Appendix A: Supplemental Methods

The performance of presence-based and process-based species distribution models under realistic conditions

Tim M. Szewczyk, Marek Petrik, Jenica M. Allen

Contents

This appendix contains supplemental methods pertaining to the virtual species, species distribution models, and scenarios.

---

1 General model structure

1.1 Integral Projection Model overview

To generate fully-known true distributions for the virtual species, we used the general structure of an Integral Projection Model (IPM) to calculate the intrinsic growth rate λ in each cell of the gridded landscape, and adapted the regression-based structure of an IPM into an individual-level, simulation-based cellular automata (CA) model to produce spatiotemporally dynamic abundance distributions.

IPMs (Easterling 2000; Merow et al. 2014, 2017) use the size distribution, z, of individuals at time t, along with a kernel $K(z', z)$, to predict the size distribution, z', at time $t + 1$:

$$n_{t+1}(z') = \int_{\Omega} K(z', z) n_t(z) dz \tag{1}$$

Here, Ω represents the range of possible sizes for the species or population. The kernel $K(z', z)$ is composed of a growth and survival component, $P(z, z')$, representing the fate of individuals from time t to $t + 1$, and a fecundity component, $F(z, z')$, representing new individuals added between time t and $t + 1$. In practice, the integral is approximated using a discretized transition matrix, and the intrinsic growth rate λ is calculated as the first eigenvalue of the transition matrix.

The P and F kernels are decomposed further into more mechanistic conditional probabilities and parameters, many of which are functions of the size distribution z. For example:

$$\begin{aligned} K(z', z) &= P(z', z) + F(z', z) \\ &= s(z)g(z'|z) + p_{flower}(z)f_{seeds}(z)p_{estab}f_{rcrSize}(z') \end{aligned} \tag{2}$$

where $s(z)$ is the survival probability of individuals based on size, $g(z'|z)$ is the probability density of size z' for an individual of size z, $p_{flower}(z)$ is the probability that an individual of size z produces flowers, $f_{seeds}(z)$

is the expected number of seeds produced by an individual of size z given that they flower, p_{estab} is the probability that a seed germinates and establishes as a new recruit, and $f_{rcrSize}(z')$ is the expected size distribution of new recruits at time $t+1$. On a gridded landscape, the population in each cell could be modelled independently using the above structure, with environmental effects incorporated by allowing the environment to influence parameter values (e.g., f_{seeds}).

1.2 Adding a seed bank

This basic IPM structure is very flexible, allowing for discrete life stages, reproduction-dependent mortality, and environmental covariates. To incorporate a seed bank where seeds that do not germinate between time t and $t+1$ may survive to $t+2$ or beyond, the fecundity kernel is altered, with seed bank B, such that:

$$n_{t+1}(z') = B_t s_{rcrB} p_{estab} f_{rcrSize}(z') + \int_\Omega [P(z',z) + F(z',z)] n_t(z) dz \tag{3}$$

$$F(z',z) = p_{flower}(z) f_{seeds}(z) s_{rcrDirect} p_{estab} f_{rcrSize}(z') \tag{4}$$

$$B_{t+1} = B_t s_{survB}(1 - s_{rcrB}) + p_{flower}(z) f_{seeds}(z)(1 - s_{rcrDirect}) s_{survB} n_t(z) dz \tag{5}$$

where B_t is the number of seeds in the seed bank at time t, s_{rcrB} is the probability a seed recruits from the seed bank, s_{survB} is the probability a seed survives in the seed bank from time t to $t+1$, and $s_{rcrDirect}$ is the probability a seed produced in year t germinates between year t and $t+1$. Two notes about the structure and definitions: 1) a seed added to the seed bank must fail to recruit in time t and must also survive from t to $t+1$, and 2) the probability of recruiting is best interpreted as the probability of germinating and is therefore separate from the probability of establishing. Both of these could be defined differently to combine each set of processes, though keeping them separate allows for seeds to perish.

