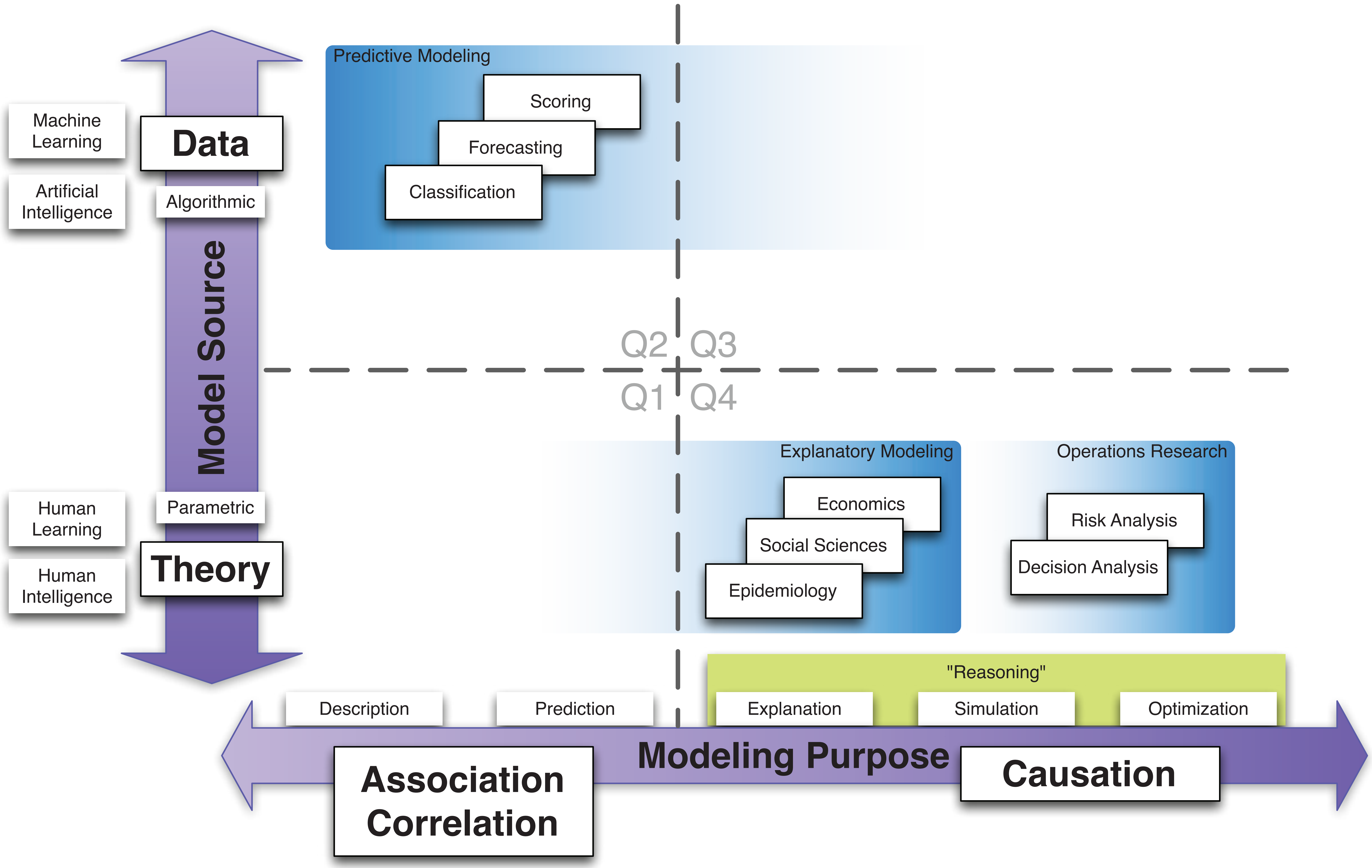
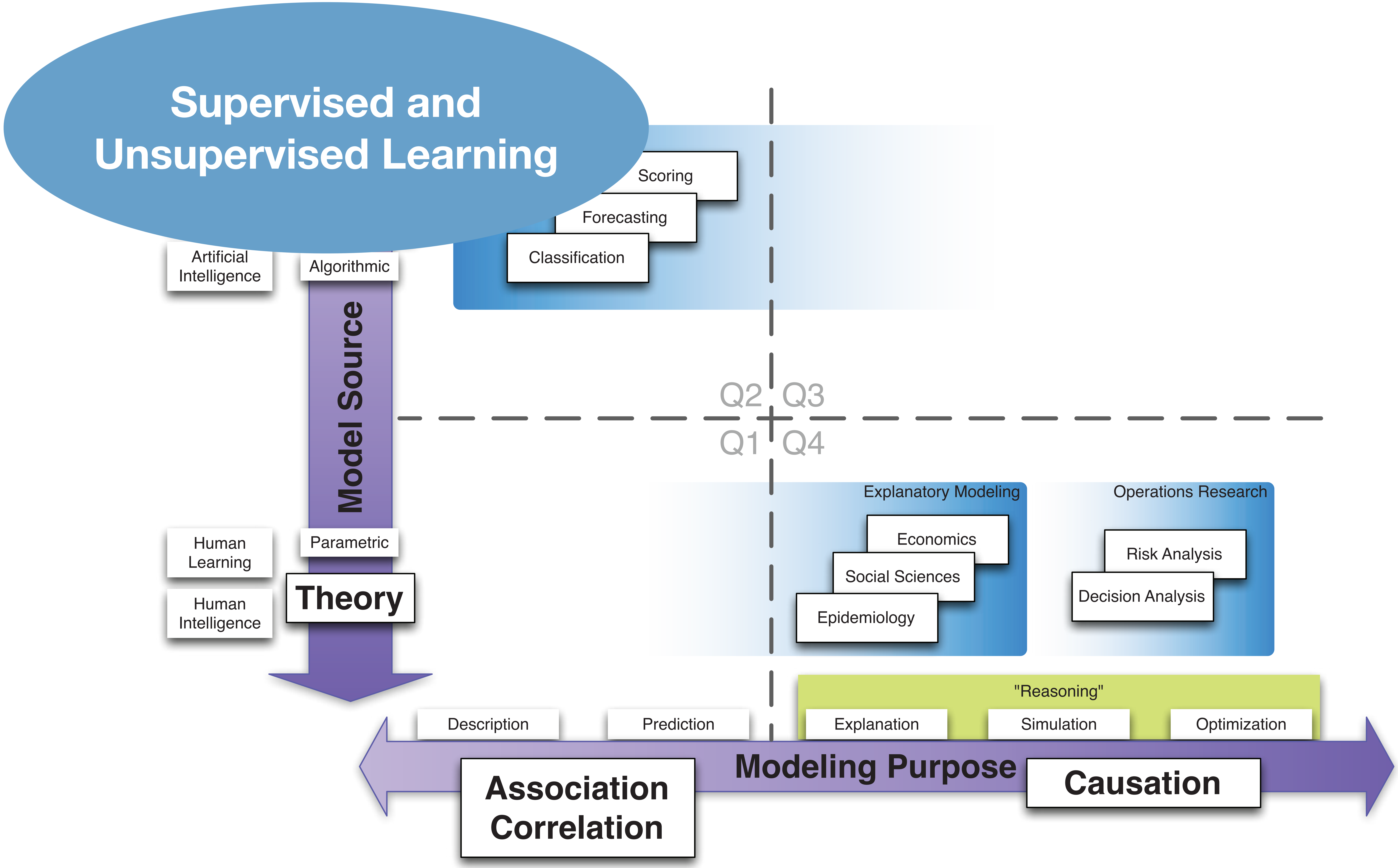
The Small Data Problem: Using Bayesian Networks in Endangered Species Policy Development

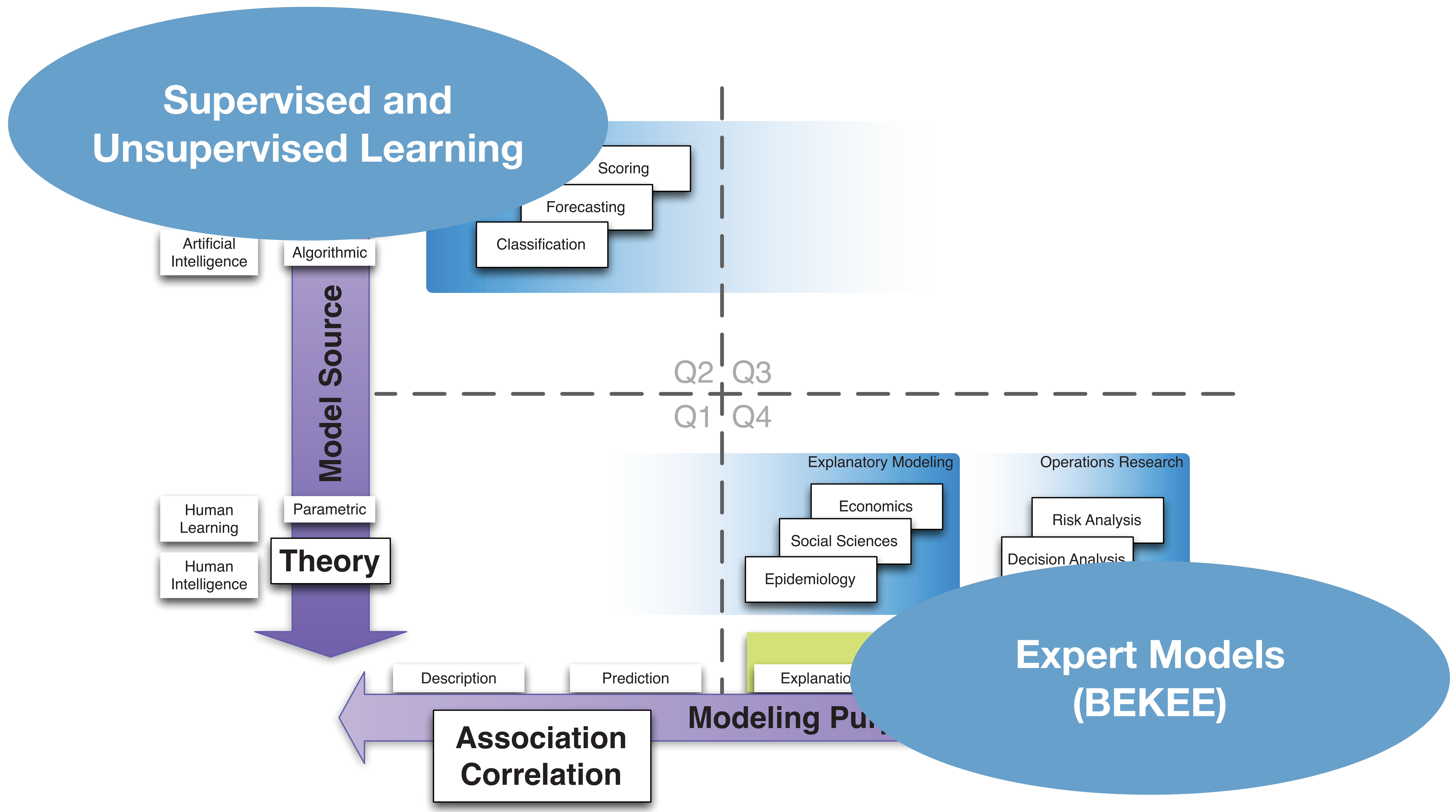
Steve Wilson, Ph.D.
Standpoint Decision Support Inc., Canada



