based on the proposed landscape modification using
177 Circuitscape (McRae *et al.* 2008), and based on the values of resistance to dispersal between pairs of
178 locations we simulate spatially-explicit metapopulation dynamics model (Hanski & Ovaskainen 2000;
179 Ovaskainen *et al.* 2002) to estimate a distribution of time until extinction for each landscape modification.
180 The largest challenge in implementing this algorithm is the space of potential modifications grows as
181 $O((nm)!)$ for an n by m raster. For most actual landscapes to which we wish to apply this method, the set
182 of possible modifications becomes uncomputably large, so we use simulated annealing to explore the
183 search space of possible modifications to estimate the modification that maximizes the time-until
184 extinction of simulated metapopulation dynamics under that hypothetical modified landscape.
185 The biggest challenge in implementing simulated annealing in this context is defining a proposal function

186 for landscape modifications. At the moment this is done by computing the minimum-spanning-tree
187 (MST) of the spatial nodes, and then proposing corridors that connect nodes that are already connected in
188 the MST.

189 The final component of this chapter is measuring the effect of land-use change on the robustness of the
190 optimized corridor.

191 **Progress**

192 The current progress is that I have an algorithm for proposing landscape modifications and a simple
193 implementation of simulated annealing. The only gap left is implementing Circuitscape estimation of
194 resistance surfaces.

195 **Chapter Four: MetacommunityDynamics.jl: a virtual laboratory for 196 community ecology**

197 **Objective**

198 The final chapter consists of a collection of modules in the Julia language for different aspects of
199 community ecology, including most of the code used for the preceding chapters. Indeed
200 MetacommunityDynamics.jl (MCD.jl) is the epicenter of this set of tools, but due to the nature of the Julia
201 language, MCD.jl is interoperable with several existing packages within the EcoJulia organization,
202 including several to which I have contributed. A diagram showing the relation between these packages is
203 shown in fig. 9.

204 We need tools to generate synthetic data from a *known* set of mechanisms and parameters to test our
205 methods for parameter inference and forecasting on this *known* system to assess the effectiveness of these
206 inference and forecasting methods.

207 [Figure 9 about here.]

208 **Methods**

209 Software is structured based on DG.jl (**cite?**) and Dispersal.jl (**cite?**).
210 Uses methods from EN.jl to generate metawebs, or can use empirical networks from Mangal.jl (**cite?**).
211 Framework based on (**Velland2010ConSyn?**), processes divided into four categories: selection, dispersal,
212 drift, speciation.

213 **Results**

214 Below (fig. 10) is a sample output of simulated food-web dynamics for a metaweb of 100 species generated
215 using the minimum-potential-niche model with connectance $C = 0.05$ and forbidden-link probability of
216 0.5. The dynamics change according to a Lotka-Volterra functional response, dispersal (with dispersal
217 distance inverse proportional to trophic-level), linear mortality, and logistic growth for any species at the
218 producer trophic-level.

219 [Figure 10 about here.]

220 **Progress**

221 The software as it exists is capable of simulating the biomass dynamics of arbitrarily large food-webs using
222 Lotka-Volterra, Holling Type-II, or Holling Type-III functional responses. It currently has methods to
223 implement Gaussian drift, and various forms of dispersal via Dispersal.jl. Also occupancy dynamics for
224 Levins metapopulations (**levins1967?**), and spatially explicit Hanski metapopulations (**hanski2001?**).
225 This is most of what needs to exist for the preceding chapters. In-progress functionality includes selection
226 (which affects growth-rate) on arbitrary environmental variables in progress, as well as traits.

227 **Discussion**

228 Describing expected/anticipated contributions of the thesis. Very important for QE. This
229 should be at least half a page.

230 **References**

- 231 Bascompte, J. & Jordano, P. (2007). Plant-Animal Mutualistic Networks: The Architecture of Biodiversity.
232 *Annual Review of Ecology, Evolution, and Systematics*, 38, 567–593.
- 233 Bauer, P., Thorpe, A. & Brunet, G. (2015). The quiet revolution of numerical weather prediction. *Nature*,
234 525, 47–56.
- 235 Beckage, B., Gross, L.J. & Kauffman, S. (2011). The limits to prediction in ecological systems. *Ecosphere*, 2,
236 art125.
- 237 CaraDonna, P.J., Petry, W.K., Brennan, R.M., Cunningham, J.L., Bronstein, J.L., Waser, N.M., et al. (2017).
238 Interaction rewiring and the rapid turnover of plantpollinator networks. *Ecology Letters*, 20, 385–394.
- 239 Chen, Y., Angulo, M.T. & Liu, Y.-Y. (2019). Revealing Complex Ecological Dynamics via Symbolic
240 Regression. *BioEssays*, 41, 1900069.
- 241 Dietze, M.C. (2017). Prediction in ecology: A first-principles framework. *Ecological Applications*, 27,
242 2048–2060.
- 243 Gravel, D., Baiser, B., Dunne, J.A., Kopalke, J.-P., Martinez, N.D., Nyman, T., et al. (2019). Bringing Elton
244 and Grinnell together: A quantitative framework to represent the biogeography of ecological
245 interaction networks. *Ecography*, 42, 401–415.
- 246 Hadfield, J.D., Krasnov, B.R., Poulin, R. & Nakagawa, S. (2014). A Tale of Two Phylogenies: Comparative
247 Analyses of Ecological Interactions. *The American Naturalist*, 183, 174–187.
- 248 Hanski, I. & Ovaskainen, O. (2000). The metapopulation capacity of a fragmented landscape. *Nature*, 404,
249 755–758.
- 250 Hill, C., DeLuca, C., Balaji, Suarez, M. & Da Silva, A. (2004). The architecture of the Earth System
251 Modeling Framework. *Computing in Science Engineering*, 6, 18–28.
- 252 Hubbell, S.P. (2001). *The unified neutral theory of biodiversity and biogeography*. Monographs in
253 population biology. Princeton University Press, Princeton.
- 254 Karger, D.N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R.W., et al. (2017). Climatologies at
255 high resolution for the earth's land surface areas. *Scientific Data*, 4, 170122.