1.3 Adding dispersal

The above equations assume isolated populations in each cell. However, for a typical plant species, dispersal occurs when seeds move from the cell where they are produced to a different cell. In an IPM with a seed bank, this will affect the fecundity kernel $F(z',z)$ and the seed bank B. Specifically, the number of seeds in the seed bank in cell i at time $t+1$ will be the number of seeds surviving in the seed bank B_i from time t to $t+1$, plus the number of seeds produced in cell i at time t that remain in cell i and do not recruit directly, plus the number of seeds entering cell i from cells $j = 1, ..., J$ as immigrants in time t that then fail to recruit directly:

$$\begin{aligned} B_{i,t+1} = \; & B_{i,t} s_{survSB}(1 - s_{rcrSB}) \; + \\ & \int_\Omega p_{flower,i}(z) f_{seeds,i}(z)(1 - p_{emig})(1 - s_{rcrDirect}) s_{survB} n_{i,t}(z) dz \; + \\ & \sum_{j=1}^{J} \int_\Omega [p_{flower,j}(z) f_{seeds,j}(z) p_{emig} n_{t,j}(z) dz] p_{SDD,ji}(1 - s_{rcrDirect}) s_{survB} \end{aligned} \tag{6}$$

$$\begin{aligned} n_{i,t+1}(z') = \; & B_{i,t} s_{rcrSB} p_{estab_i} f_{rcrSize}(z') \; + \\ & \int_\Omega [s_i(z) g_i(z'|z) + p_{flower_i}(z) f_{seeds_i}(z)(1 - p_{emig}) s_{rcrDirect} p_{estab_i} f_{rcrSize}(z')] n_{i,t}(z) dz \; + \\ & \sum_{j=1}^{J} \int_\Omega [p_{flower,j}(z) f_{seeds,j}(z) p_{emig} n_{t,j}(z) dz] p_{SDD,ji} s_{rcrDirect} p_{estab,i} f_{rcrSize}(z') \end{aligned} \tag{7}$$

for each cell i which is a target cell of each cell j of J cells, where the integral describes the seed production in each cell j and $p_{SDD,ji}$ is the probability that a seed dispersed from j lands in i. Note that, as above, seeds added to the seed bank must survive overwinter as well as fail to recruit directly.

---

2 Regression equations for the virtual species

We used the above IPM structure, including a seed bank and dispersal, as the basis for the virtual species. Thus, the population in each cell i has size distribution z, where z' is the size distribution the next year, and $\boldsymbol{z_i}$ is a matrix with columns of 1 (intercept) and z. For the IPMs, we used the following regression equations to populate discretized transition matrices for each cell to then calculate the deterministic intrinsic growth rate λ. For the $\mathrm{CA_i}$ models, we used the same regression equations to generate expected values for each individual, and then drew simulated outcomes from the specified distribtutions.

2.1 Survival

Annual survival, $\boldsymbol{s_i}$, was modelled for each individual in cell i as a binary outcome (0: mortality; 1: survival) following a Bernoulli distribution with probability $\boldsymbol{\psi_{si}}$ such that:

$$\boldsymbol{s_i} \sim Bern(\boldsymbol{\psi_{si}}) \tag{8}$$

$$logit(\boldsymbol{\psi_{si}}) = \boldsymbol{z_i}\beta_s + \boldsymbol{X_i}\theta_s \tag{9}$$

where β_s is a vector of covariates for size, $\boldsymbol{X_i}$ is a set of cell-level environmental covariates, and θ_s is a vector of responses to the environmental covariates.

2.2 Growth

The size distribution, $\boldsymbol{z_i'}$ of individuals in cell i for time $t+1$ was distributed normally about the vector of expected sizes, $\boldsymbol{\mu_{gi}}$ with standard deviation σ_g, such that

$$\boldsymbol{z_i'} \sim Norm(\boldsymbol{\mu_{gi}}, \sigma_g) \tag{10}$$

$$\boldsymbol{\mu_{gi}} = \boldsymbol{z_i}\beta_g + \boldsymbol{X_i}\theta_g \tag{11}$$

where β_g is a vector of covariates for size, $\boldsymbol{X_i}$ is a set of cell-level environmental covariates, and θ_g is a vector of responses to the environmental covariates.

2.3 Flowering

Individual flowering, $\boldsymbol{l_i}$, was modelled for each individual in cell i as a binary outcome (0: no flowers; 1: flowers) following a Bernoulli distribution with probability $\boldsymbol{\psi_{li}}$ such that:

$$\boldsymbol{l_i} \sim Bern(\boldsymbol{\psi_{li}}) \tag{12}$$

$$logit(\boldsymbol{\psi_{li}}) = \boldsymbol{z_i}\beta_l + \boldsymbol{X_i}\theta_l \tag{13}$$

where β_l is a vector of covariates for size, $\boldsymbol{X_i}$ is a set of cell-level environmental covariates, and θ_l is a vector of responses to the environmental covariates.