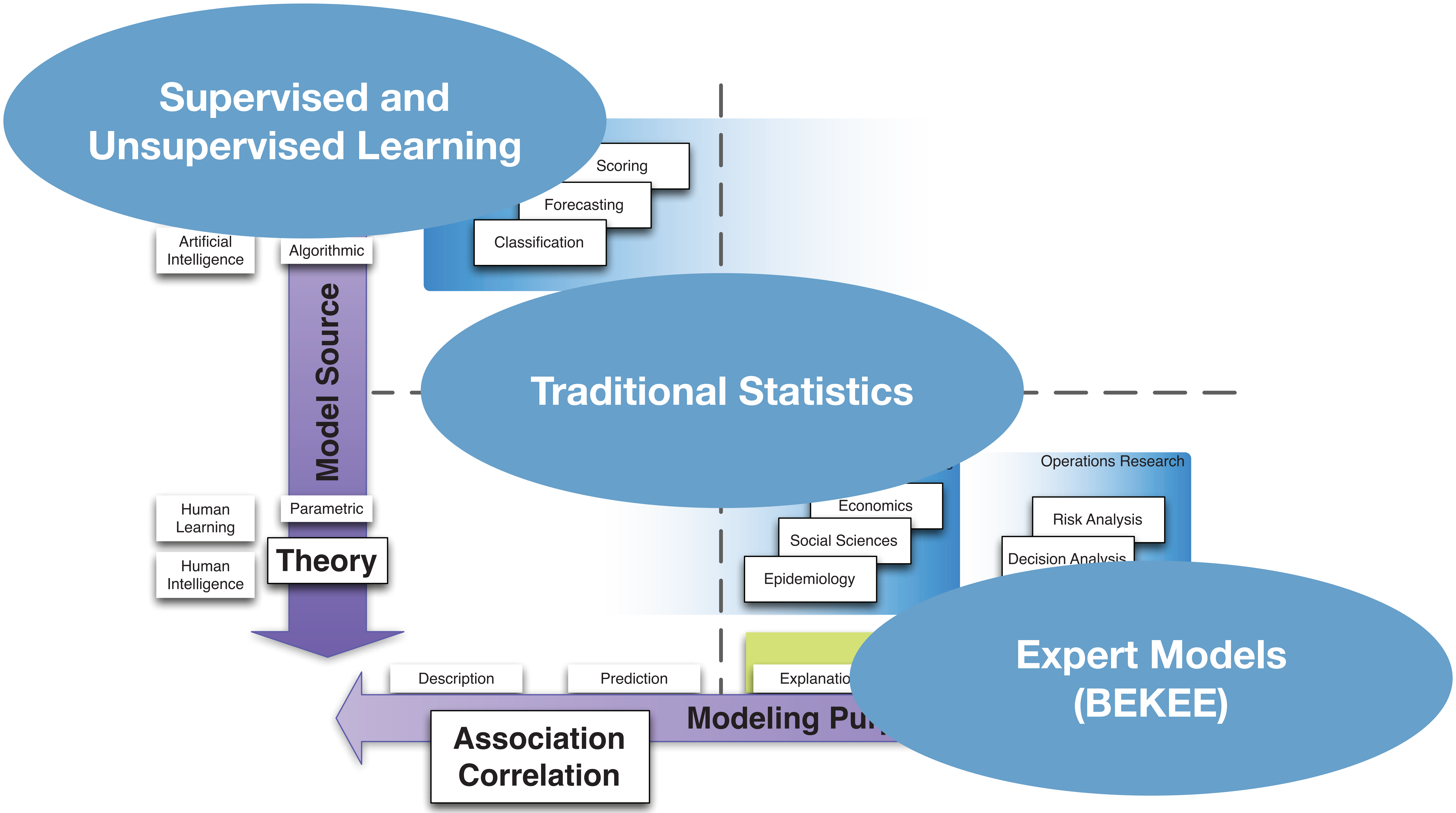
Conrady & Jouffe (2015)



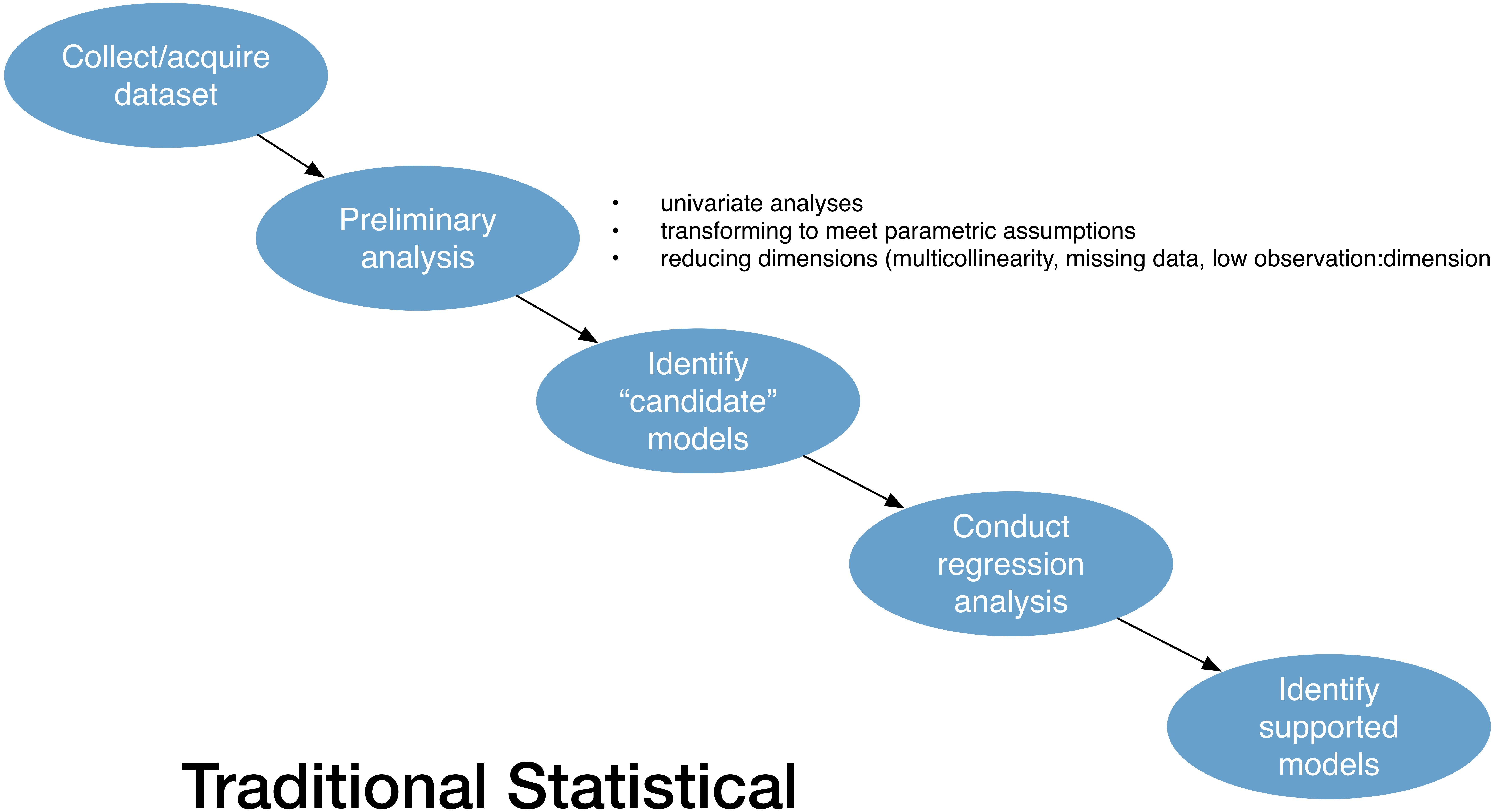
Conrady & Jouffe (2015)



Conrady & Jouffe (2015)



Conrady & Jouffe (2015)



- univariate analyses
- transforming to meet parametric assumptions
- reducing dimensions (multicollinearity, missing data, low observation:dimension ratio)

- AIC or other information-theoretic scoring

Traditional Statistical Workflow (Observational Data)


Collect/acquire dataset
Preliminary analysis
• univariate analyses
• transforming to meet parametric assumptions
• reducing dimensions (multicollinearity, missing data, low observation:dimension ratio)
Identify "candidate" models
Conduct regression analysis
Identify supported models
• AIC or other information-theoretic scoring
Make stuff up about causes
Implement made-up stuff as policy
Traditional Statistical Workflow (Observational Data)

"… I see no greater impediment to scientific progress than the prevailing practice of focusing all our mathematical resources on probabilistic and statistical inferences while leaving causal considerations to the mercy of intuition and good judgment."

