# Towards Synthetic Ecosystems of Disease Vectors

Elaine Liu [Department of Statistics, Carnegie Mellon University], Abby Smith [Department of Statistics, Carnegie Mellon University], Samuel Ventura [Department of Statistics, Carnegie Mellon University]

INTRODUCTION

INFECTIOUS DISEASE VECTORS

- **Vector (Epidemiology):** any agent (animal, or microorganism) that carries and transmits an infectious pathogen into another living organism

- **Common Vectors:** hematophagous arthropods
  - **Mosquitos** Dengue, Chikungunya, Zika, Mayaro and Yellow Fever
  - **Ticks** Lyme, Rocky Mountain Spotted Fever
  - Flies, Sand Flies, Lice, Fleas
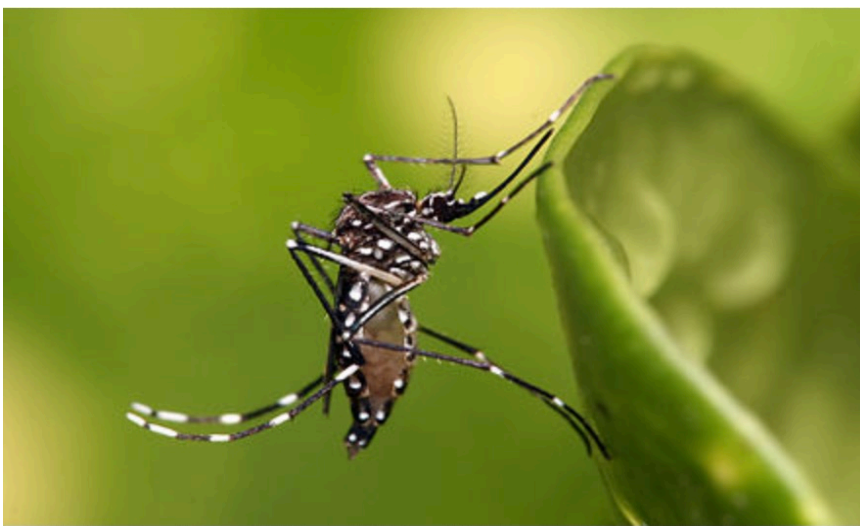
AGENT-BASED MODEL (ABM)

- **ABM:** Simulate autonomous agents' interactions in constrained environments [1]

- **Example ABM:** Synthetic Populations and Ecosystems of the World (SPEW) for studying the spread of infectious diseases among human populations [2]

