The performance of presence-based and process-based species distribution models

Tim M. Szewczyk[a,b], Marek Petrik[b], Jenica M. Allen[a]

[a]*Department of Natural Resources and the Environment, University of New Hampshire*
[b]*Department of Computer Science, University of New Hampshire*

Abstract

Abstract here
Keywords: keywords

1. Introduction

1. General introduction about species distribution models and the historical prevalence of occurrence-based SDMs.

2. The rise of process-based SDMs (or at least the rise in advocacy for process-based SDMs), including some specific examples of their use, since that was the main question raised at ISEM.

3. Expected benefits of process-based SDMs

4. Potential downsides of process-based SDMs

We evaluate the performance of an occurrence-based SDM and three process-based SDMs for two virtual species. We investigate the ability of each SDM to predict the true distributions under ideal conditions, as well as under various realistic data and modelling scenarios. Specifically, we ask: 1) Do process-based models outperform MaxEnt overall? 2) are process-based or occurrence-based SDMs more susceptible to particular data deficiencies? And 3) do common modelling errors negate any improved performance observed in the process-based SDMs?

2. Methods

We evaluated the performance of four species distribution models (SDMs) in predicting the ranges of two simulated species under a variety of data quality and modelling scenarios (Fig. 1), including one occurrence-based method (MaxEnt) and three process-based methods (Integral Projection Model: IPM; individual-level cellular automata: CA_i; and population-level cellular

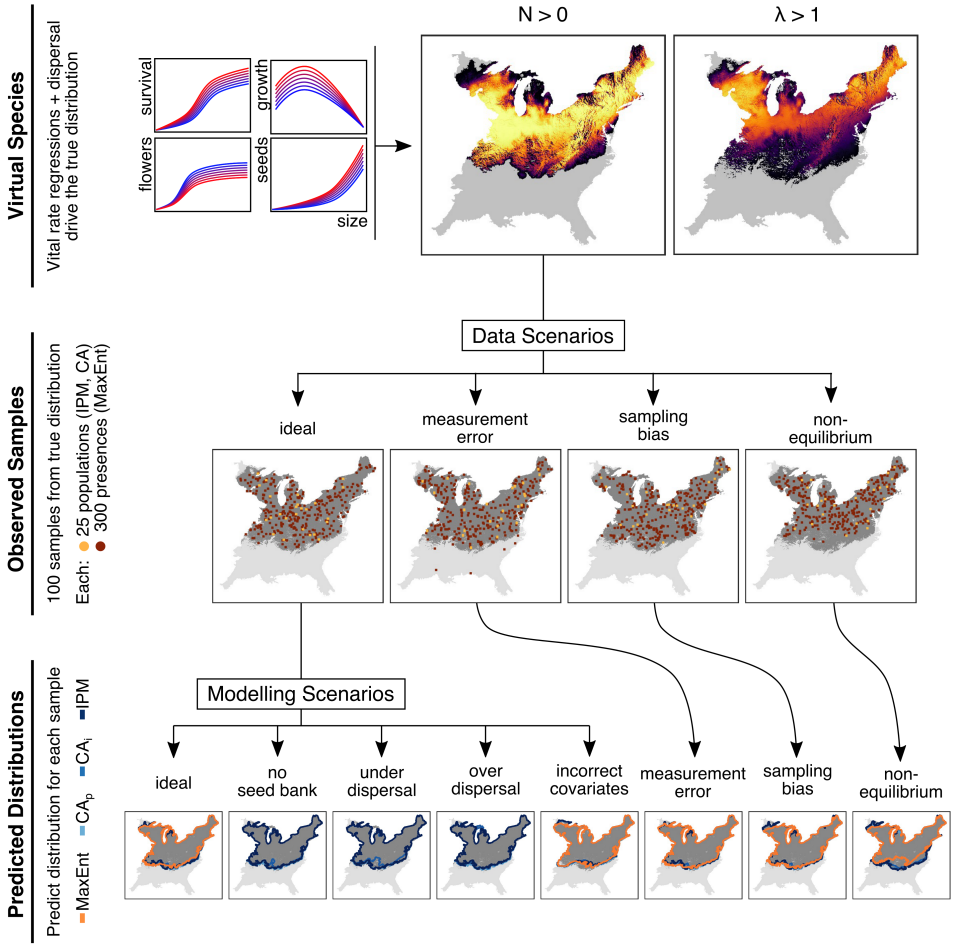

Figure 1: Simulation, sampling, and prediction process. Distributions were simulated for virtual species based on environmental effects on vital rates and demographic processes.