- 256 Levin, S.A. (1992). The Problem of Pattern and Scale in Ecology: The Robert H. MacArthur Award
257 Lecture. *Ecology*, 73, 1943–1967.
- 258 MacDonald, A.A.M., Banville, F. & Poisot, T. (2020). Revisiting the Links-Species Scaling Relationship in
259 Food Webs. *Patterns*, 1.
- 260 Makiola, A., Compson, Z.G., Baird, D.J., Barnes, M.A., Boerlijst, S.P., Bouchez, A., et al. (2020). Key
261 Questions for Next-Generation Biomonitoring. *Frontiers in Environmental Science*, 7.
- 262 McRae, B.H., Dickson, B.G., Keitt, T.H. & Shah, V.B. (2008). Using Circuit Theory to Model Connectivity
263 in Ecology, Evolution, and Conservation. *Ecology*, 89, 2712–2724.
- 264 Ovaskainen, O., Sato, K., Bascompte, J. & Hanski, I. (2002). Metapopulation Models for Extinction
265 Threshold in Spatially Correlated Landscapes. *Journal of Theoretical Biology*, 215, 95–108.
- 266 Ovaskainen, O., Sato, K., Bascompte, J. & Hanski, I. (2002). Metapopulation Models for Extinction
267 Threshold in Spatially Correlated Landscapes. *Journal of Theoretical Biology*, 215, 95–108.
- 268 Pennekamp, F., Iles, A.C., Garland, J., Brennan, G., Brose, U., Gaedke, U., et al. (2019). The intrinsic
269 predictability of ecological time series and its potential to guide forecasting. *Ecological Monographs*, 89,
270 e01359.
- 271 Petchey, O.L., Pontarp, M., Massie, T.M., Kéfi, S., Ozgul, A., Weilenmann, M., et al. (2015). The ecological
272 forecast horizon, and examples of its uses and determinants. *Ecology Letters*, 18, 597–611.
- 273 Peterman, W.E. (2018). ResistanceGA: An R package for the optimization of resistance surfaces using
274 genetic algorithms. *Methods in Ecology and Evolution*, 9, 1638–1647.
- 275 Strydom, T., Bouskila, S., Banville, F., Barros, C., Caron, D., Farrell, M.J., et al. (2021a). Food web
276 reconstruction through phylogenetic transfer of low-rank network representation.
- 277 Strydom, T., Catchen, M.D., Banville, F., Caron, D., Dansereau, G., Desjardins-Proulx, P., et al. (2021b). A
278 *Roadmap Toward Predicting Species Interaction Networks (Across Space and Time)* (Preprint).
279 EcoEvoRxiv.
- 280 Thompson, R.M. & Townsend, C.R. (2000). Is resolution the solution?: The effect of taxonomic resolution
281 on the calculated properties of three stream food webs. *Freshwater Biology*, 44, 413–422.
- 282 Urban, M.C., Travis, J.M.J., Zurell, D., Thompson, P.L., Synes, N.W., Scarpa, A., et al. (2021). Coding for
283 Life: Designing a Platform for Projecting and Protecting Global Biodiversity. *BioScience*.

²⁸⁴ Williams, R.J. & Martinez, N.D. (2000). Simple rules yield complex food webs. *Nature*, 404, 180–183.