- **Proposed Disease Pathogen Agent Modeling Framework:**
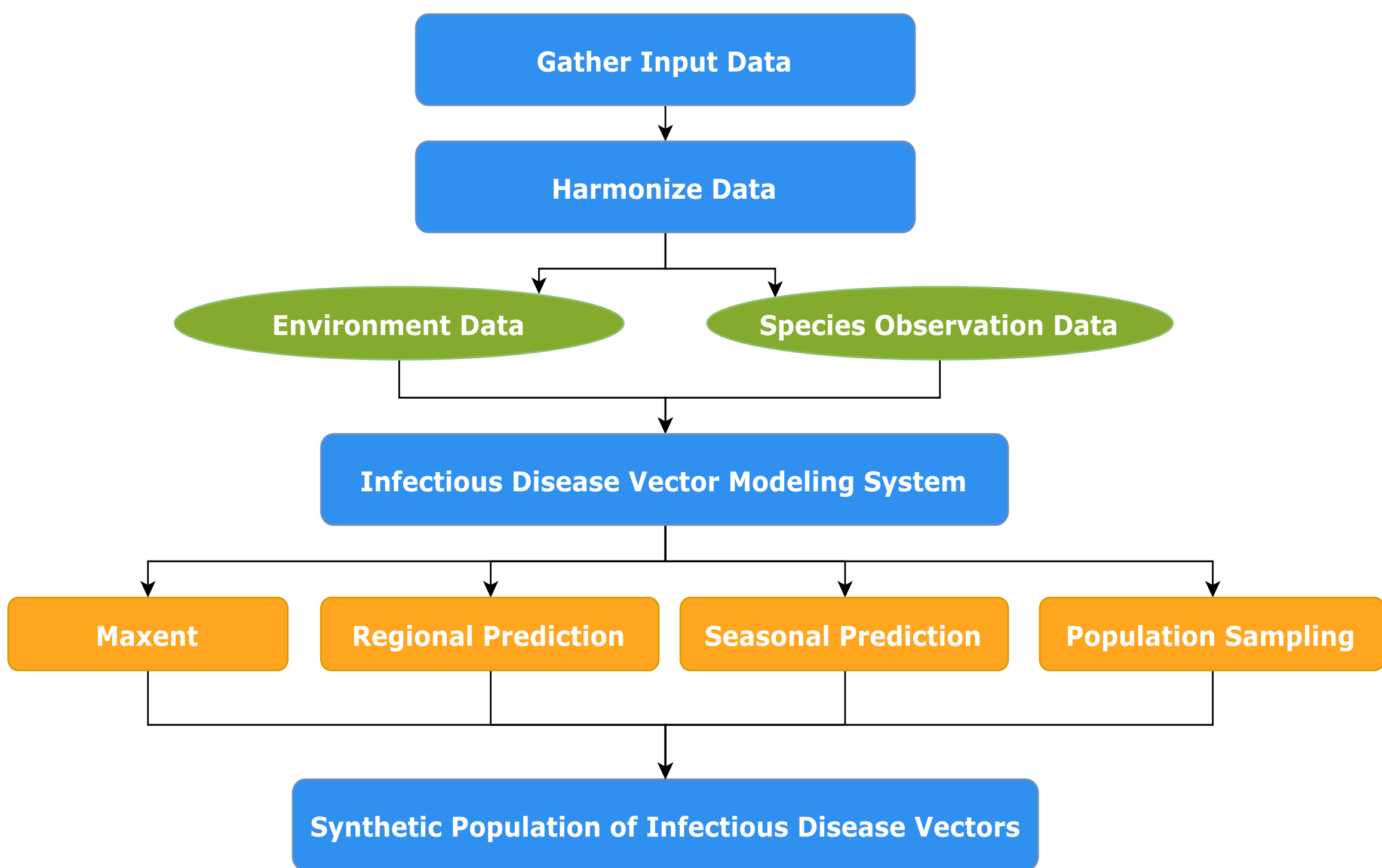

MAXENT

- **Modeling probability distribution for a species** based on the maximum-entropy approach [3]

- **Input Data:**
  - Species Presence Data: geographical locations of observed species from Global Biodiversity Information Facility (GBIF), which provides species observance data around the world.
  - Environment Factors: a set of global climate layers (gridded climate data) from WorldClim, which provides monthly precipitation and temperature information with the spatial resolution down to 1 km^2.

- **Output Data:** two types of probability grid cell outputs
  - **Raw Probability:** proportion of potential distribution each site represents
  - **Logistic Probability:** the probability of occurrence

$$P_{raw} = \frac{f_1(z)}{f(z)} = e^{\eta(z)} \qquad P_{logistic} = Pr(y=1|z) = \tau e^{\eta(z)-r}/(1-\tau+\tau e^{\eta(z)-r})$$