2.4 Seeds

The number of seeds produced, $\boldsymbol{d_i}$ by each flowering individual in cell i for time t was Poisson distributed about the vector of expected seed counts, $\boldsymbol{\mu_{di}}$, such that

$$\boldsymbol{d_i} \sim Poisson(\boldsymbol{\mu_{di}}) \tag{14}$$

$$log(\boldsymbol{\mu_{di}}) = \boldsymbol{z_i}\beta_d + \boldsymbol{X_i}\theta_d \tag{15}$$

where β_d is a vector of covariates for size, $\boldsymbol{X_i}$ is a set of cell-level environmental covariates, and θ_d is a vector of responses to the environmental covariates. In each cell i, the seeds produced stay in the cell, emigrate through short distance dispersal, or perish. Seeds that survive may either enter the seed bank or recruit directly.

2.5 Dispersal

The total number of immigrant seeds arriving in a cell, D_i is calculated as the sum of the seeds from nearby cells that disperse from cell j to cell i, such that:

$$D_i = \sum_{j=1}^{J} \sum_{n=1}^{N_j} \boldsymbol{d_j} p_{emig} p_{ji} \tag{16}$$

where J is the number of cells dispersing into i, N_j is the number of individuals in cell j, d_j is the vector of seed numbers produced by individuals in cell j, p_{emig} is the probability that seeds produced in cell j emigrate, and p_{ji} is the probability that a seed emigrating from cell j is dispersed to cell i.

2.6 Recruits

The number of recruits, $n_{rcr,i}$ in cell i in time t, may originate either from the seed bank, from seeds produced in cell i in year t, or from immigrant seeds arriving in year t, such that:

$$n_{rcr,i} = B_i p_{rcrB} p_{est} + \sum_{n=1}^{N_i} \boldsymbol{d_i} (1 - p_{emig}) p_{rcrDirect} p_{est} + D_i p_{rcrDirect} p_{est} \tag{17}$$

$$\boldsymbol{z'_{rcr,i}} \sim Norm(\mu_{rcr.z}, \sigma_{rcr.z}) \tag{18}$$

where B_i is the number of seeds in the seed bank, p_{rcrB} is the probability that a seed germinates from the seed bank, p_{est} is the probability that a new seedling establishes, $p_rcrDirect$ is the probability that a seed germinates in the year it was produced, $z'_{rcr,i}$ is the size distribution of new recruits in year $t+1$, $\mu_{rcr.z}$ is the mean recruit size, and $\sigma_{rcr.z}$ is the standard deviation of recruit size.

2.7 Seed Bank

Finally, the seed bank for year $t+1$ is calculated as the sum of the seeds remaining in the seed bank, the seeds produced in cell i and entering into the seed bank, and the immigrant seeds entering into the seed bank, such that:

$$B'_i = B_i (1 - p_{rcrB}) s_B + \sum_{n=1}^{N_i} \boldsymbol{d_i} (1 - p_{emig}) (1 - p_{rcrDirect}) s_B + D_i (1 - p_{rcrDirect}) s_B \tag{19}$$

where s_B is the probability that a seed survives in the seed bank to the next year.

---

3 Species Distribution Model details

This section of the appendix contains additional information regarding the structure of the species distribution models. The full R code is available from https://github.com/Sz-Tim/sdmMethodComp.

3.1 IPM

The Integral Projection Models for each species followed the structure described in the previous section. For each dataset (100 datasets x 2 species x 4 data scenarios), we fit regressions for each vital rate regression using the R package *MuMIn* to compare all combinations of climatic variables. The full models were:

```
# pr(survival)
glm(surv ~ size + temp + temp_sq + precip + precip_sq, family="binomial")
# growth | survival
lm(size_next ~ size + temp + temp_sq + precip + precip_sq)
# pr(flowering | survival)
glm(flower ~ size + temp + temp_sq + precip + precip_sq, family="binomial")
# number of seeds | flowering
glm(n_seed ~ size + temp + temp_sq + precip + precip_sq, family="poisson")
# pr(germination)
glm(germ ~ temp + temp_sq + precip + precip_sq, family="binomial")
```

The size distribution for recruits was taken as the sample mean and standard deviation of recruits in each cell. Parameters relating to seed bank dynamics and dispersal were assumed to be known, and the effects of this assumption were evaluated in the modelling mis-specification scenarios.