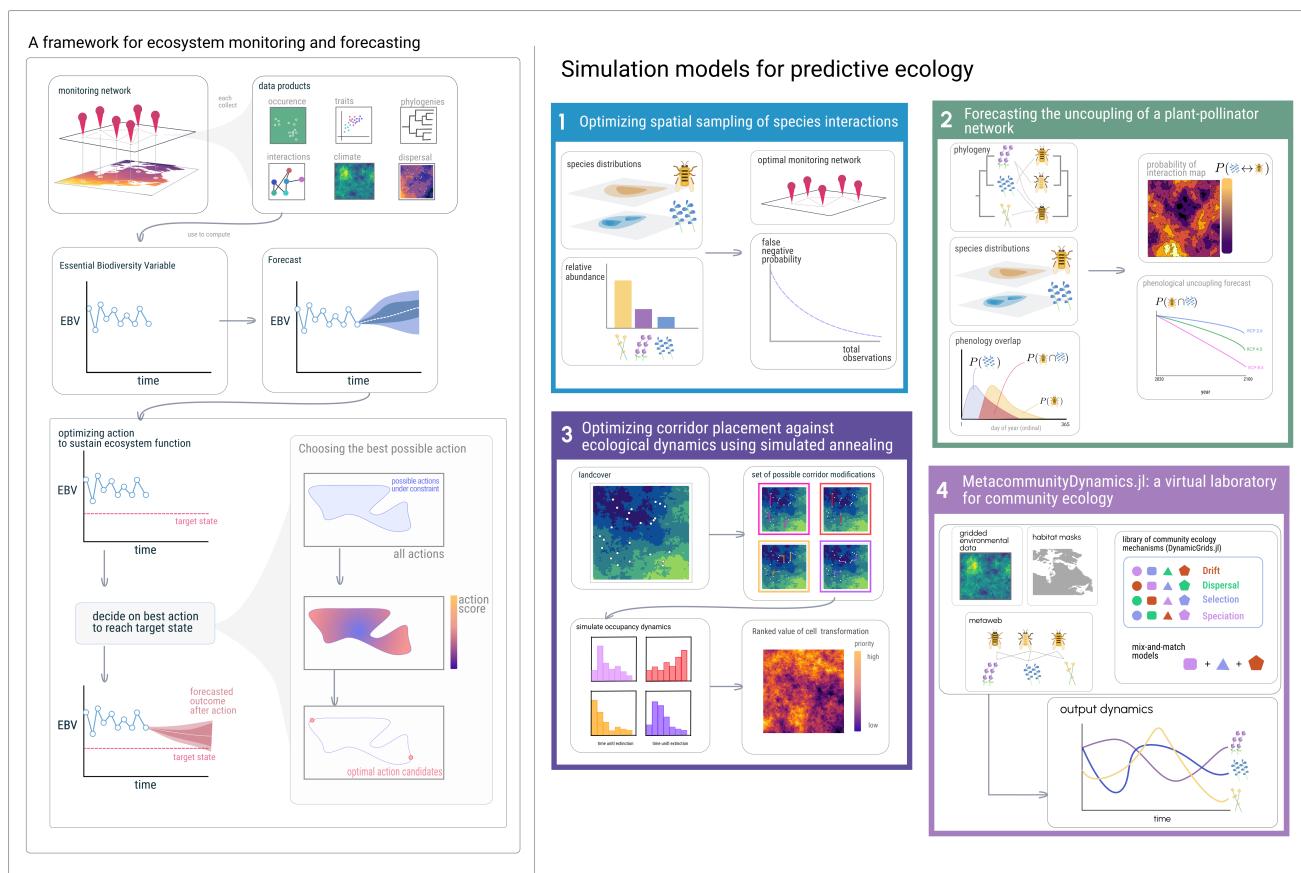


Figure 1: thesis concept

Species A occurs?

		true		false								
		true	Species A observed?									
		true	Species B observed? <table border="1"> <tr> <td>true</td><td>co-occurrence true-positive</td><td>Interaction observed? true false</td><td>co-occurrence false-negative</td></tr> <tr> <td>false</td><td>co-occurrence false-negative</td><td>interaction true-positive interaction false-negative</td><td>occurrence false-negative</td></tr> </table>	true	co-occurrence true-positive	Interaction observed? true false	co-occurrence false-negative	false	co-occurrence false-negative	interaction true-positive interaction false-negative	occurrence false-negative	
true	co-occurrence true-positive	Interaction observed? true false	co-occurrence false-negative									
false	co-occurrence false-negative	interaction true-positive interaction false-negative	occurrence false-negative									
Species B occurs?	true	co-occurrence true-negative		occurrence true-negative								
false												

Figure 2: A taxonomy of occurrence, co-occurrence, and interaction false negatives in data