$f_1(z)$: conditional density at presence sites

$f(z)$: marginal density across the study area

$\eta(z) = \alpha + \beta * h(z)$, a linear score of feature vector $h(z)$

τ: probability of presence at sites with typical conditions for a species
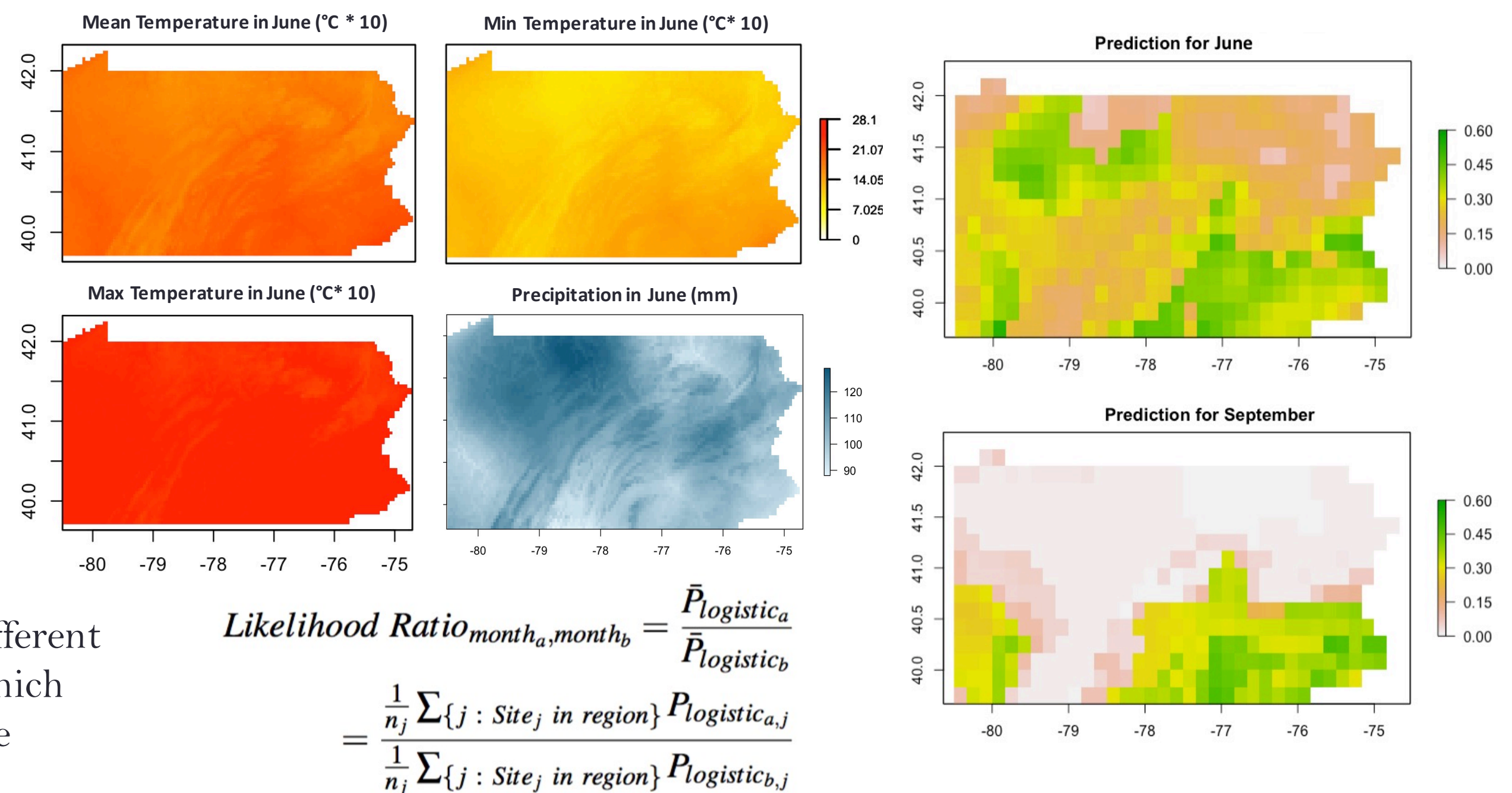
MODEL DETAILS

FROM GLOBAL TO LOCAL PREDICTIONS

- Take the name of a geographical region by a string vector
  - e.g. c("USA"), c("USA", "Pennsylvania"), c("USA", "Pennsylvania"), c("USA", "Pennsylvania", "Allegheny")
- Extract geographical boundaries from Global Administrative Areas (GADM) as spatial polygons
- Masking the prediction grids with geographical boundaries to extract regional distribution
- **Transform Raw Probability Predictions:**