automata: CA_p). IPMs predict a deterministic intrinsic growth rate λ in each cell of the landscape based on regressions of vital rates with environmental variables and individual sizes. In contrast, CA models are simulation-based and explicitly model spatio-temporal dynamics among cells. The CA_i model simulates the growth, survival, and fecundity of individuals, while the CA_p model uses population averages to summarize similar processes. See additional details for the process-based models in Appendix 1.

2.1. Virtual Species Simulation

We simulated two virtual species in the eastern United States, based approximately on Japanese barberry (*Berberis thunbergii*), an invasive shrub with bird-dispersed seeds, and garlic mustard (*Alliaria petiolata*), an invasive biennial with seeds dispersed by non-volant animals and water. Both species have previously been the focus of process-based SDMs (CITE). To simulate each species, we defined relationships between environmental variables, individual size, and vital rates (Fig. 1, Appendix 1) on a gridded landscape spanning the Eastern Temperate Forest and Northern Forest ecozones within the United States (~5x5km resolution: 115,105 cells, CITE). Vital rates included annual individual growth, annual survival, flowering probability, seed production, and germination probability. For environmental covariates, we used the minimum temperature of the coldest month and the precipitation in May, following previous work (CITE PNAS, Chelsa). Based on these regressions, we calculated the intrinsic growth rate λ in each cell following the methods used for IPMs, where each cell was treated as a population. Cells with $\lambda > 1$ were considered to be capable of containing persistent populations, constituting the first definition of the true distribution for each species.

To simulate populations, we initialized 10 random cells with 10 individuals, with individual sizes drawn from a uniform distribution constrained by the allowable range for each species. Then, we calculated the expected survival probability, growth rate, flowering probability, and seed production for each individual according to the vital rate regressions (Appendix 1). Then, we drew random values from the appropriate distributions for each vital rate for each individual, implemented short- and long-distance dispersal of seeds according to the dispersal mode of each species, and generated new individuals in each cell from germinating seeds. Populations in each cell were subject to density dependence through reduced seedling establishment (CITE) after the abundance

exceeded a predetermined threshold. The populations were simulated through 300 years, at which point the distribution appeared stable, indicating that equilibrium had been reached. The cells with $N > 0$ in the final year thus constituted the second definition of the true distribution for each species.

2.2. Sampling and Modelling Scenarios

We evaluated four sampling scenarios: *ideal* (year 300; uniform sampling probability among occupied cells), *non-equilibrium* (year 100; uniform sampling probability among occupied cells), *geographic bias* (year 300; sampling probability $\propto$ human and road density in occupied cells), and *measurement error* (year 300; MaxEnt: 3% of observations drawn from full landscape regardless of occupancy; process-based: error added to growth, seed production, and abundance observations). For each species, we generated 100 sets of sampled cells for each scenario. For MaxEnt, each sample consisted of 300 cells (CITE), while for the process-based SDMs, each sample consisted of 25 cells (CITE). For process-based models, a maximum of 1000 (CA_p) or 100 (CA_i, IPM) individuals were sampled from within each selected grid cell to reflect real-world logistical limitations (Appendix 1, https://github.com/Sz-Tim/sdmMethodComp).

In addition, we assessed the effect of four scenarios of modeling misspecification: *incorrect covariates*, *no seed bank* (process-based only), *over-estimated dispersal* (process-based only), *under-estimated dispersal* (process-based only). Each of these misspecifications was implemented using the *ideal* datasets. For *incorrect covariates*, the SDMs were fit with correlated but more general climate variables (mean annual temperature, annual precipitation) rather than the more specific covariates used in the generative process and for all other scenarios (minimum temperature of the coldest month, precipitation in May).