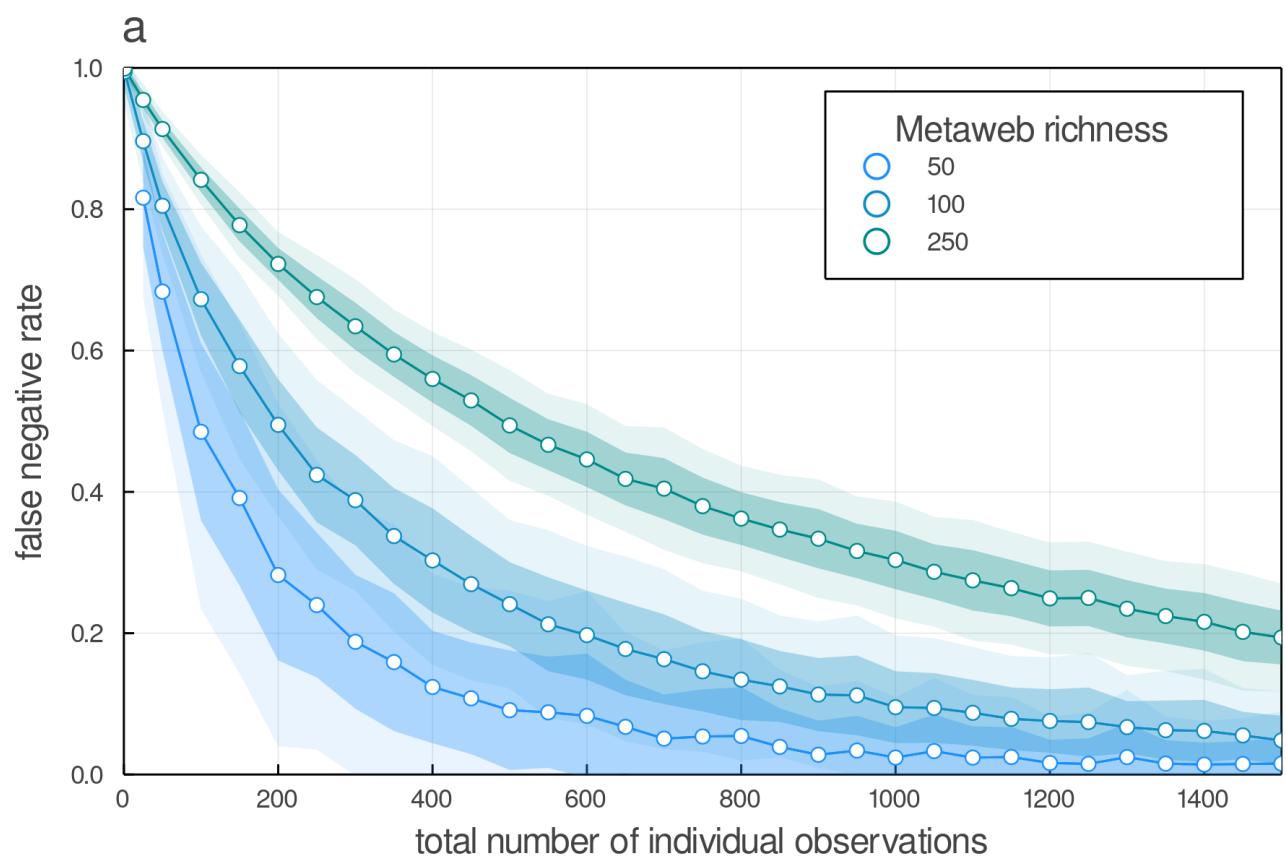


Figure 3: foo

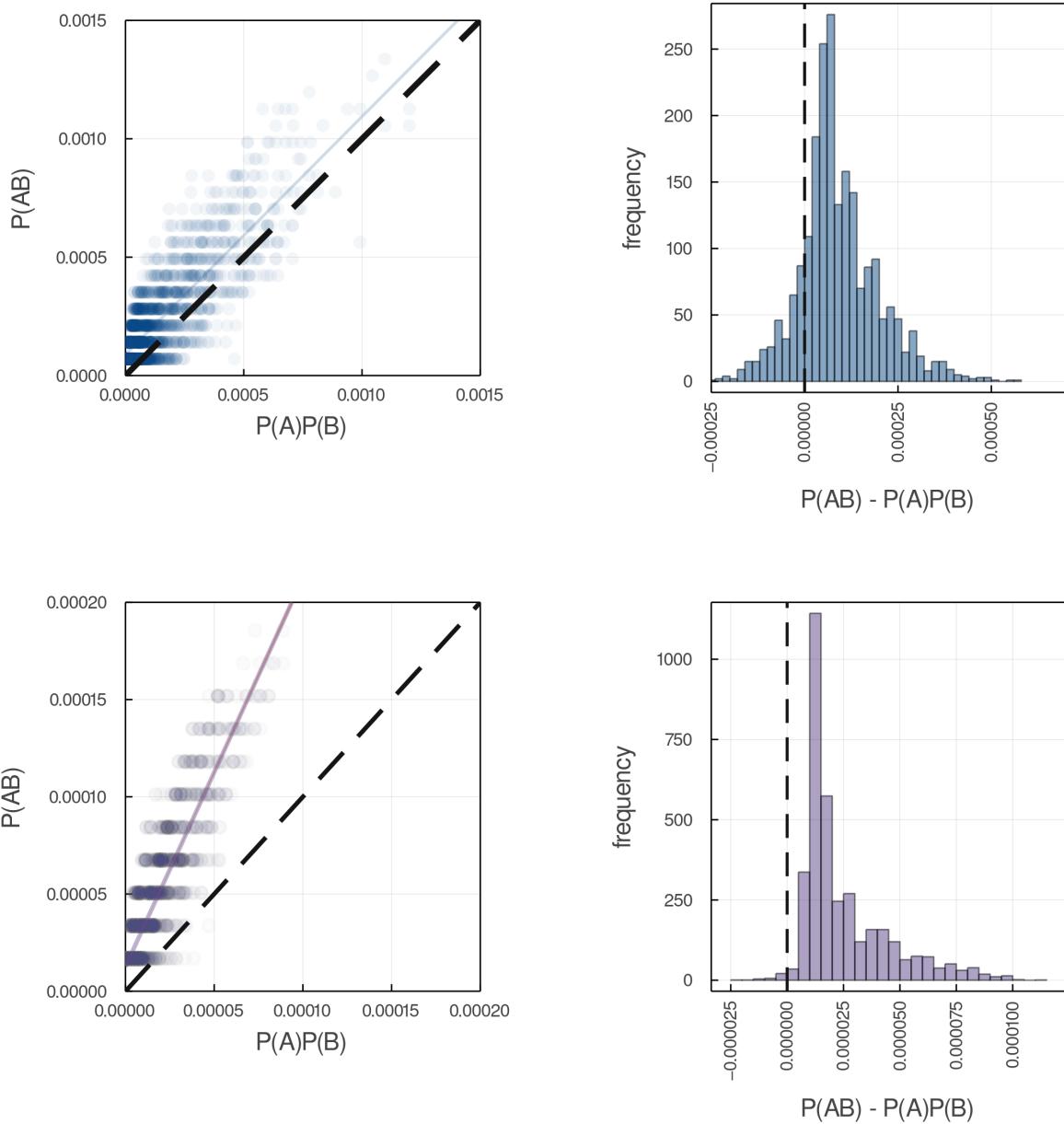


Figure 4: Demonstrates positive associations in co-occurrence

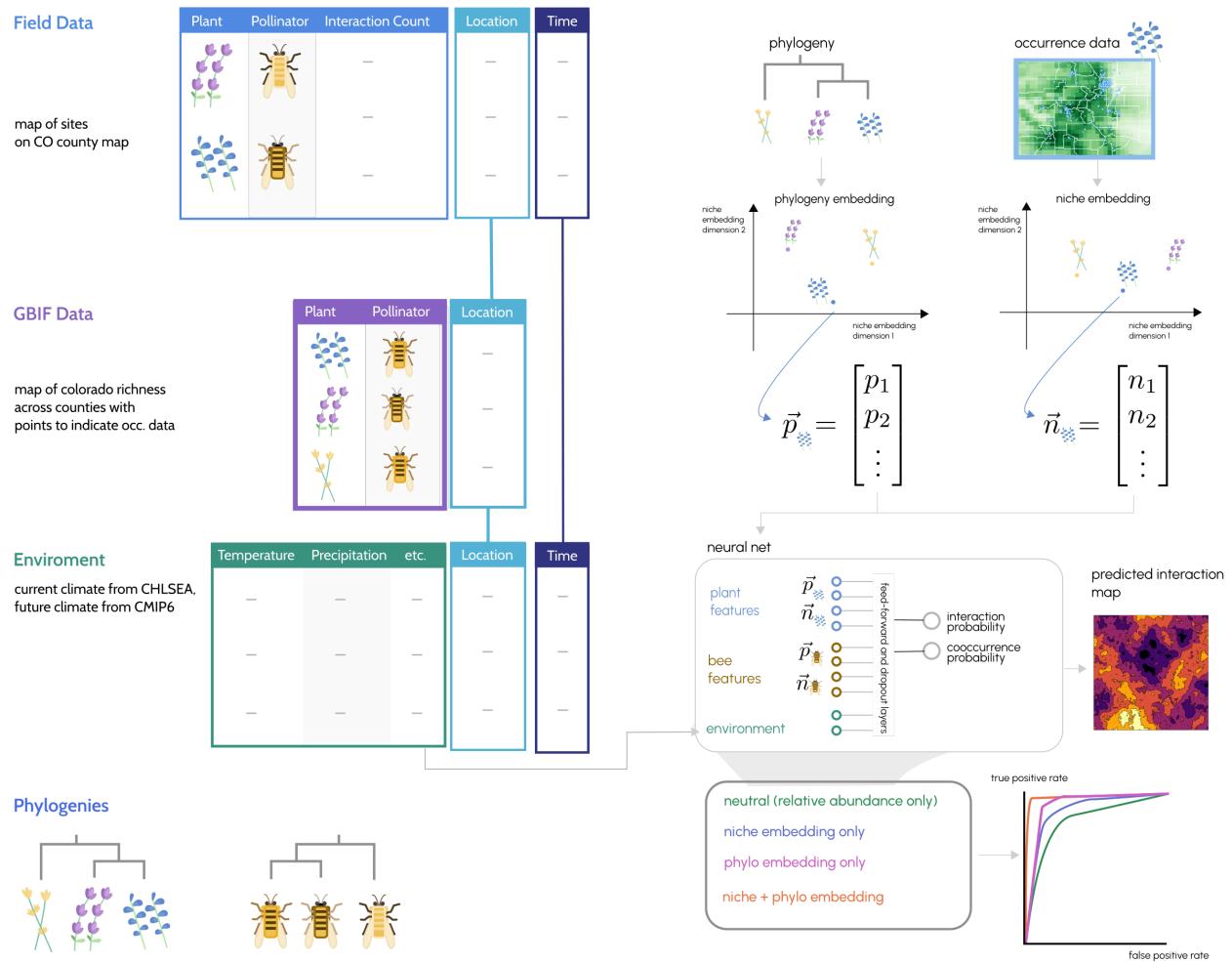


Figure 5: Chapter One conceptual figure. Left: the sources of data and how they can be synthesized. Right: The flow from data to interaction prediction using a few different interaction prediction models.