$$P_{presence_i|presence_{local}} = \frac{P_{presence_i \, and \, presence_{local}}}{P_{presence_{local}}}$$
$$= \frac{P_{presence_i \, and \, local}}{\sum_{\{j \, : \, Site_j \, in \, local\}} P_{presence_j \, and \, presence_{local}}} = \frac{P_{raw_1}}{\sum_{\{j \, : \, Site_j \, in \, local\}} P_{raw_j}}$$
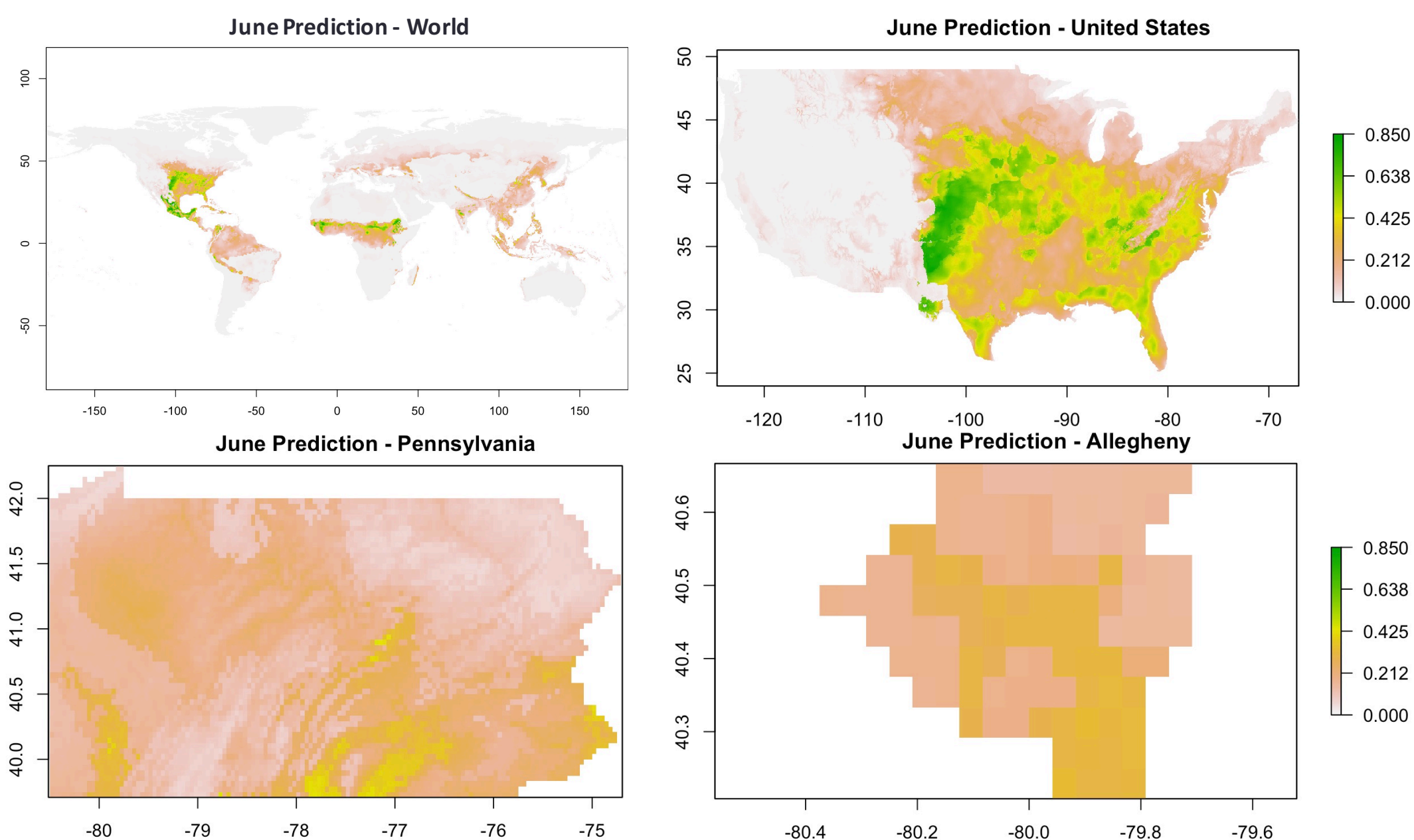