For each scenario (Fig 1), the SDMs were parameterized using the corresponding sampled datasets. To select environmental covariates for each vital rate regression for each process-based SDMs, we compared all combinations of the climatic variables and their squares with the R package *MuMIn*. For the CA_p and CA_i models, simulations were run for 300 years to generate predicted distributions. Using each definition of the true simulated ranges (i.e., $\lambda > 1$ and $N > 0$), we calculated the True Skill Statistic (TSS) for each of the 100 predicted distributions per scenario per SDM. We calculated TSS as $sensitivity + specificity - 1$, where $sensitivity$ is the proportion of

SDM	Best	Mean	Median	Worst
True range: $\lambda > 1$				
CA_p	2	2.9	2.5	4
CA_i	2	2.9	3	4
IPM	1	1.5	1	4
MaxEnt	1	2.7	2.5	4
True range: $N > 0$				
CA_p	1	1.9	2	3
CA_i	1	1.6	1	4
IPM	3	3.6	4	4
MaxEnt	1	2.9	3	4

Table 1: Summary of ranks across species and all scenarios applicable to all four SDMs. For each scenario and species, SDMs were ranked based on the median TSS.

true presences that are predicted correctly, and *specificity* is the proportion of true absences that are predicted correctly, such that TSS ranges from −1 (all cells predicted incorrectly) to 1 (all cells predicted correctly).

3. Results

All models performed well in recovering the true λ-based ($\lambda > 1$) and N-based distributions ($N > 0$) for both species across all scenarios (all TSS medians > 0.77), though no single model performed best universally (Fig. 2). Across scenarios applicable to all four SDMs (Table 1), the IPM best predicted the λ-based distributions (mean rank: 1.5) and the CA models best predicted the N-based distributions (mean rank: CA_i=1.6, CA_p=1.9). MaxEnt was intermediate for both range boundaries on average (mean rank: λ-based=2.7, N-based=2.9)

For MaxEnt, TSS decreased when fit using samples from non-equilibrium populations, driven by smaller predicted ranges as indicated by a decrease in sensitivity (Fig. SXX:Sensitivity-Ideal) and a smaller increase in specificity (Fig. SXX:Specificity-Ideal). In contrast, non-equilibrium