3.2 CA_i

The individual-level cellular automaton model (CA_i) simulates individuals in each cell of the landscape using the structure of the IPM. Effectively, the fitted regressions are used to generate expected values for each individual, accounting for size and environment, and random values are then drawn from the corresponding distributions to simulate outcomes. For example, to assign survival, a value is drawn from a binomial distribution with the probability calculated from the fitted survival GLM. Thus, we simulated local processes within each cell. Regional processes were represented as the dispersal of seeds both within short distance dispersal neighborhoods and to a pre-determined number of random cells in the landscape. The CA_i model is thus a simulation-based adaptation of the IPM. However, this requires imposing density dependence to both maintain realism and to work within the confines of modern computing capabilities. We implemented density dependent seedling establishment such that $p_{est,i,t} = min(p_{est,i}, N_{rcr}/(p_{germ}(D_i + B_i + d_i(1 - p_{emig}))))$, following the naming convention described above, effectively placing a cap on the number of new recruits (Ellner & Rees 2006).

3.3 CA_p

The population-level cellular automaton model (CA_p) simulates stage-structured populations in each cell of the landscape using population-level values for demographic rates. Thus, rather than simulating, for example, the survival of each individual, the CA_p model simulates the total expected number of surviving individuals in each cell. We follow the model structure and sequence from a previously published model focused on the invasive shrub glossy buckthorn (*Frangula alnus*) in the northeastern United States (Szewczyk et al. 2019). This includes the same life cycle processes as in the IPM and CA_i: survival, growth (=aging), flowering, seed production, short and long distance dispersal, seed bank dynamics, and germination. However, individuals were not tracked or predicted, but rather averages or totals within age categories (shrub: 4 age categories; biennial: 2 age categories). We implemented density dependence as a ceiling-type carrying capacity, where the carrying capacity K was predicted for each cell. In the samples used to parameterize the regressions (consisting of 3 years of data), cells were assumed to be approximately at carrying capacity if the calculated

$lambda_t = N_t/N_{t-1}$ changed less than 5% between consecutive years. Thus, we parameterized the following full models, including year as a random effect:

```
# carrying capacity
glmer(K ~ temp + temp_sq + precip + precip_sq + (1|yr), family="poisson")
# adult survival (shrub only)
glmer(surv_adult ~ temp + temp_sq + precip + precip_sq + (1|yr), family="binomial")
# juvenile survival
glmer(surv_juv ~ temp + temp_sq + precip + precip_sq + (1|yr), family="binomial")
# pr(flowering | survival)
glmer(flower ~ size + temp + temp_sq + precip + precip_sq + (1|yr), family="binomial")
# number of seeds | flowering
glmer(n_seed ~ size + temp + temp_sq + precip + precip_sq + (1|yr), family="poisson")
# pr(germination)
glm(germ ~ temp + temp_sq + precip + precip_sq, family="binomial")
```

As in the IPM and CA_i, dispersal included long distance and short distance dispersal of seeds. Identical assumptions and implementations were used for the CA_p model as described above.

---

4 Scenario details

This section of the appendix contains additional information regarding the data and modelling scenarios. The full R code is available from https://github.com/Sz-Tim/sdmMethodComp. Include the table showing the exact implementation of each scenario.

4.1 Data scenarios

4.1.1 Ideal

In the *ideal* scenario, each data set consisted of cells sampled from the equilibrium distribution at the end of the virtual species simulation (tmax = 300). Cells with N > 5 were selected with uniform probability. For MaxEnt, 300 occupied cells were drawn for each data set. For the process-based SDMs, 25 cells were drawn. Within each of the 25 cells, up to 100 (CA_i, IPM) or 1000 (CA_p) individuals were randomly selected to be used in the vital rate regressions, with stratification among small, medium, and large individuals (CA_i, IPM) or juveniles and adults (CA_p). For cells with fewer individuals than the maximum, all individuals were included regardless of size. The final three years of the simulation were used to parameterize the CA_p model, while only the final two years were used for the CA_i model and IPM. The data available included abundance, survival, size, flowering status, fruit production, and germination rates. For estimating carrying capacity in the CA_p model, we assumed cells where N changed less than 5% between years were approximately at carrying capacity, and used the average abundance across years for those cells.