TIME DEPENDENT PREDICTIONS

- Many infectious disease vectors are highly sensitive to environmental conditions such as temperature and precipitation.
- It is therefore important to **study seasonal changes** in the vector populations as environment factors change throughout a year.
- This is achieved by:
  - Subsetting environment layers by month
  - Subsetting species occurrence data by month
  - Fitting separate models for each month
- **Compare likelihood** of a species' presence in different months using mean logistic probability ratios, which reflects the true probability of a species' presence

$$Likelihood \, Ratio_{month_a, month_b} = \frac{\bar{P}_{logistic_a}}{\bar{P}_{logistic_b}}$$
$$= \frac{\frac{1}{n_j} \sum_{\{j \, : \, Site_j \, in \, region\}} P_{logistic_{a,j}}}{\frac{1}{n_j} \sum_{\{j \, : \, Site_j \, in \, region\}} P_{logistic_{b,j}}}$$
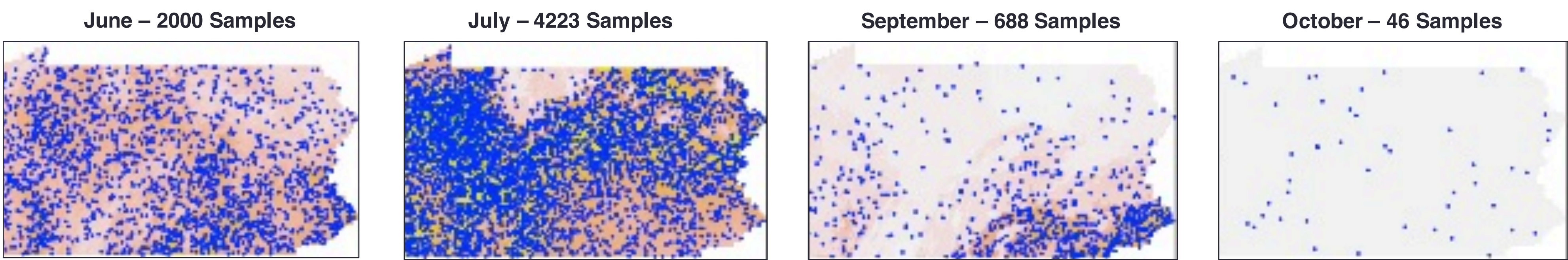
POPULATION SAMPLING

- **Monthly Sample Size:** Given a specific sample size in a reference month, calculate the corresponding sample size for other months based on logistic probabilities, which reflect true probabilities of presence

$$N_{month_a} = \frac{\bar{P}_{logistic_a}}{\bar{P}_{logistic_{ref}}} * N_{ref}$$

- **Sampling Methodology:** Sampling according to transformed regional raw probabilities, which reflect relative likelihood of a presence instance coming from each grid in the region
  - Drawing grid cells with replacement according to transformed raw probabilities
  - Adding to each sampled cell a uniformly distributed within-cell deviation from the cell center
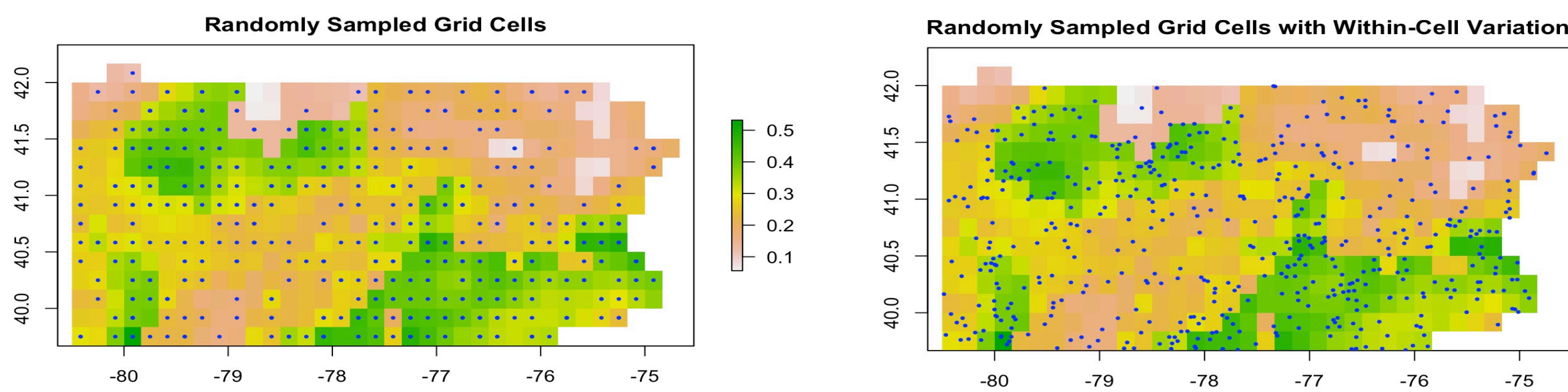