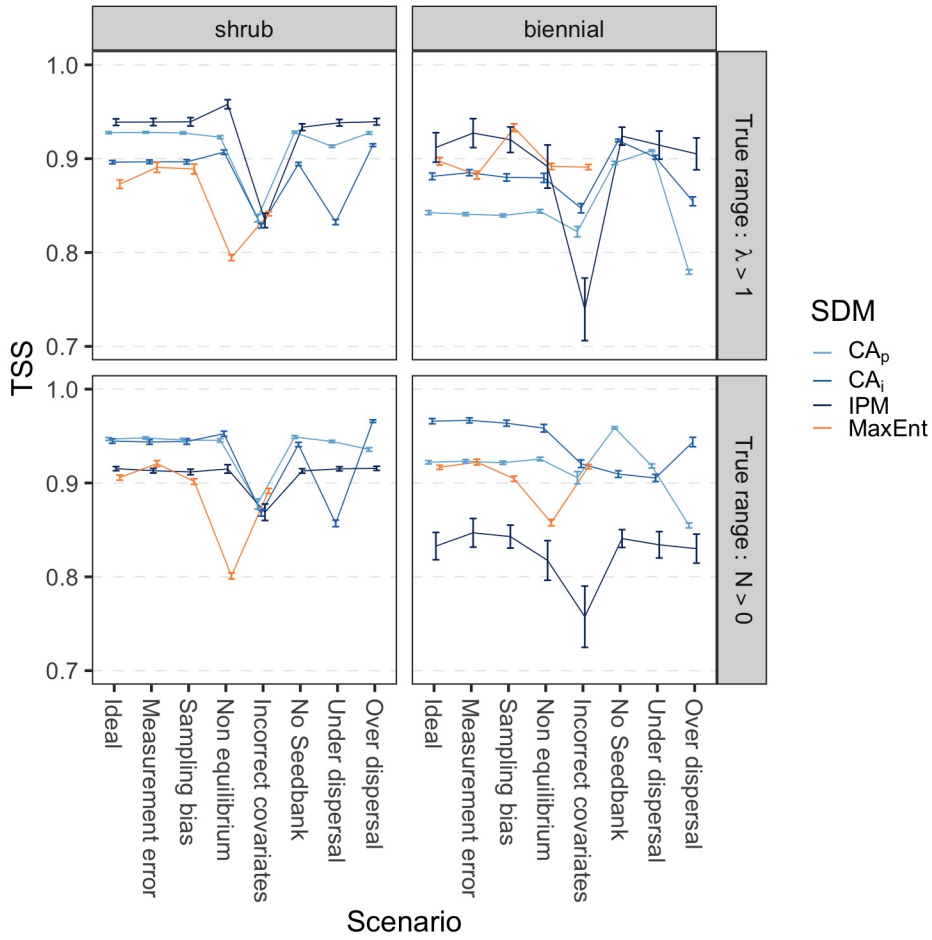

Figure 2: True skill statistic (TSS) mean and 95% confidence intervals across 100 sampled datasets for each SDM and scenario, compared to true distributions defined by $\lambda > 1$ and $N > 0$. Scenarios include: no sampling or modeling issues (ideal), sampling issues (measurement error, sampling bias, non-equilibrium), and modeling issues (incorrect covariates, no seed bank, under dispersal, over dispersal).

samples had minimal effect on TSS for the CA_p model, with a modest improvement for the long-lived shrub in the IPM and CA_i models driven by increases in sensitivity (Fig. SXX:Sensitivity-Ideal). Measurement error and sampling bias had negligible effects on the process-based models, and small to moderate effects on the performance of MaxEnt (Fig. 2, Fig. SXX:Sensitivity-Ideal, Fig. SXX:Specificity-Ideal).

Modeling with incorrect covariates was universally problematic for model performance, though it had a greater impact on TSS in the shrub than the biennial species (Fig. 3). Incorrect covariates did not lead to systematic over-prediction or under-prediction, but rather a decrease in both sensitivity and specificity, particularly for the process-based models (Fig. SXX:Sensitivity-Ideal, Fig.

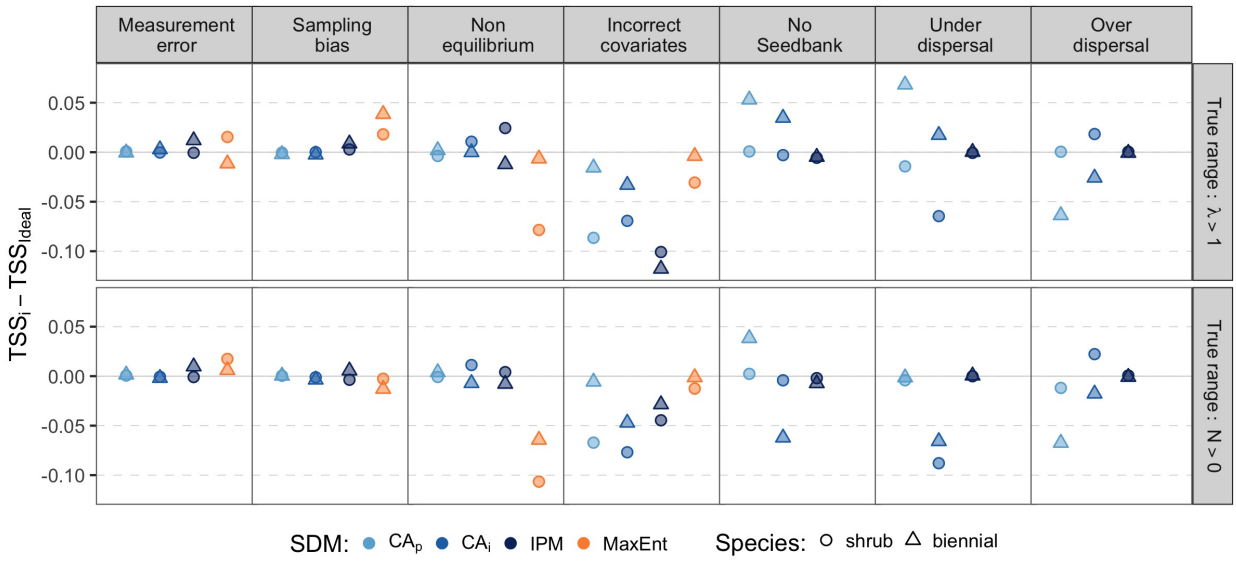

Figure 3: Effect of scenario on median true skill statistic (TSS) relative to the '*ideal*' scenario. Positive values indicate improved predictive ability, while negative values indicate decreased predictive ability.