4.1.2 Non-equilibrium

The sampling regime under the *non-equilibrium* scenario was identical to that of the *ideal* scenario, except that samples were drawn from the distribution 100 years after introduction rather than the equilibrium distribution of 300 years.

4.1.3 Geographic bias

The *geographic bias* sampling scenario represents the bias in sampling locations common in ecological studies, where data are more likely to be acquired in locations that are easier to reach. To establish realistic patterns of data collection, we used previously published data on invasive plant occurrences in the United States (Allen & Bradley 2016) along with human population density (Manson et al. 2017) and the primary road network (US Census Bureau 2017). In each cell of the landscape, we binarized invasive plant observation (0: no observations, 1: at least 1 observation), the total human population density, and the total length of primary roads. We fit a logistic regression including quadratic terms for each covariate. Finally, we used the predicted probability of observation from the regression as the probability that each cell was sampled from the virtual species distributions.

4.1.4 Measurement error

The *measurement error* sampling scenario added noise to the observations used to fit each SDM. Because the type of data required differs drastically between SDM methods, we adjusted the mechanism for adding error accordingly. In each case, datasets were first generated as described in the *ideal* scenario. To mimic errors in geolocation or identification for MaxEnt, we randomly selected 3% of the cells in each dataset, replacing them with cells chosen randomly from the landscape regardless of occupancy status. To mimic error in field measurements for the population-level datasets (i.e., CA_p), we added noise to the observations of abundance ($N_{obs} = Norm(N_{true}, N_{true}{*}0.02)$) and seed production ($seed_{obs} = Norm(seed_{true}, seed_{true}{*}0.05)$). To mimic error in field measurements for the individual-level datasets (i.e., CA_i, IPM), we added noise to the observations of growth ($g_{obs} = Norm(g_{true}, 0.1)$) and seed production ($seed_{obs} = Norm(seed_{true}, seed_{true} * 0.05)$). Thus, counting error for seeds and individuals increased proportionally to the true values, while measurement error for size was constant.