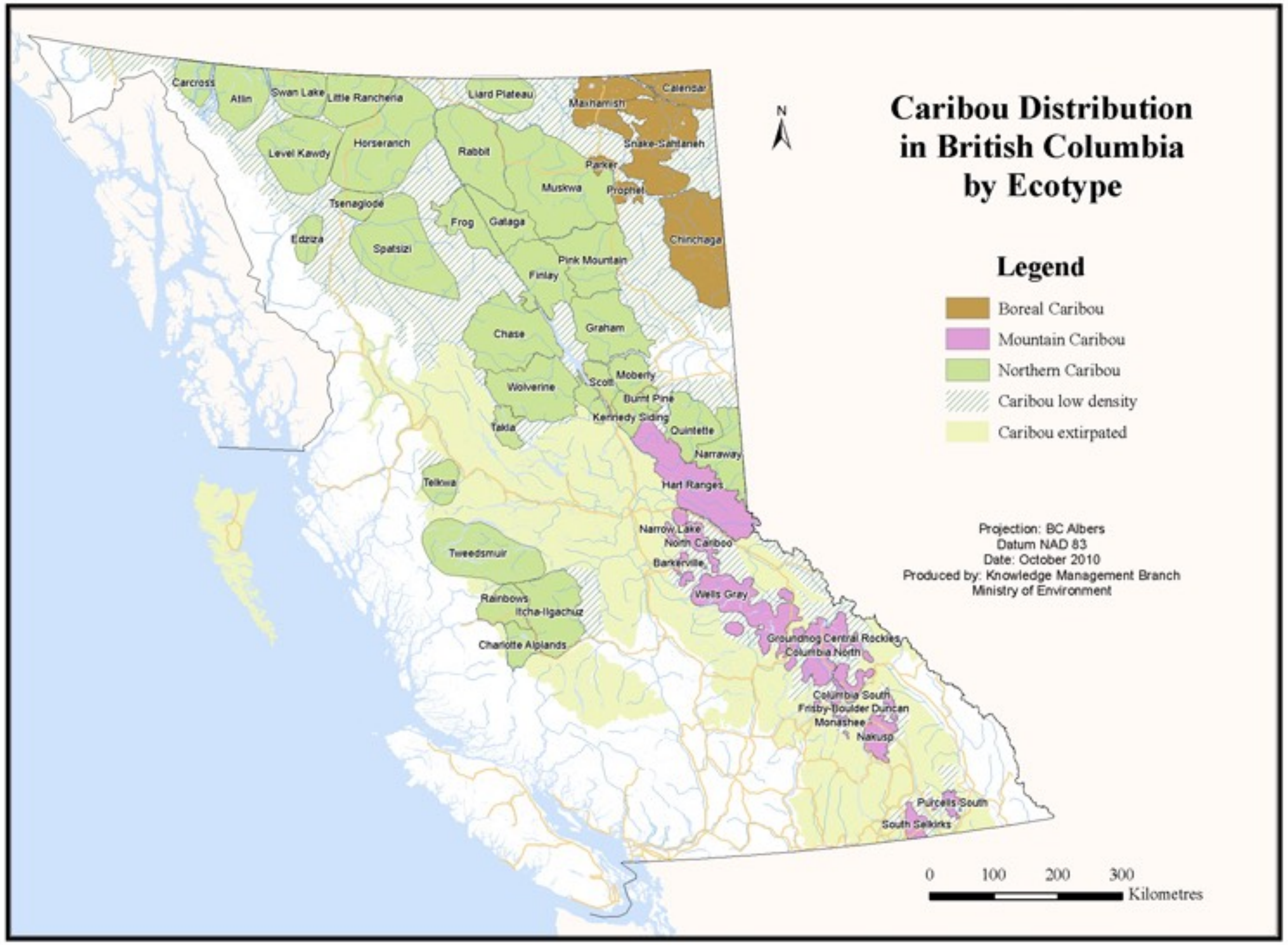
Pearl (1999)

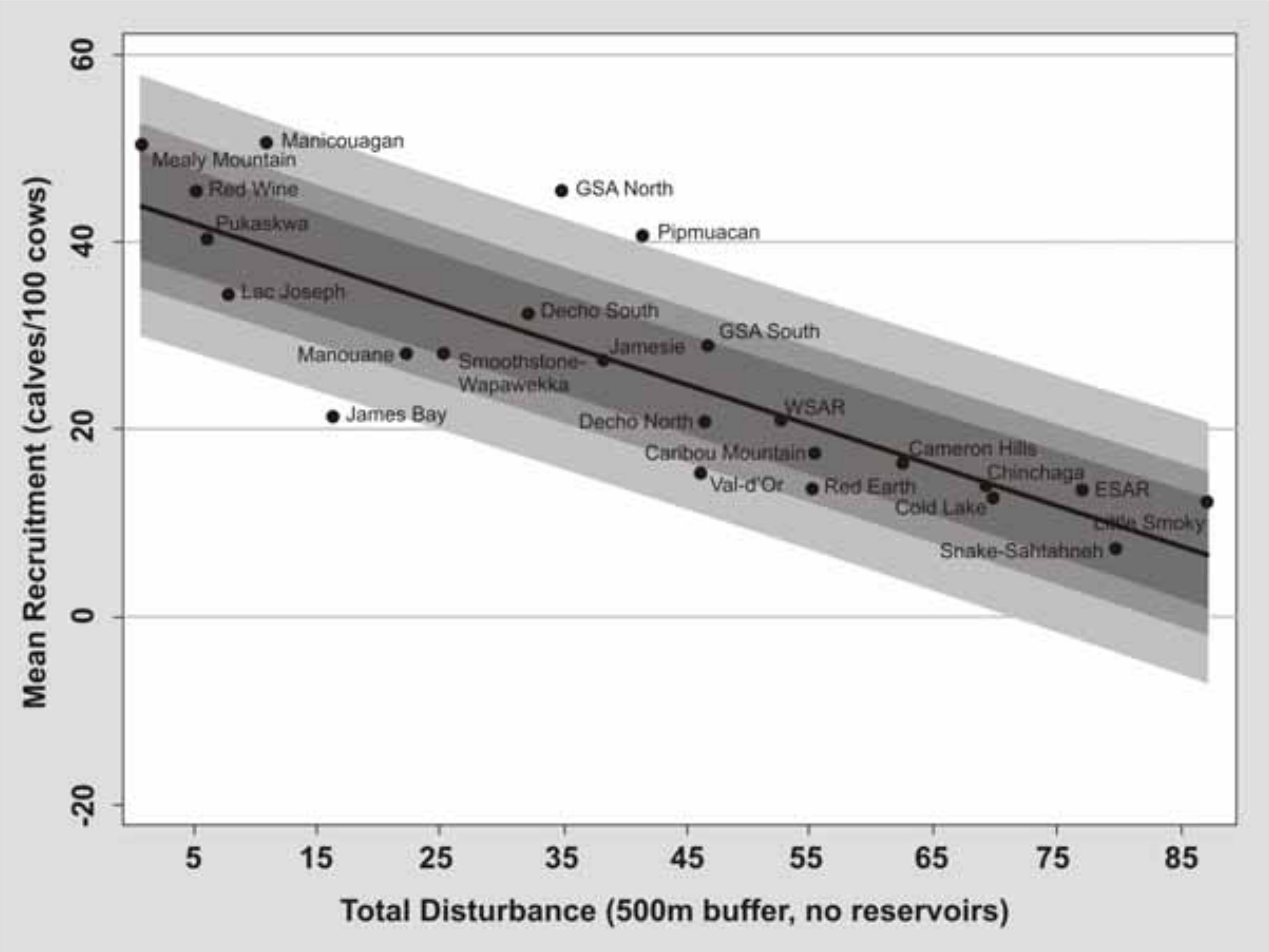
Problems with the Traditional Workflow

- No causal analysis!
- The analytically weakest results drive policy
- Inferences are data-limited