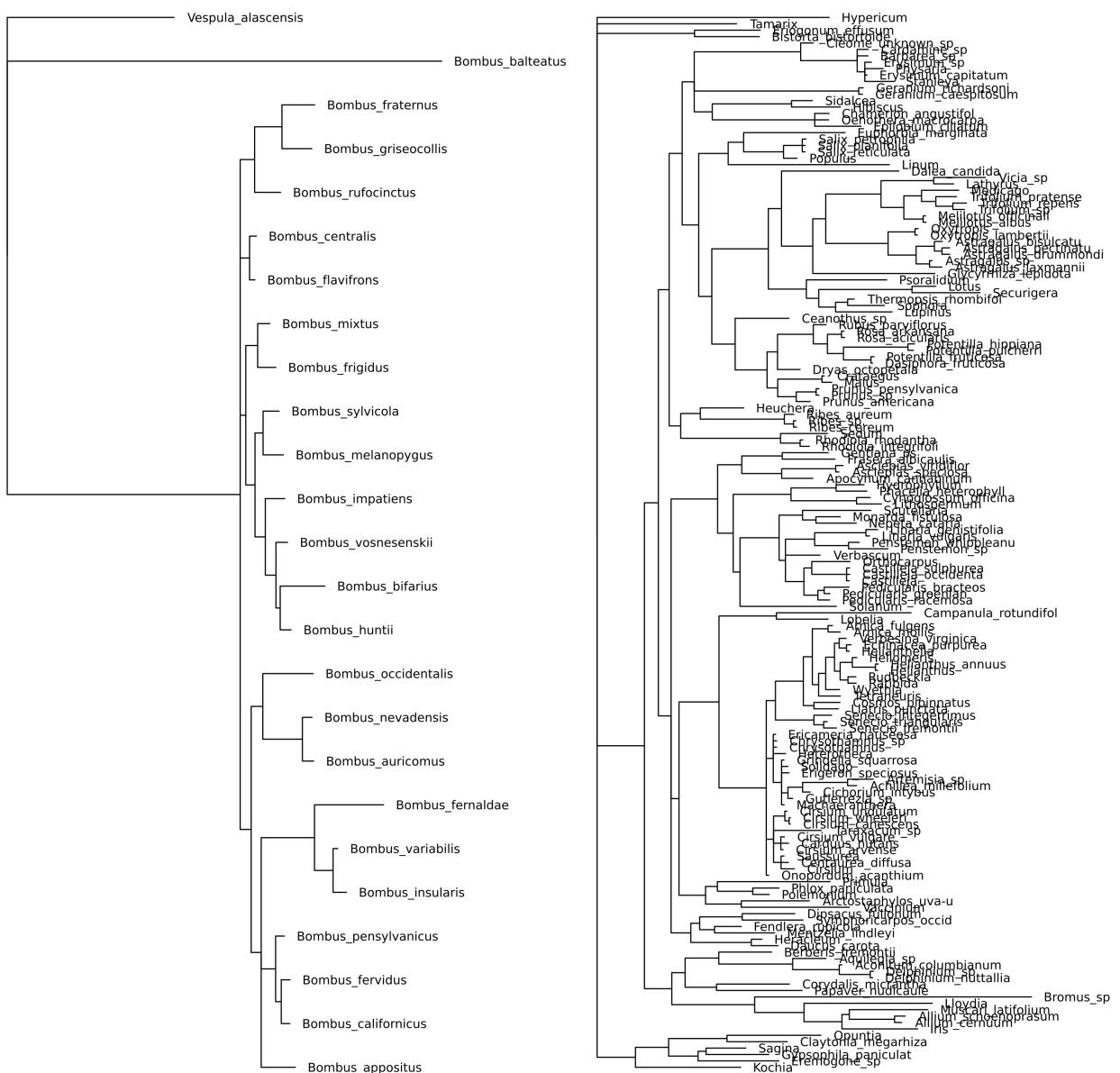


Figure 6: Phylogeny for both bumblebee species (left) and flower species (right)