SIMULATION RESULT DEMO

Simulation results for yellow fever mosquitos (Aedes Aegypti) in Pennsylvania, drawing 2000 samples in June:

June – 2000 Samples   July – 4223 Samples   September – 688 Samples   October – 46 Samples

# NEXT STEPS

REAL POPULATION SIZES

- To simulate a synthetic population of infectious disease vectors, it is essential to use the **true population size** for sampling:
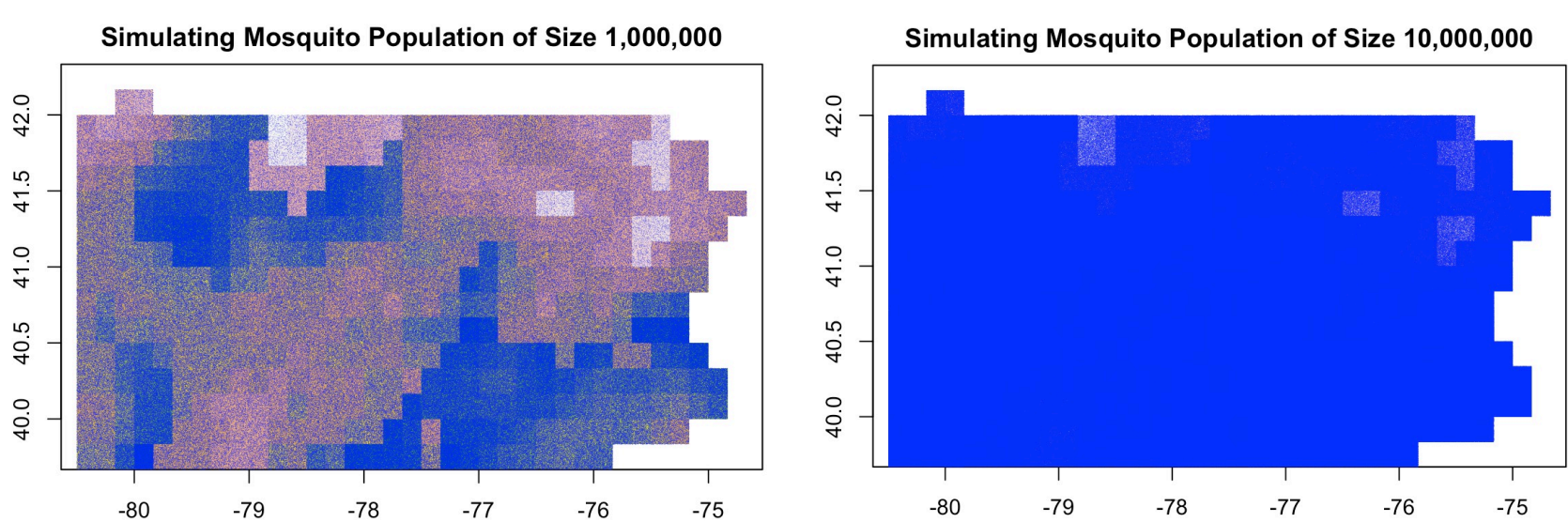
- Currently, there is no publicly available source specifying actual population counts for individual species