“Small Data” Problems

- Many dimensions
- Few observations
- RCT experiments impractical

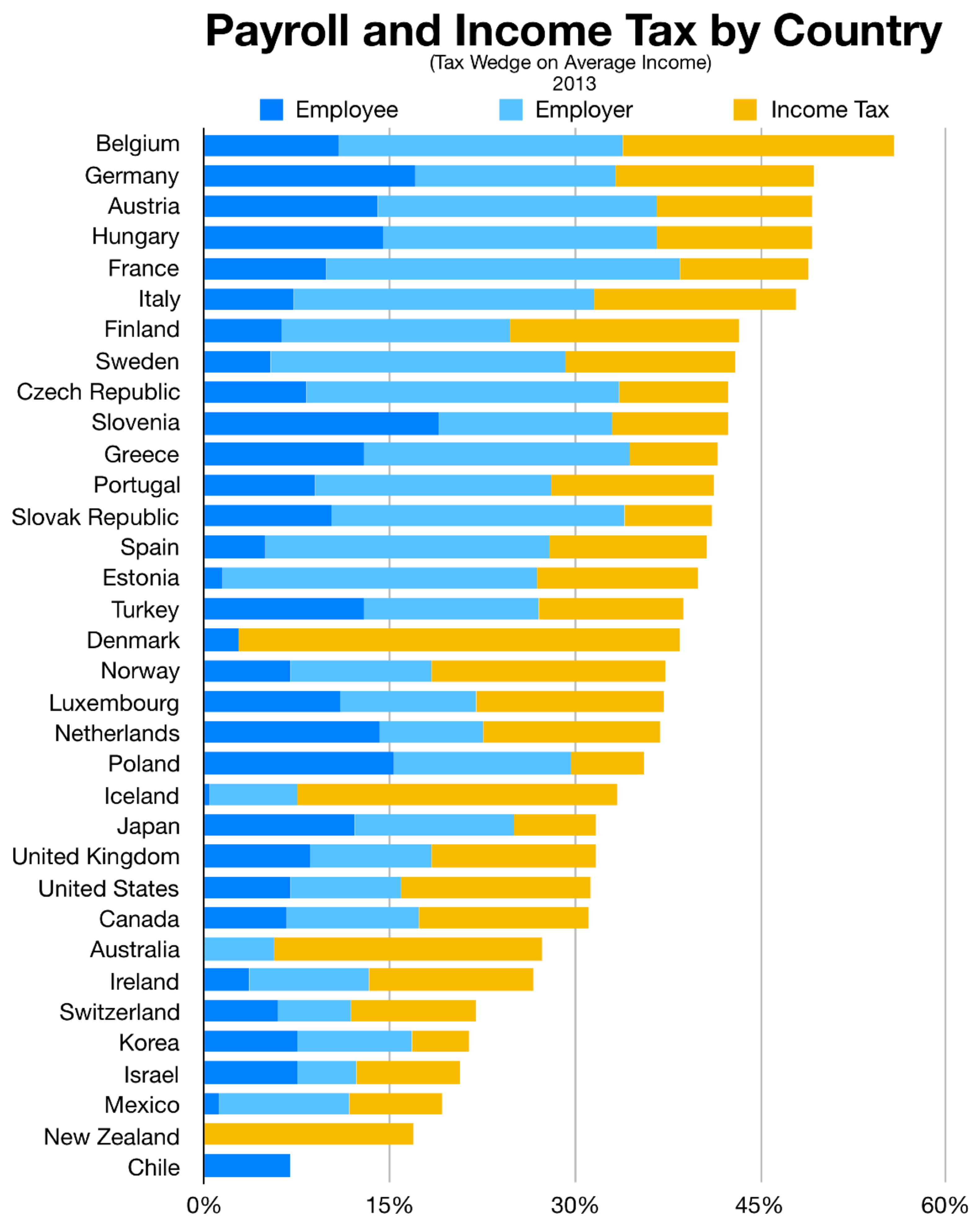




- “We only had 24 data points, so we couldn’t consider any other variates in our analysis”
- “That might be and important factor but collecting data on it is infeasible”
- But look at the r-squared!

“Small Data” Problems

- Jurisdictions
- Sports teams
- Medical trials
- Endangered species
- Climate change



Med Sci Sports Exerc. 1993 Jan;25(1):127-31.

Effect of time zone and game time changes on team performance: National Football League.

Jehue R, Street D, Huizenga R.

Abstract

To determine the effect of time zone and game time changes on NFL team performance, win-loss records from 1978-1987 were analyzed. Twenty-seven NFL teams were grouped by time zone and possible anti-jet lag adjustments. Among all intra-time zone rivals, home teams won 56.6%, away teams won 43.8%, for a home vs away winning percentage change of -12.8% ($P < 0.001$). West teams ($N = 5$) displayed fluctuations in home vs away team performance in association with trans-meridian travel. The change in winning percentage was found to be 0.0% vs West teams, -14.1% vs Central teams ($N = 8$) ($P < 0.05$), -16.3% vs East ($N = 14$) ($P < 0.05$) for West teams ($N = 4$) flying about 42 h pregame and +2.3% vs East for the one West team advancing practices 3-4 h to match East coast game time in addition to 48 h pregame flights. For night games within the same time zone, home vs away team winning percentage changed -23.8% ($P < 0.01$). West teams displayed uniformly high home winning percentages (75.0% and 68.4%) when playing Central and East teams, respectively, with little or no fall in away winning percentages (67.7% and 68.8%). For day games, a 3-h phase advance may decrease West coast team performance. In one small subset, anti-jet lag adjustments appeared to eliminate the expected decrement in performance. For night games, West coast teams, whether home or away, appear to be at a distinct advantage over East and Central teams.

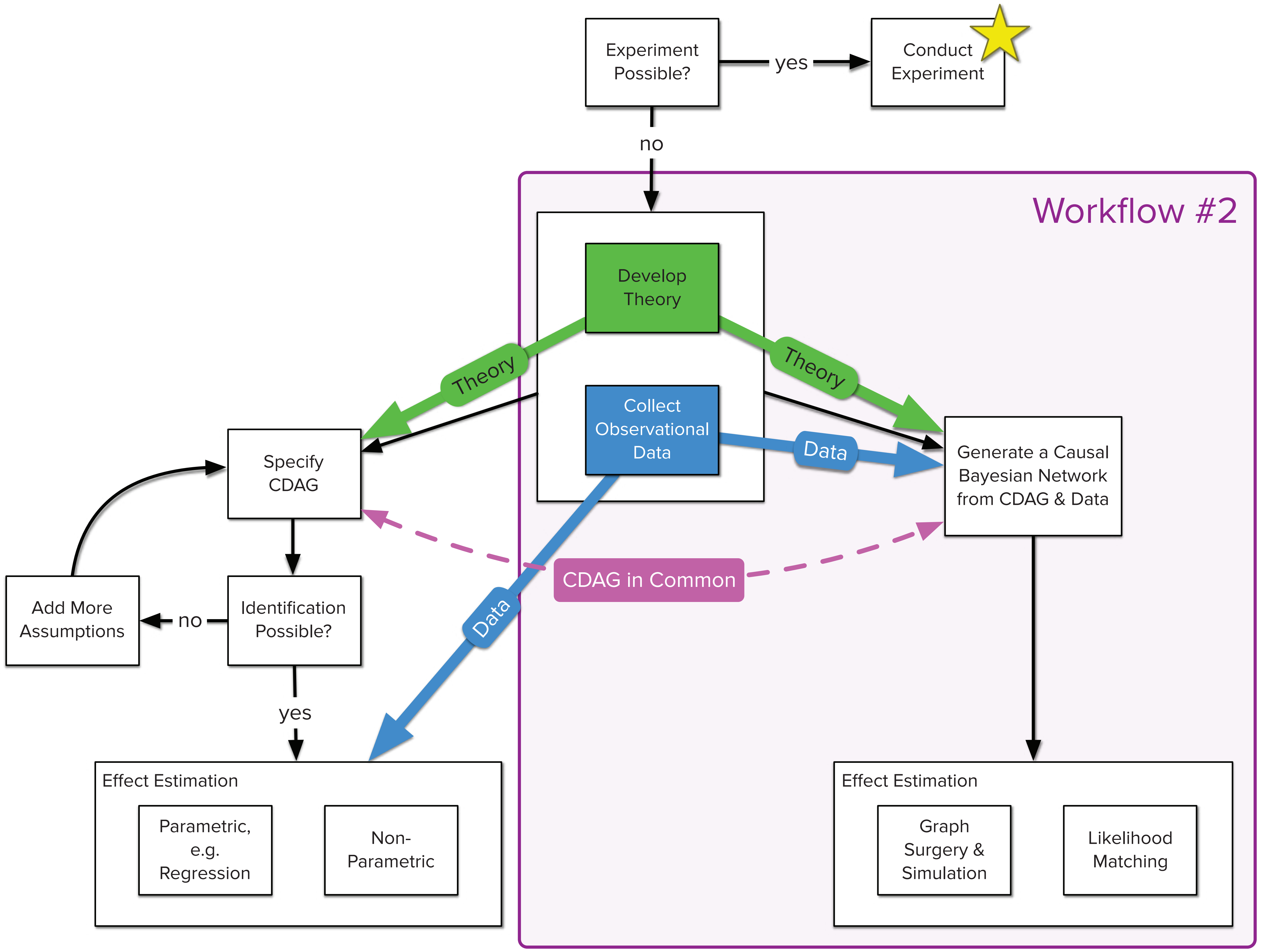
What's the Alternative Workflow?

- Conduct a formal causal analysis
- Address the small data problem

The Data-Free Option

- **Build and parameterize a model based on expert elicitation**
- Experts want models to reflect available data, but this can lead to biases
- “Small experts” can be as big a problem to manage as “small data”
- Empiricism trumps expertise

What's our Workflow?



Conrady & Jouffe (2015)