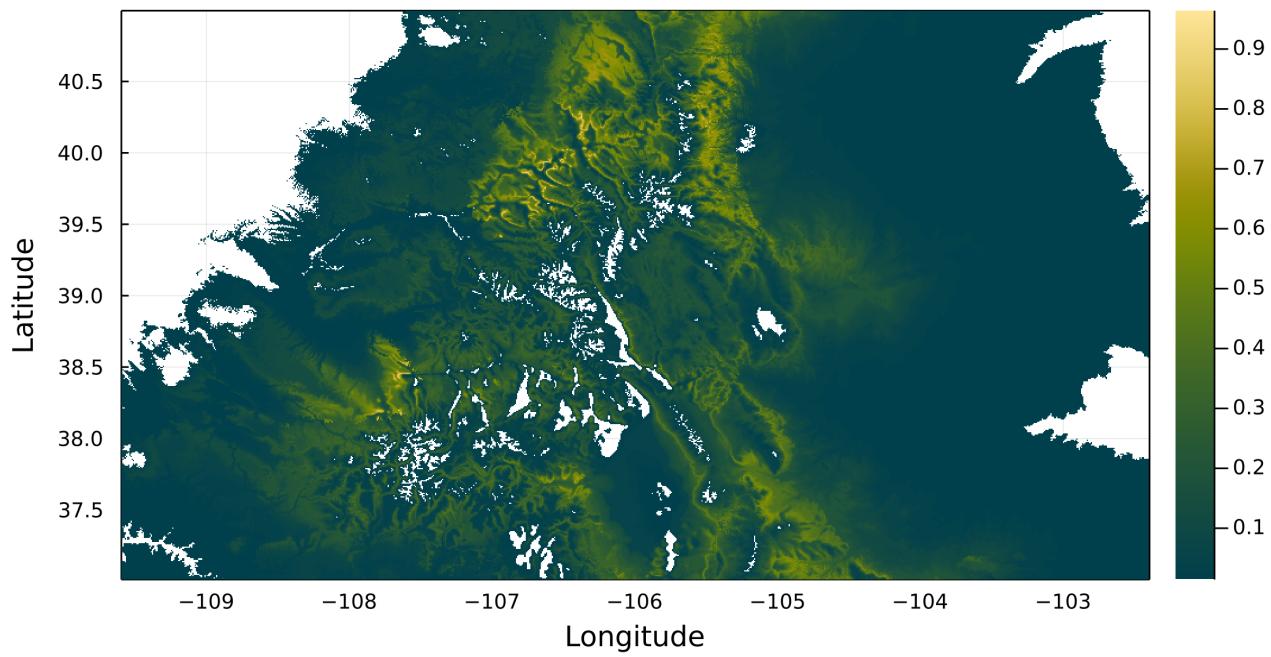


Figure 7: Example SDM for *Achillea millefolium*

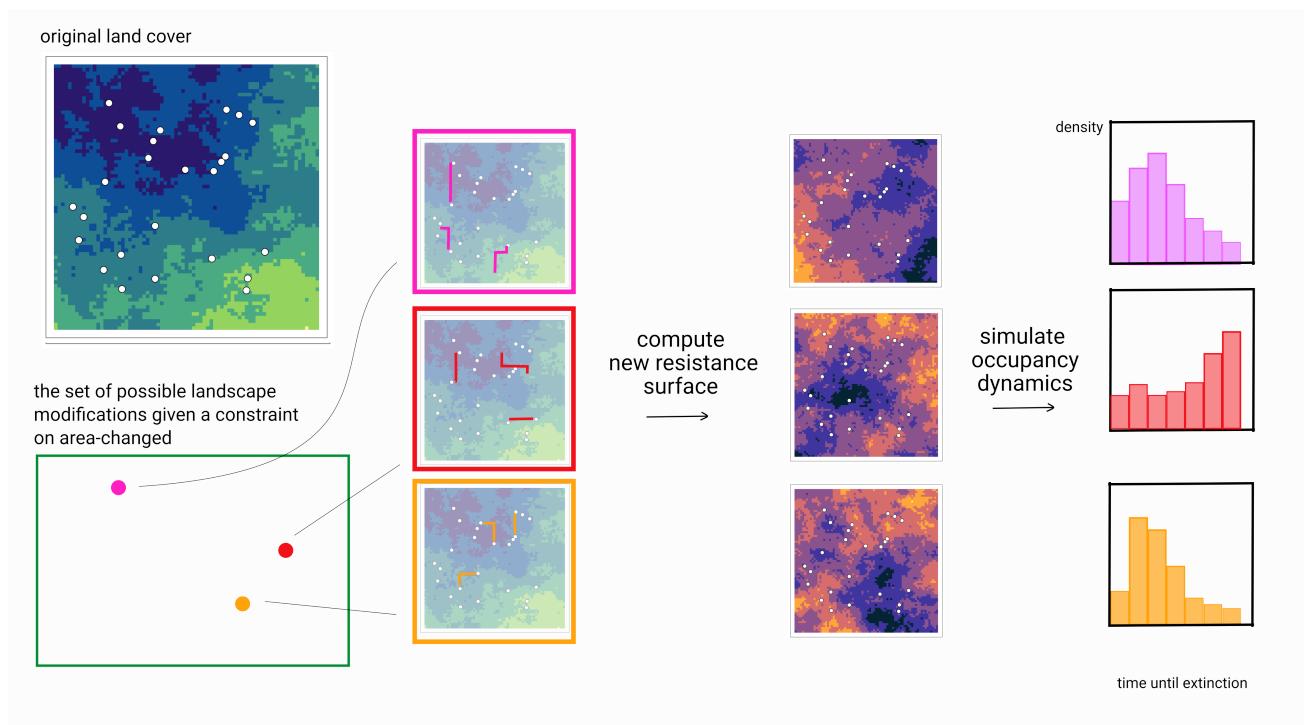


Figure 8: foo

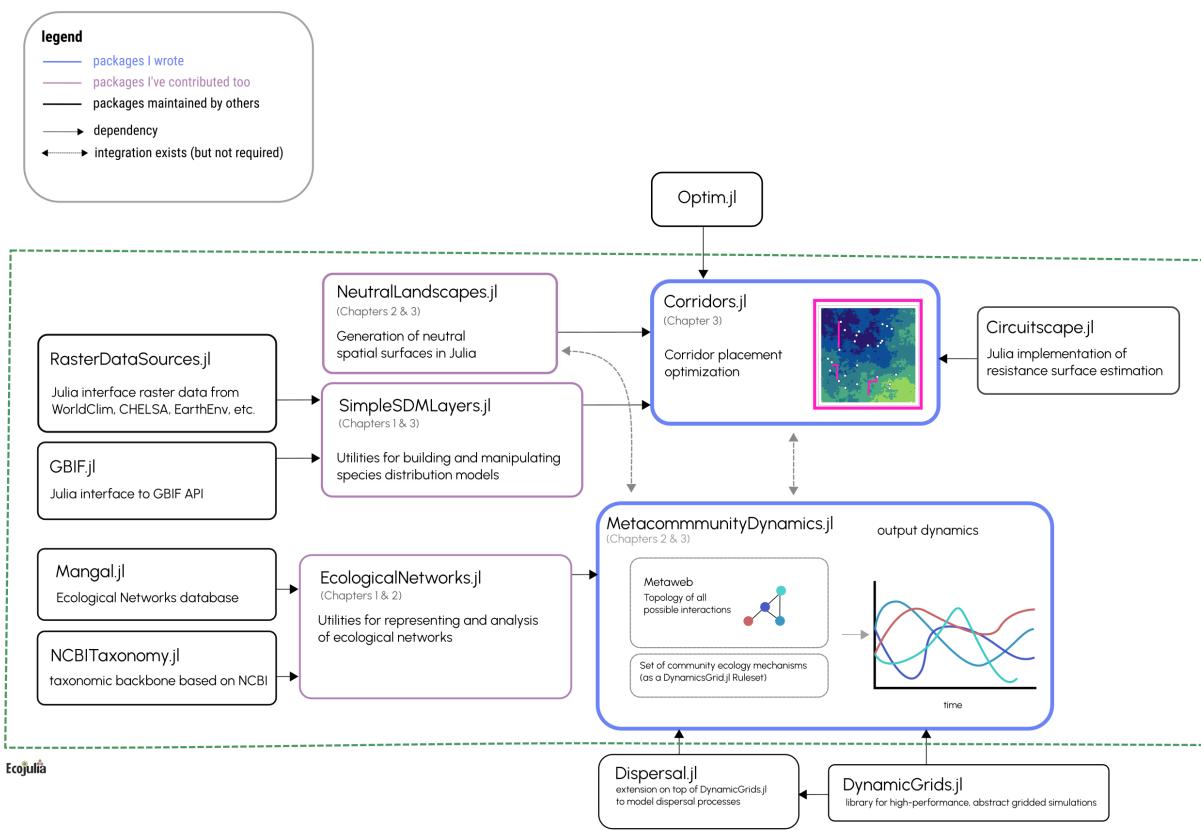