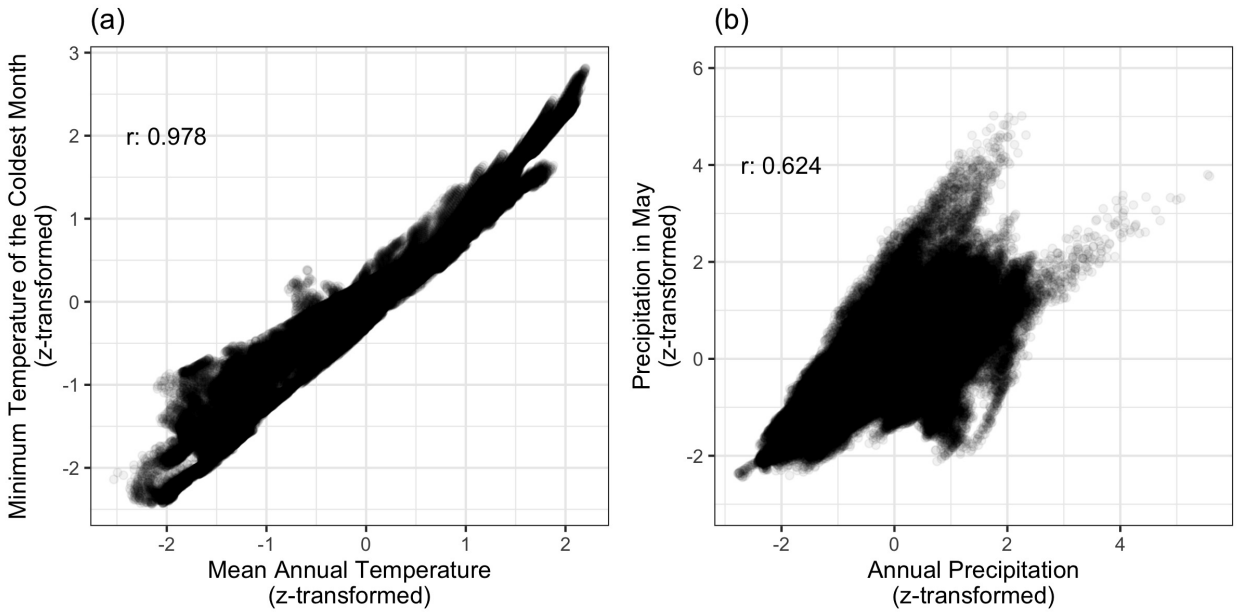

Figure A.1: Scatter plots of z-transformed environmental covariates. (a) Mean Annual Temperature and Minimum Temperature of the Coldest Month, and (b) Annual Precipitation and Precipitation in May. The more general variables are highly correlated with the more specific variables across the eastern United States at a 5x5km resolution. The true demographic rates were related to the more specific variables, while the more general were used in the *incorrect covariates* modelling scenario.

4.2 Modelling scenarios

All modelling scenarios were performed using the *ideal* data sets. Rather than addressing common issues with the data used to inform the SDMs, the focus was on the structure and implementation of the SDMs themselves.

4.2.1 Incorrect covariates

In the *incorrect covariates* modelling scenario, we fit all SDMs with correlated but more general climatic variables than were used in the generative process for the virtual species. We used Mean Annual Temperature in place of the Minimum Temperature of the Coldest Month (r: 0.978; Fig. A.1a), and Annual Precipitation in place of the Precipitation in May (r: 0.624; Fig. A.1b). Though correlated, the geographic distribution of the difference is highly structured for both temperature and precipitation (Fig. A.2).

4.2.2 No seed bank

This scenario applied only to the process-based SDMs. Seed bank dynamics are difficult to quantify, and are poorly known in general relative to other aspects of species' life cycles and ecology. A simple solution to this problem is to assume negligible effects of the seed bank on species' distributions or that any important effects are compensated for by being implicitly included in estimates of related processes such as direct germination. In this modelling scenario, we assume in each process-based SDM that seed bank mortality is complete such that all seedlings in year t originate from seeds produced in year $t-1$.

4.2.3 Under dispersal

This scenario applied only to the process-based SDMs. Dispersal occurred via both short distance and long distance processes, and both were reduced in the *under dispersal* scenario. Because two virtual species had different dispersal strategies, where the shrub had more effective short distance dispersal while the