Draft the casual DAG

Causal identification

Specify with expert knowledge

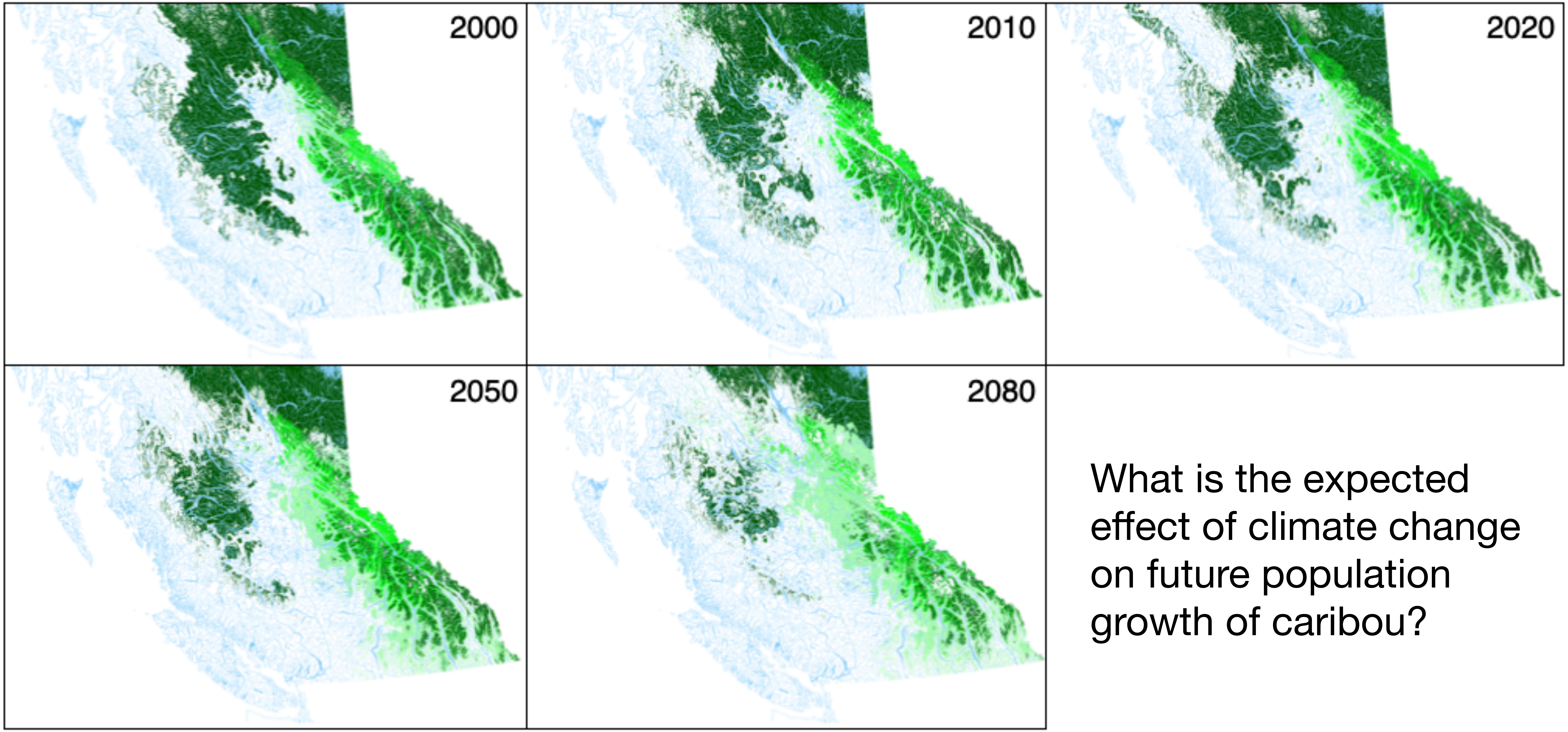
Update parameters with available data

Review for internal consistency

Estimate effects

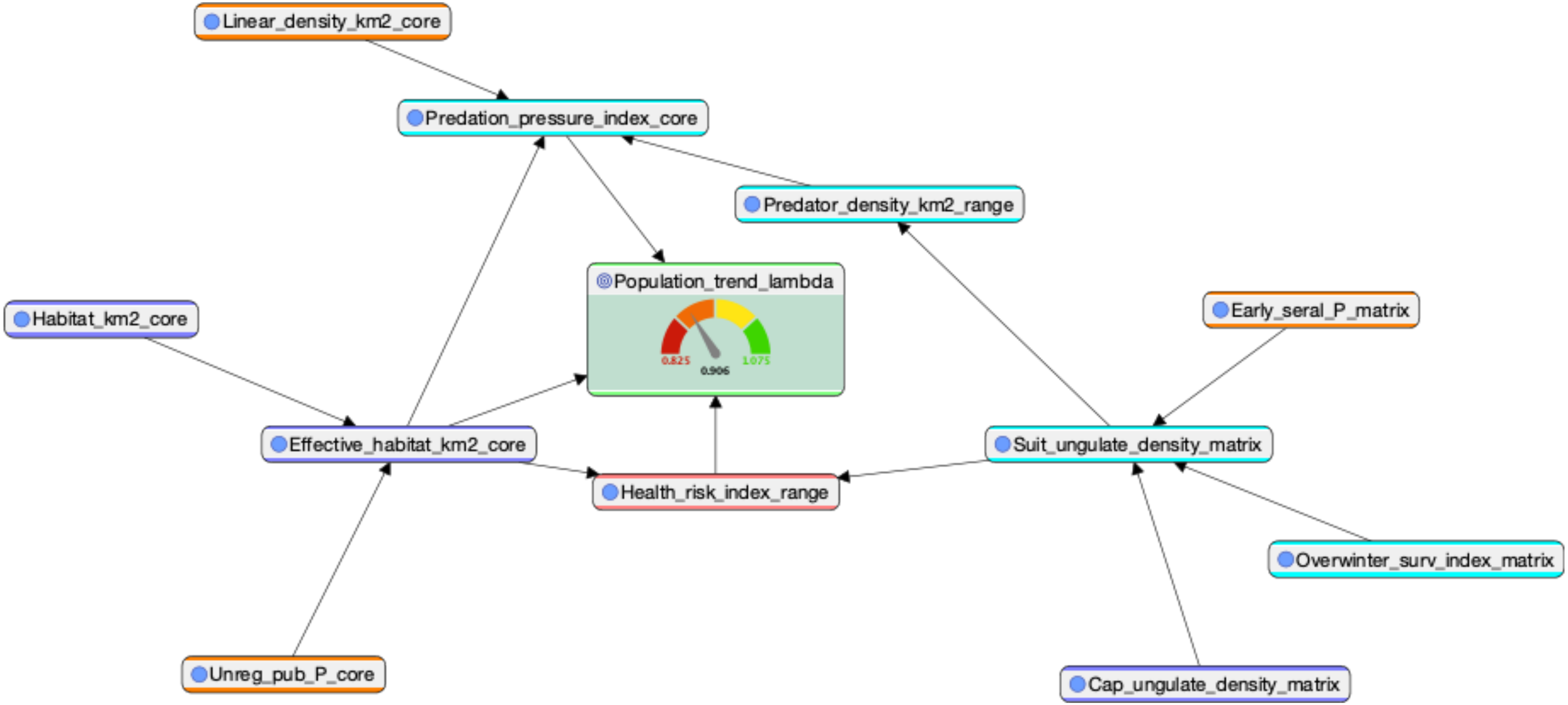
What’s our Workflow?



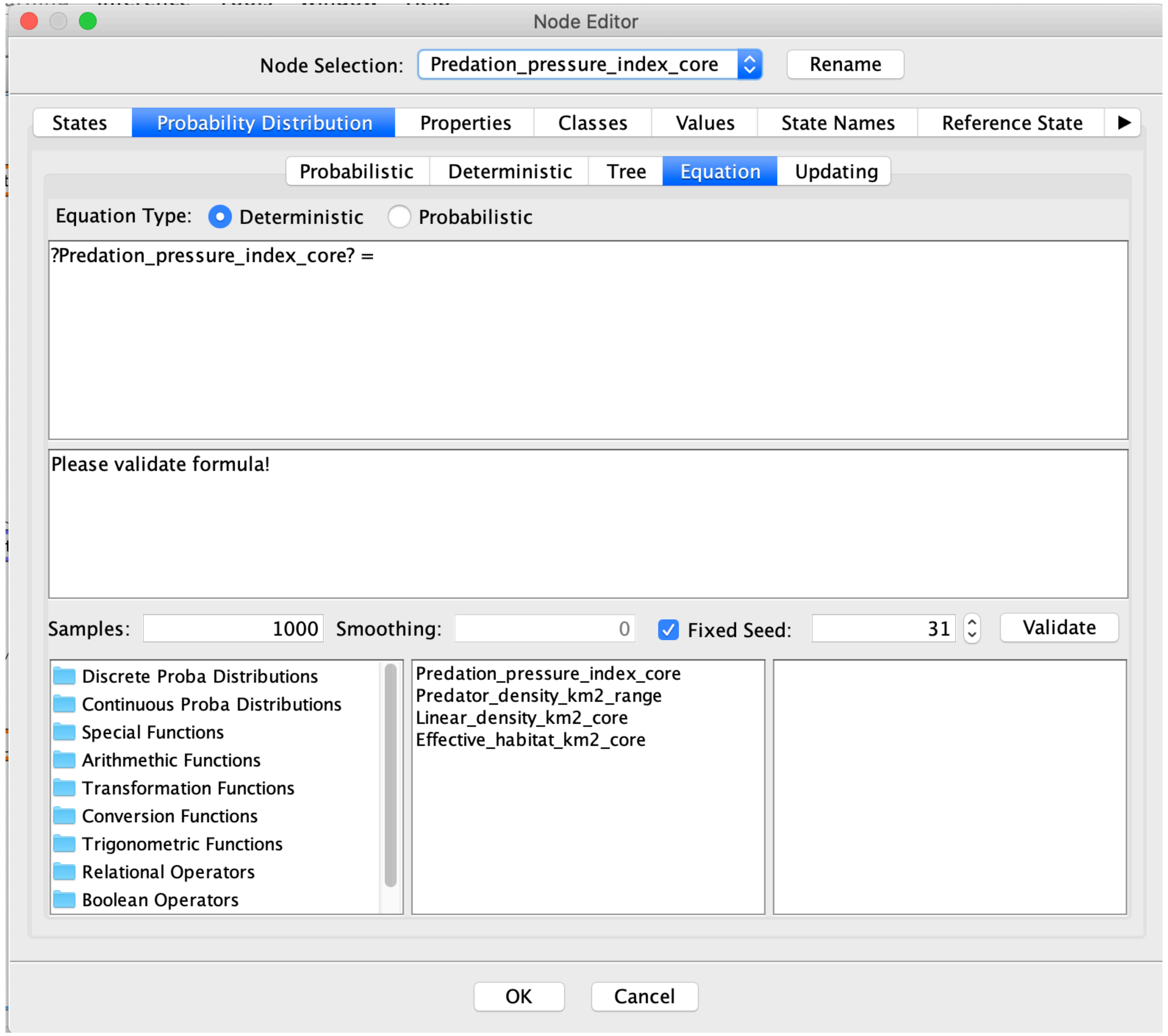


What is the expected effect of climate change on future population growth of caribou?