CLIMATE CHANGE EFFECTS

- Modifying environmental layers to **study the effects of various climate changes** (such as global warming) on infectious disease vector populations:
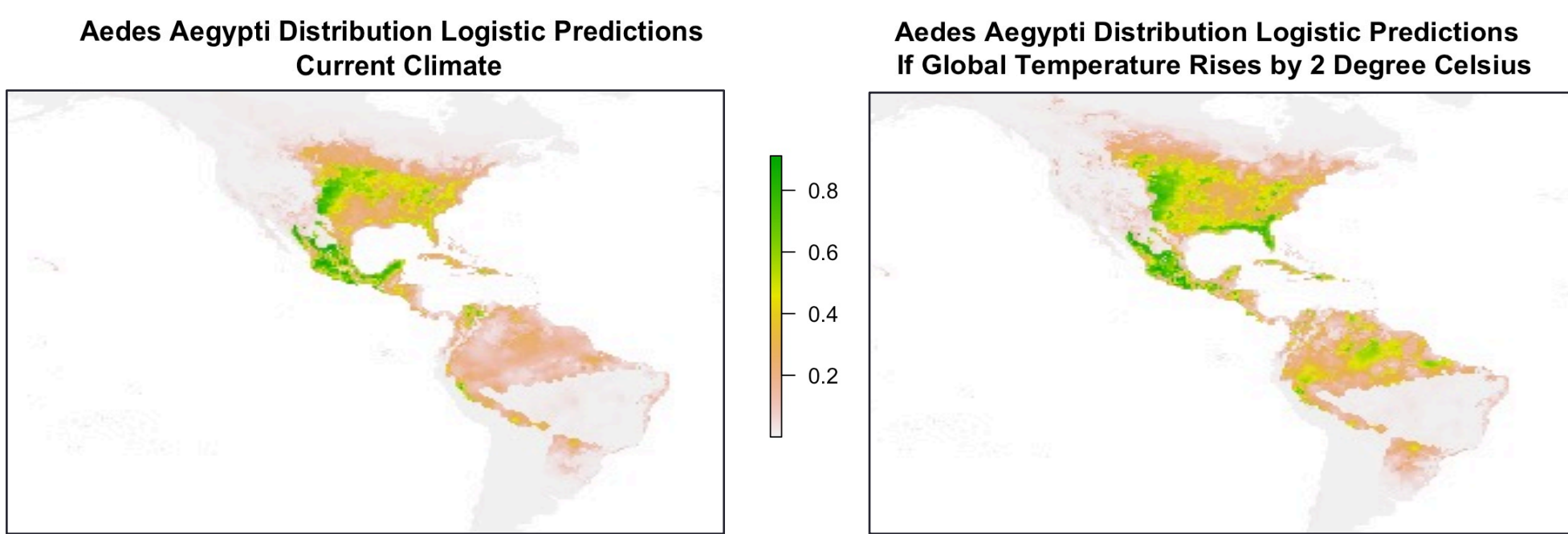
CUSTOMIZATION

- **Flexible input & output format:**
  - Allowing users to supply input data instead of using the default data sources for more updated and relevant results
  - Improving the system's compatibility with various input formats for region name and species name

- **Efficient prediction:**
  - Replacing global predictions with regional predictions, which may save substantial running time especially when users are only interested in modeling disease vector populations for a limited number of regions
  - Incorporating this alternative prediction approach as a speed-up option

CONCLUSIONS

Proposed Infectious Disease Vector Modeling System:

- Generating **synthetic populations of non-human disease vectors** to complement human-based synthetic ecosystems such as SPEW

- Predicting presence probability for **any species, any region** and **any month** of the year, based on species observation data and environmental factors

- Reflecting **seasonal changes** in species population for use in more realistic, time-dependent ABMs

ACKNOWLEDGEMENTS

1. Agent-based model - Wikipedia. (n.d.). Retrieved April 3, 2017, from https://en.wikipedia.org/wiki/Agent-based_model
2. Elith, J., Phillips, S. J., Hastie, T., Dudík, M., Chee, Y. E., & Yates, C. J. (2011). A statistical explanation of MaxEnt for ecologists. Diversity and distributions, 17(1), 43-57.
3. Gallagher, S., Richardson, L., Ventura, S. L., & Eddy, W. F. (2017). SPEW: Synthetic Populations and Ecosystems of the World. arXiv preprint arXiv:1701.02383.