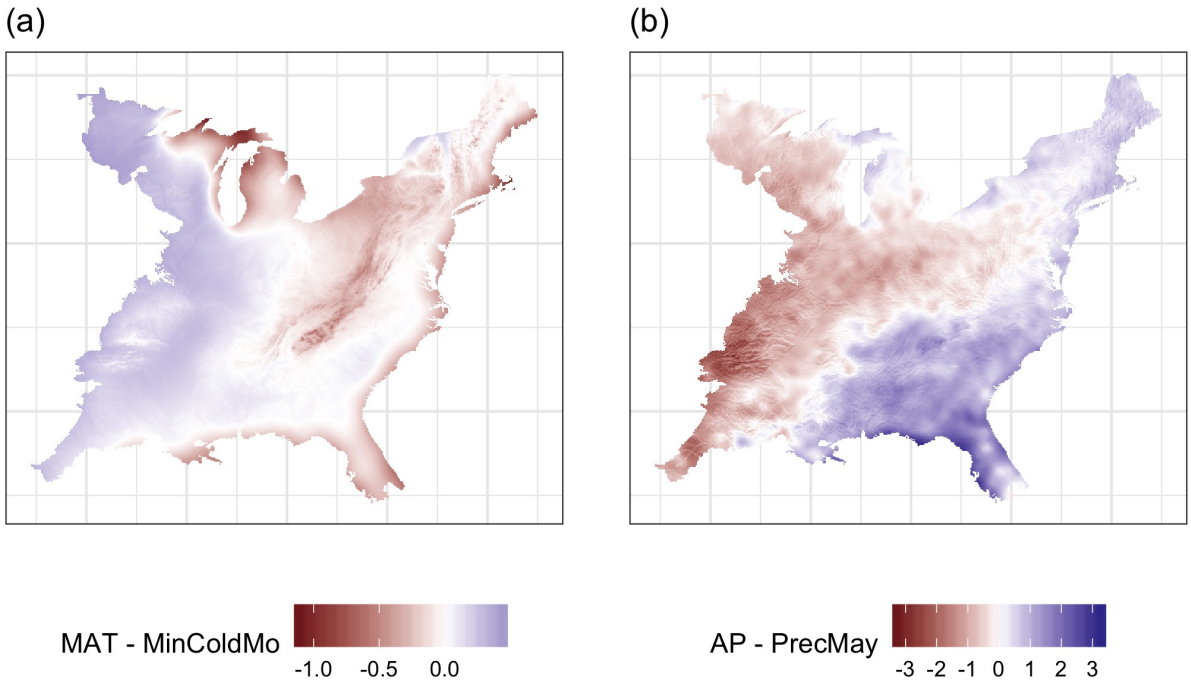

Figure A.2: Map of the difference between z-transformed environmental covariates. (a) Mean Annual Temperature and Minimum Temperature of the Coldest Month and (b) Annual Precipitation and Precipitation in May. Though highly correlated, the residuals show strong spatial patterning in each case.

biennial relied more heavily on long distance dispersal, we modified dispersal proportionally. For long distance dispersal, the number of annual events was decreased to 20% of the true value. For short distance dispersal, the rate of the exponential kernel was decreased by 50%. Consequently, the total number of seeds produced was unaffected, but chance dispersal events to new regions were less likely, and seeds were more likely to be dispersed to nearby cells within the short distance dispersal neighborhood.

4.2.4 Over dispersal

This scenario applied only to the process-based SDMs. For the *over dispersal* scenario, the model misspecification was reversed compared to the *under dispersal* scenario. Thus, for long distance dispersal, the number of annual events was increased to 5 times the true value. For short distance dispersal, the rate of the exponential kernel was increased to 2 times the true value. Consequently, the total number of seeds produced was unaffected, but chance dispersal events to new regions were more likely, and cells toward the periphery of the short distance dispersal neighborhood were more likely to receive seeds.

---

5 References

Allen, J. M., & Bradley, B. A. (2016). Out of the weeds? Reduced plant invasion risk with climate change in the continental United States. Biological Conservation, 203(November), 306–312. https://doi.org/10.1016/j.biocon.2016.09.015

Easterling, M. R., Ellner, S. P., & Dixon, P. M. (2000). Size-specific sensitivity: Applying a new structured population model. Ecology, 81(3), 694–708. https://doi.org/10.1890/0012-9658(2000)081%5B0694:SSSAAN%5D2.0.CO;2

Ellner, S. P., & Rees, M. (2006). Integral Projection Models for Species with Complex Demography. The American Naturalist, 167(3), 410–428. https://doi.org/10.1086/499438

Manson, S., Schroeder, J., Van Riper, D., & Ruggles, S. (2017). IPUMS National Historical Geographic Information System: Version 12.0 [Database]. https://doi.org/http://doi.org/10.18128/D050.V12.0

Merow, C., Dahlgren, J. P., Metcalf, C. J. E., Childs, D. Z., Evans, M. E. K., Jongejans, E., ... Mcmahon, S. M. (2014). Advancing population ecology with integral projection models: A practical guide. Methods in Ecology and Evolution, 5(2), 99–110. https://doi.org/10.1111/2041-210X.12146

Merow, C., Bois, S. T., Allen, J. M., Xie, Y., & Silander, J. A. (2017). Climate change both facilitates and inhibits invasive plant ranges in New England. Proceedings of the National Academy of Sciences, 114(16), E3276–E3284. https://doi.org/10.1073/pnas.1609633114

Szewczyk, T. M., Lee, T., Ducey, M. J., Aiello-Lammens, M. E., Bibaud, H., & Allen, J. M. (2019). Local management in a regional context : Simulations with process-based species distribution models. Ecological Modelling, 413(August), 108827. https://doi.org/10.1016/j.ecolmodel.2019.108827

U.S. Census Bureau. (2017). 2017 TIGER/Line Shapefiles Technical Documentation. Retrieved from https://www.census.gov/geo/maps-data/data/pdfs/tiger/tgrshp2012/TGRSHP2012_TechDoc.pdf