Developing the Causal Model



Specify Priors

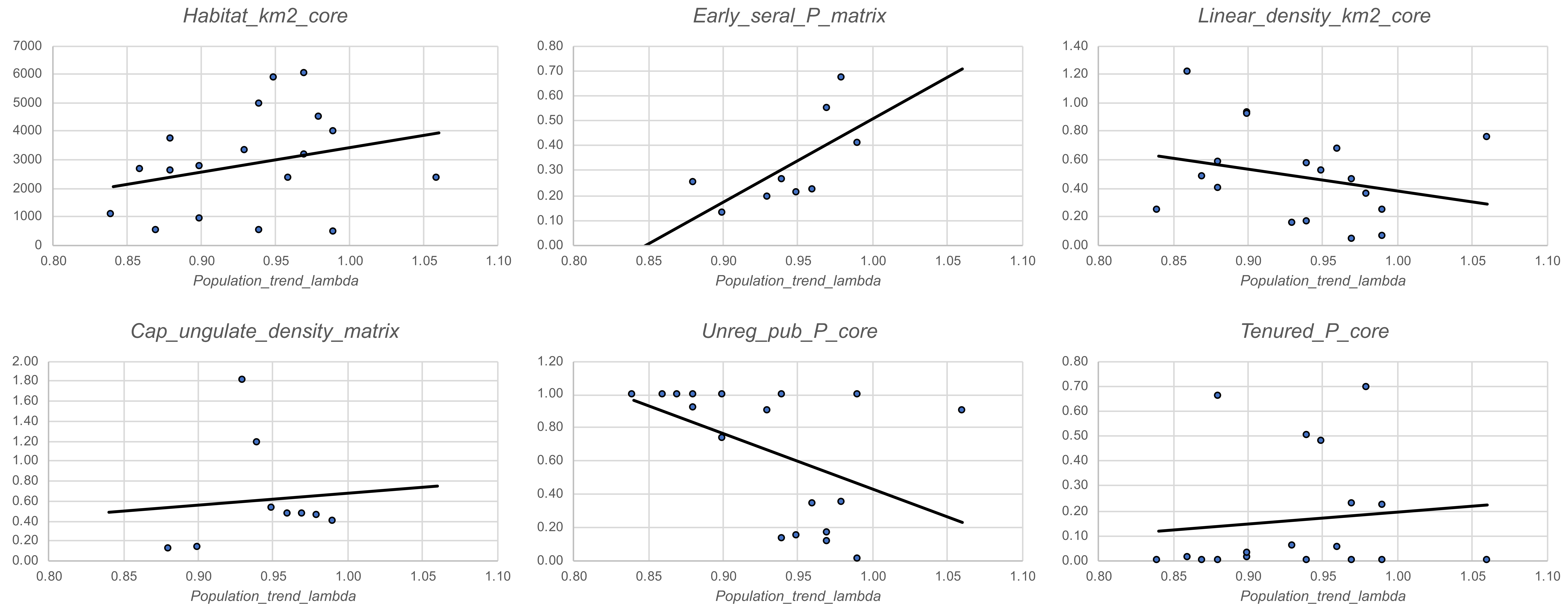


Specify Priors

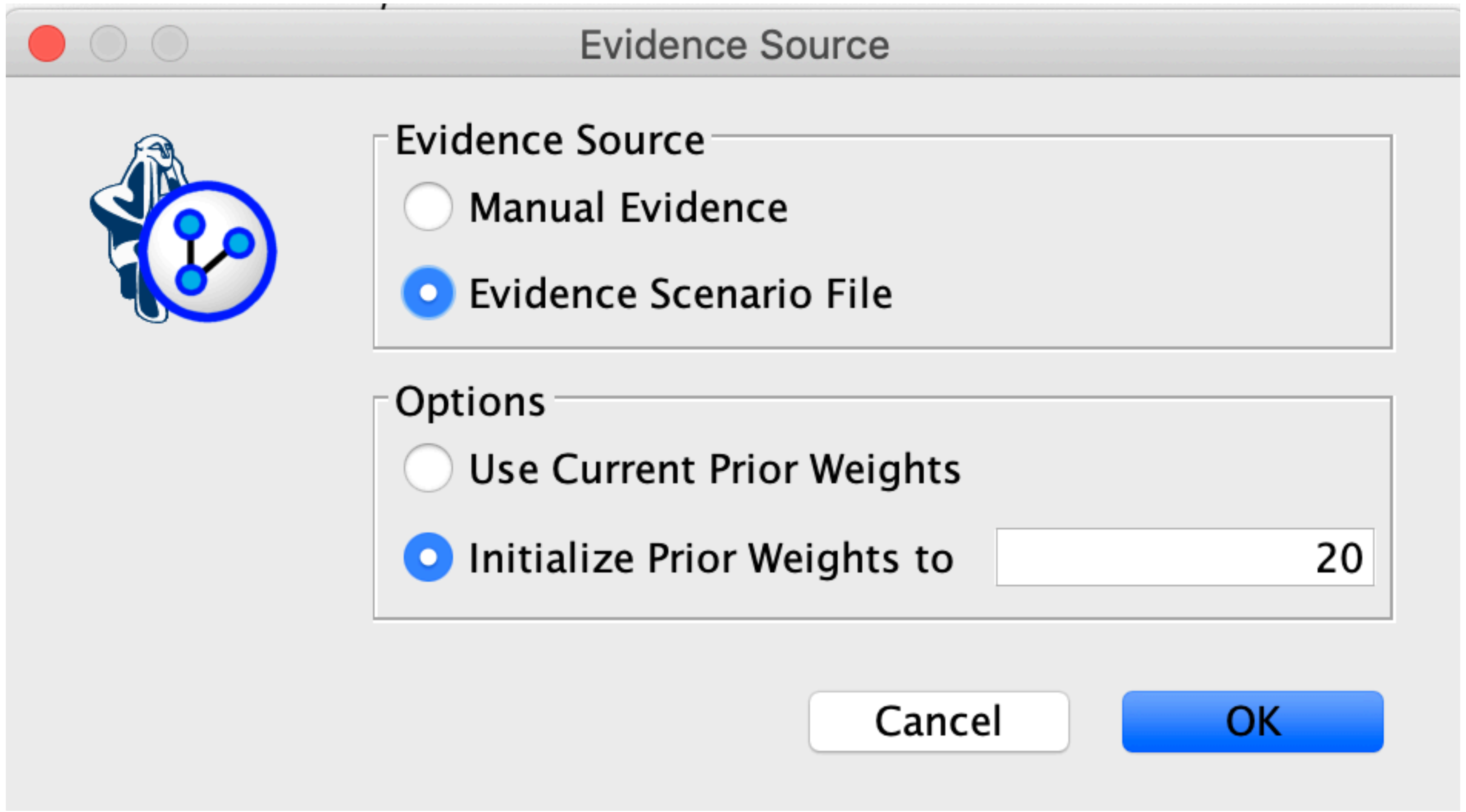
Factor 1	Factor 2	Low	Moderate	High
Low	Low	99.000	0.500	0.500
	Moderate	49.500	49.500	1.000
	High	49.500	1.000	49.500
Moderate	Low	49.500	49.500	1.000
	Moderate	0.500	99.000	0.500
	High	1.000	49.500	49.500
High	Low	49.500	1.000	49.500
	Moderate	1.000	49.500	49.500
	High	0.500	0.500	99.000

- Use 3-4 states to capture non-linearities
- Capture assumed direction of the relationship & the relative importance of parents
- Assign plausible intervals to states