SXX:Specificity-Ideal). Excluding the seed bank in the process-based models had minimal effect on TSS for the shrub, but variable effects on TSS for the biennial. Relative to 'ideal', all process-based models showed reduced sensitivity and increased specificity for the biennial, indicating a smaller predicted range. The IPM was generally more resistant to mischaracterized dispersal compared to the simulation-based CA_i and CA_p models. Underestimation of dispersal parameters led to decreased predicted ranges, and overestimation to increased predicted ranges relative to 'ideal'. The effects were more extreme for the biennial than for the shrub.

The process-based SDMs generate predictions for a number of biological quantities of interest. For instance, the IPM predicts λ in each cell of the landscape. Discrepancies in predicted λ values were biased downward in the northern portion of the study region (FIGURE 4), more strongly so for the shrub. Similarly, the CA models predict the abundance N for each species. Discrepancies in predicted abundance from the CA_p model showed a spatial banding of upward bias at the range margins for both species (FIGURE 4). Immediately interior to the range margin, the CA_p and CA_i model showed upward bias for the shrub (FIGURE 4).

4. Discussion

This is the first paragraph summarizing some general stuff. The process-based SDMs provide a lot more information, and they do, in fact, perform a bit better on average. They are less impacted by samples from non-equilibrium populations, so that is in line with expectations. However, they are more susceptible to mismatches in the covariates than MaxEnt, and the structural errors in the models don't necessarily have predictable effects.

Expand on the summary, bringing in past work on non-equilibrium distributions. Since process-based models aim to describe the biological processes that ultimately produce an emergent distribution, it really doesn't matter whether the samples are from a non-equilibrium species as long as the populations contain a sufficiently broad span of ages or sizes. This points to the fact that in parameterizing process-based models, care should be taken to select populations with a mix of ages or sizes rather from a range of environmental conditions than focusing on the geographic distribution per se, in essence capturing more of the parameter space. Thus, for non-equilibrium species, a well-described process-based SDM should do better than an occurrence-based model. . However, process-based models appear to be sensitive to several parts of the modelling process. First, they are much more sensitive to covariate choice, and by generalization, to the alignment of the modelled relationship between vital rates and the environment with reality. The flexible structure of MaxEnt, perhaps along with the top-down perspective, makes it much less influenced by errors in the covariates. Even with covariates that are highly correlated with the true covariates, there is still a large decrease in predictive ability for the process-based models.

Range boundaries can be defined in several ways. For the true distribution of each species, we use two definitions: abundance-based (cells with $N > 0$) and λ-based (cells with $\lambda > 1$). The process-based SDMs compared here each predict a specific type of range boundary. IPMs predict λ in each cell of the landscape, while the simulation-based CA models predict abundance. For MaxEnt, the range boundary definition is less explicit. The geo-located presences used as input imply an abundance-based range, but predicted distributions are frequently constrained by a data-based threshold (e.g., relative suitability values that encapsulate 95% of the observed data) to exclude both erroneous data and sink populations, which implies an effort to predict a λ-based range. Generally, process-based SDMs explicitly predict one type of range or another, while occurrence-based SDMs may be rather ambiguous. There may or may not be a big difference between the two range

definitions, but it will depend on the species. For example, the magnitude of the difference should vary with the number or distribution of marginal cells (lots of marginal cells mean large areas with middling occupancy probabilities), the typical annual stochasticity in the population abundances (more stochasticity means more discrepancies are likely), and dispersal limitation (high dispersal or a long introduction history means fewer areas that have not been reached, but also possibly a larger proportion of sink populations).

5. Acknowledgments

This project was funding through National Science Foundation award IIS-1717368.

6. Bibliography