Figure 9: The structure of the software libraries used as part of MCD.jl

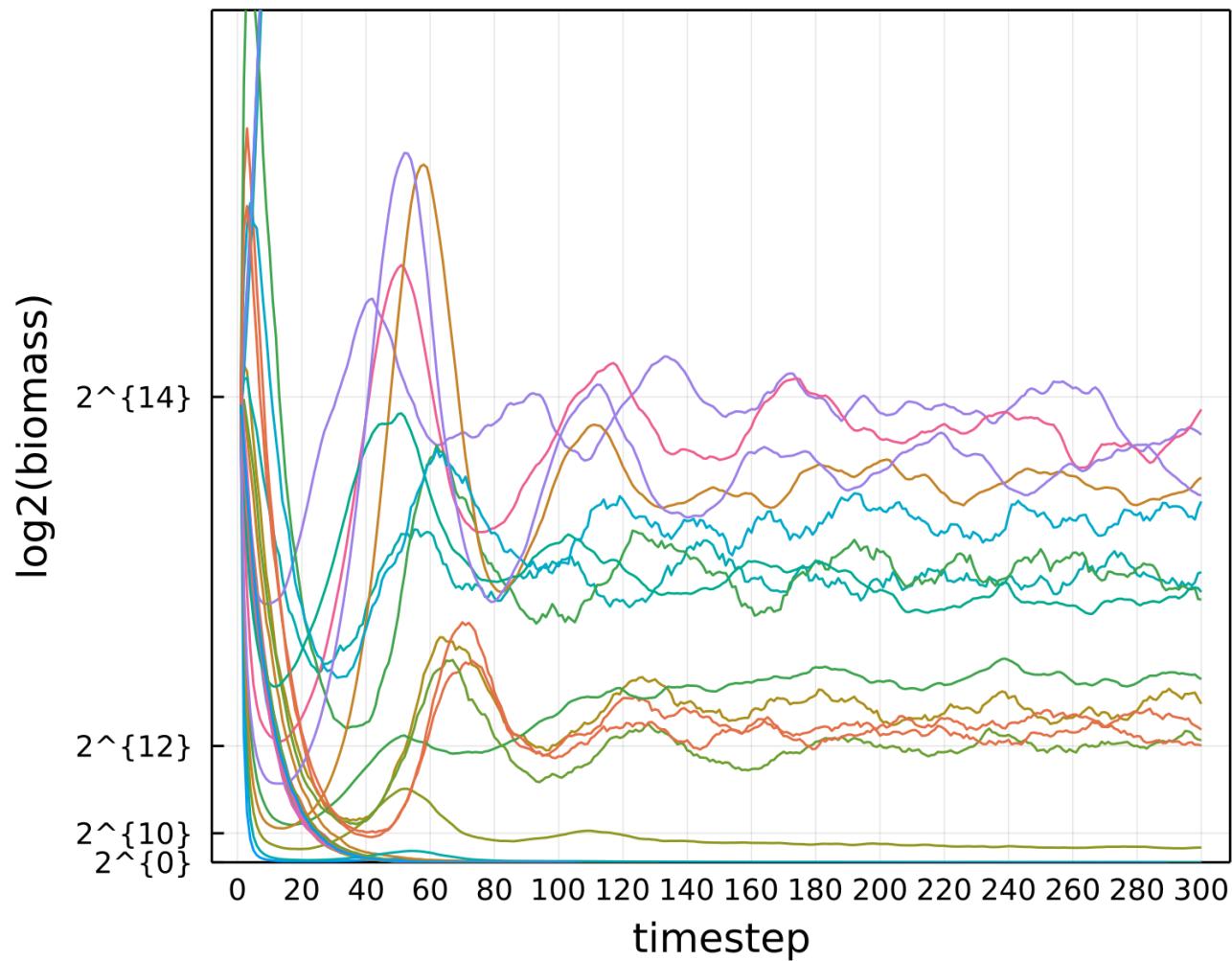


Figure 10: Sample output of simulated food web dynamics from MetacommunityDynamics.jl