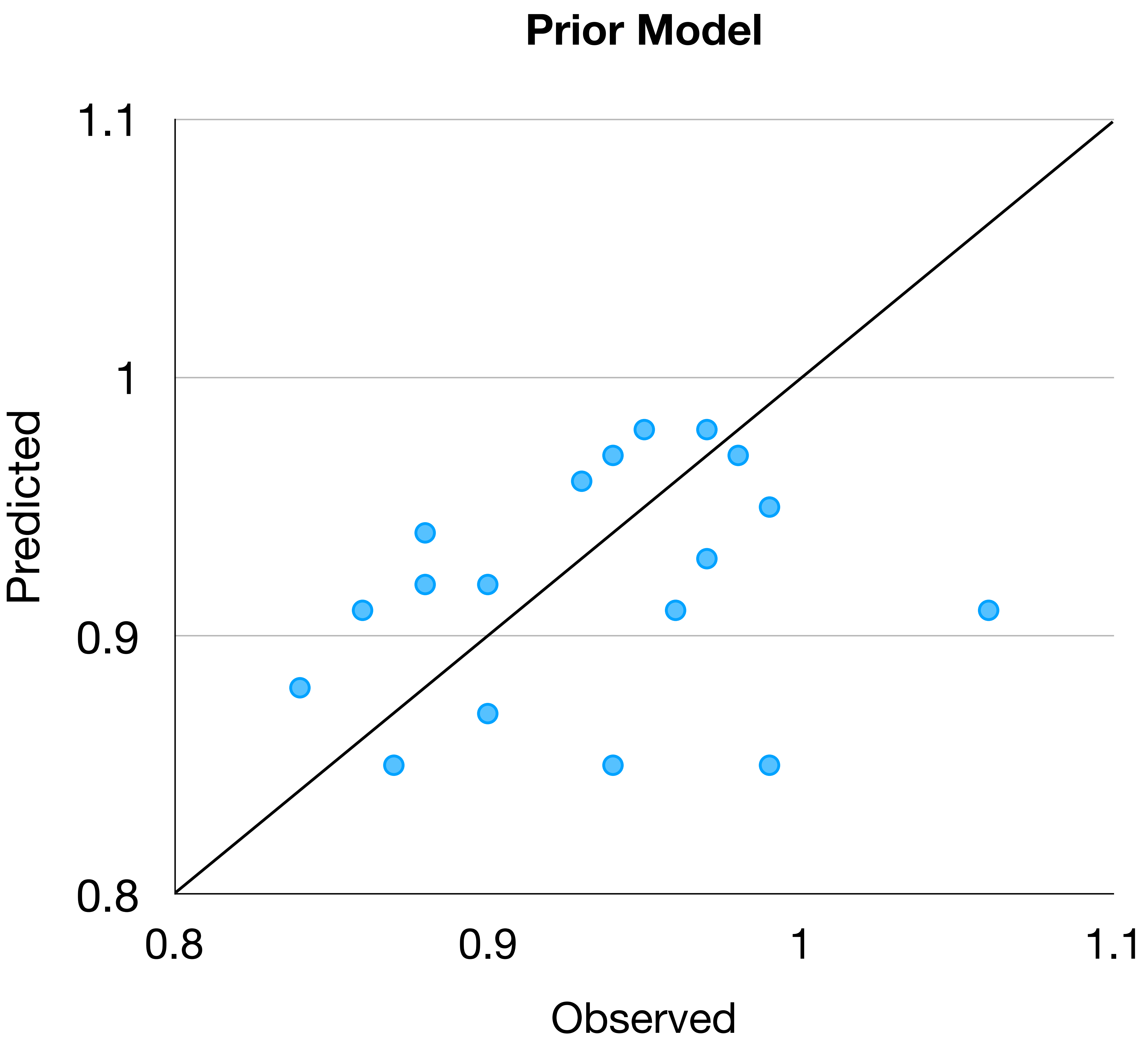
Update Parameters

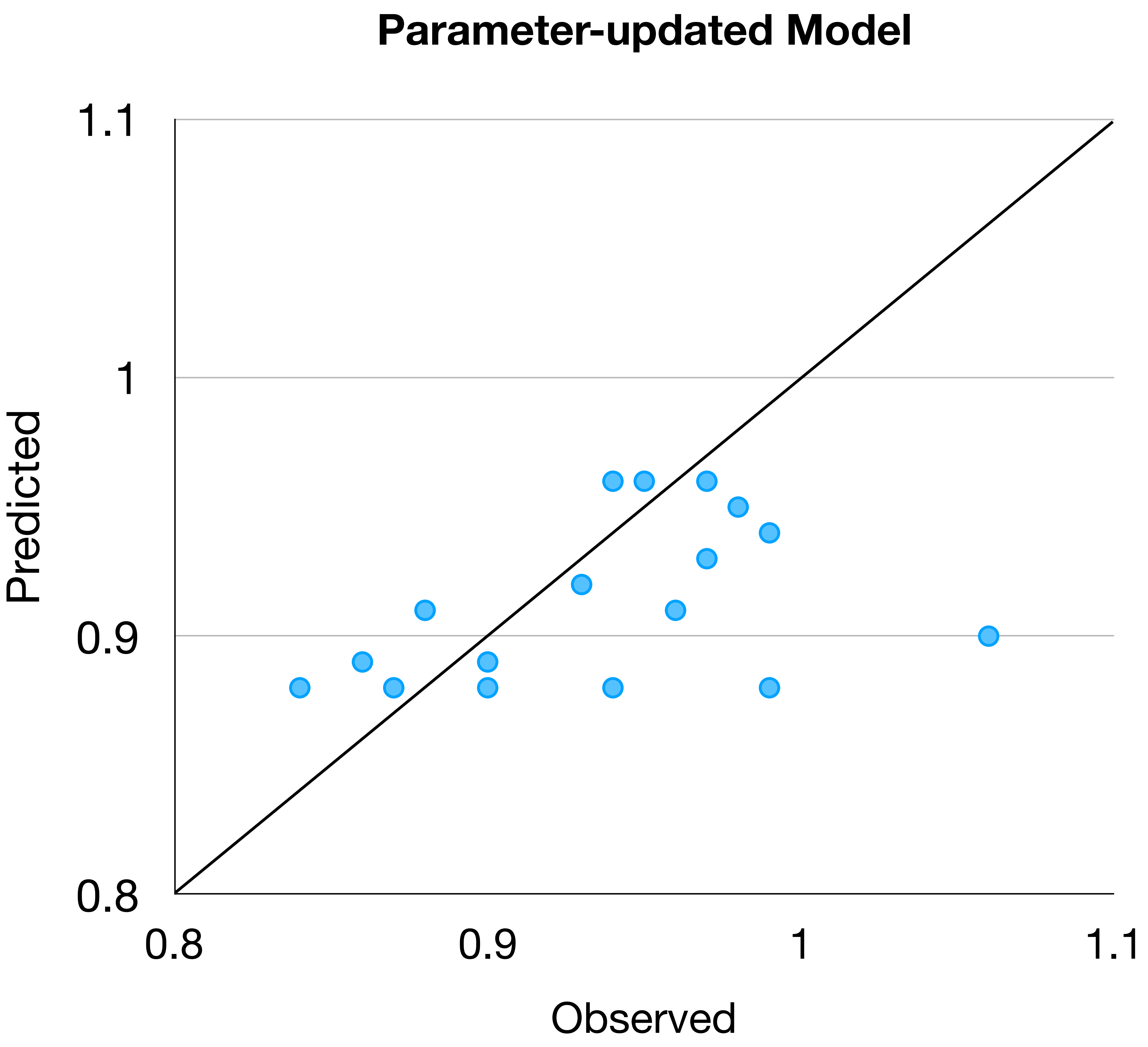


Inference>Parameter Updating

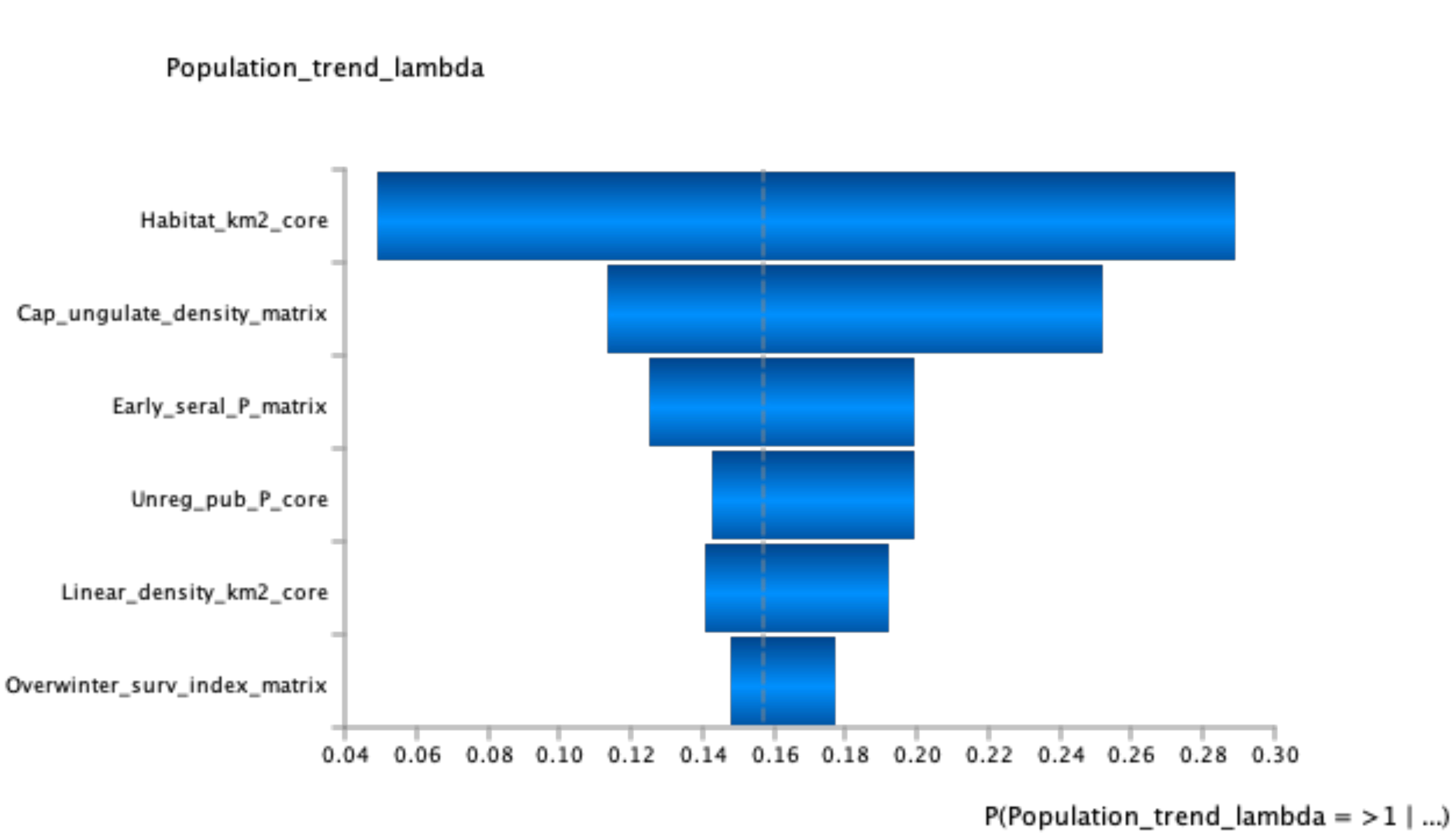


Internal Consistency

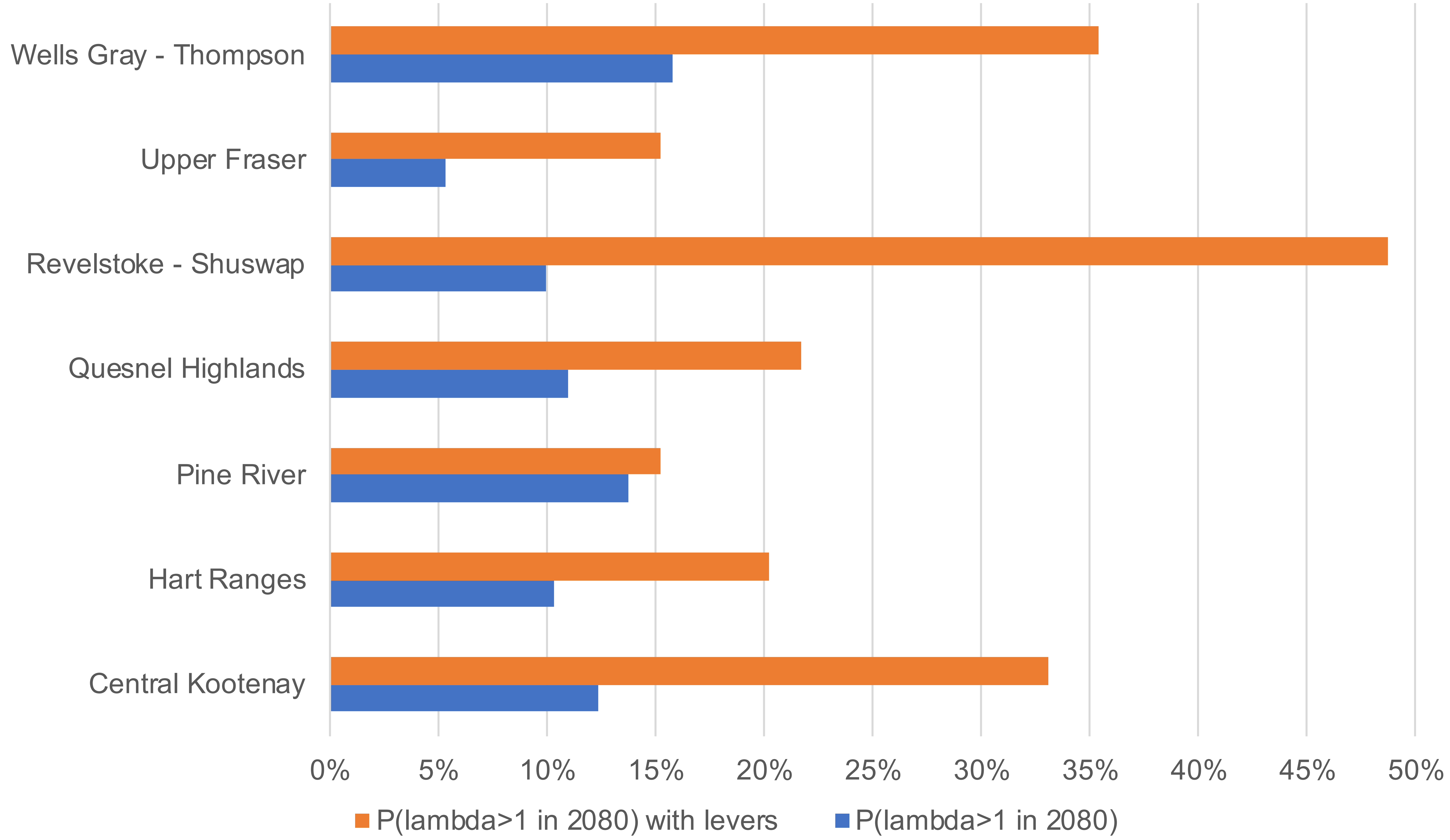


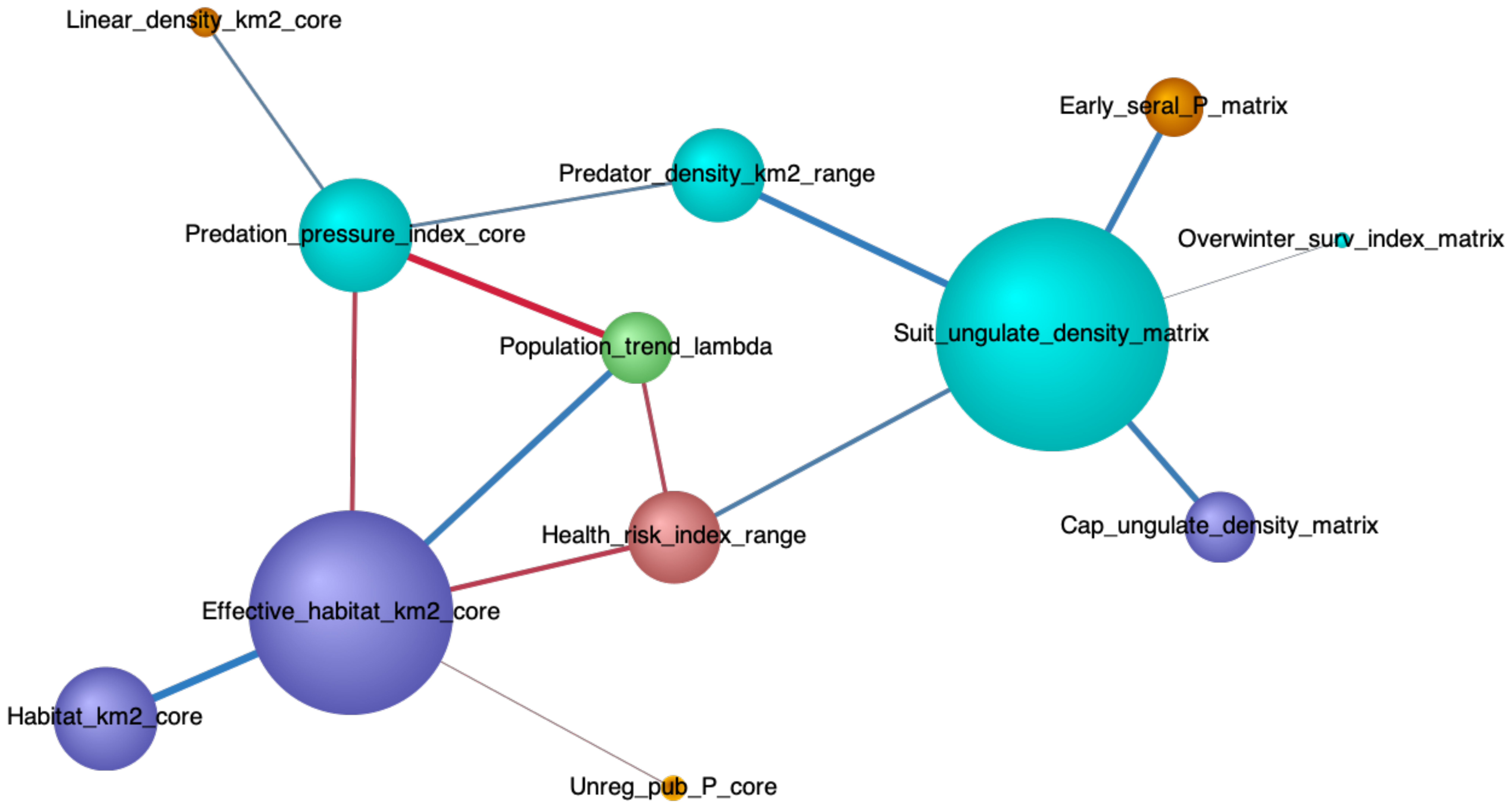


Estimate Effects



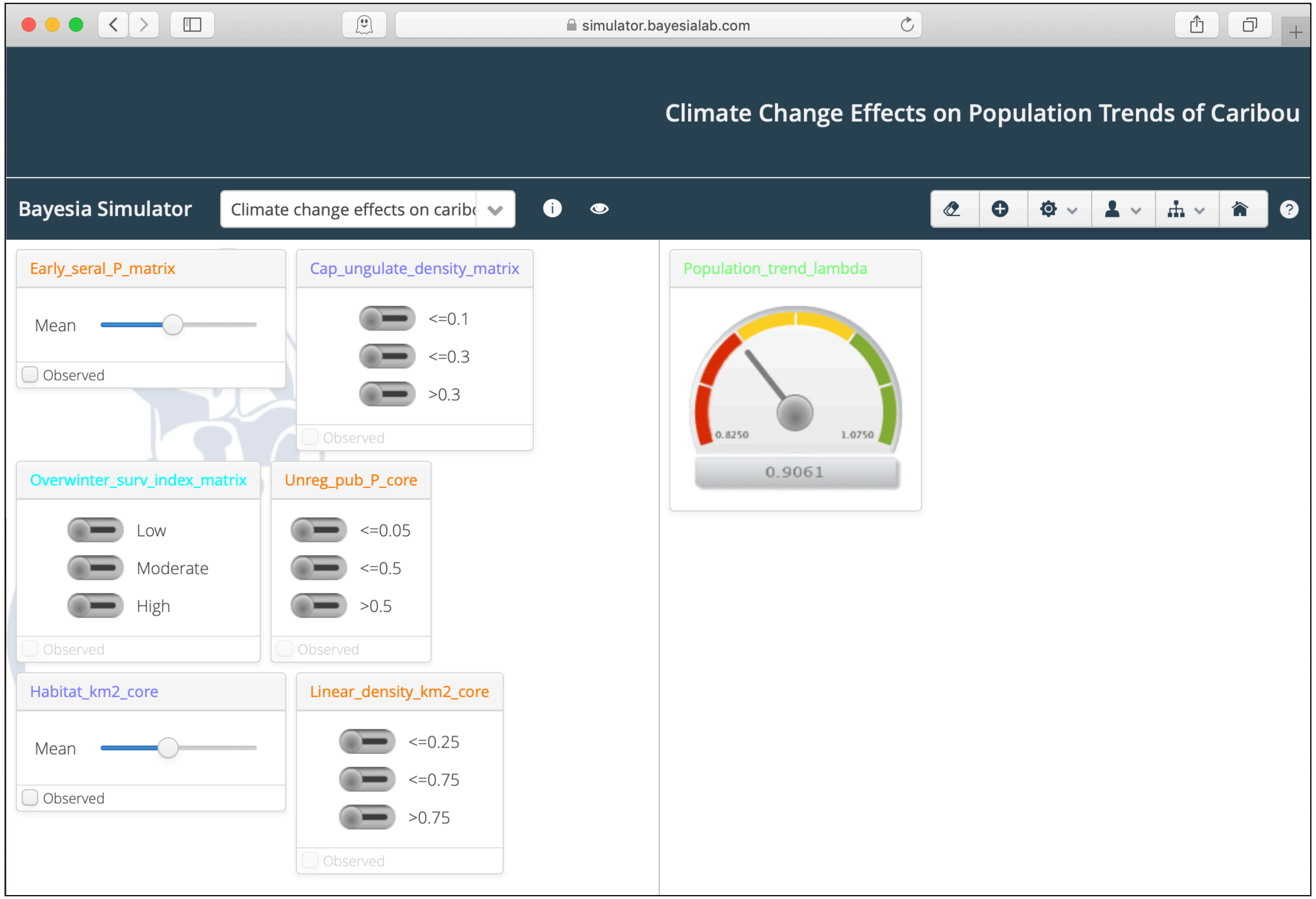
Effects of Management





Linear_density_km2_core
Early_seral_P_matrix
Predator_density_km2_range
Overwinter_surv_index_matrix
Predation_pressure_index_core
Population_trend_lambda
Suit_ungulate_density_matrix
Health_risk_index_range
Cap_ungulate_density_matrix
Effective_habitat_km2_core
Habitat_km2_core
Unreg_pub_P_core

Simulator



simulator.bayesialab.com
Climate Change Effects on Population Trends of Caribou
Bayesia Simulator
Climate change effects on caribo
Early_seral_P_matrix
Mean
Observed
Cap_ungulate_density_matrix
<=0.1
<=0.3
>0.3
Observed
Population_trend_lambda
0.8250
1.0750
0.9061
Overwinter_surv_index_matrix
Low
Moderate
High
Observed
Unreg_pub_P_core
<=0.05
<=0.5
>0.5
Observed
Habitat_km2_core
Mean
Observed
Linear_density_km2_core
<=0.25
<=0.75
>0.75
Observed

Conclusions

- Evolving workflow aimed at replacing the traditional statistical approach with a causal paradigm

- Blend expert knowledge and data to address the “small data” problem