Modeling Species Ranges

Sydne Record and Noah Charney

According to two separate analyses by the National Aeronautics and Space Administration (NASA) and the National Oceanic and Atmospheric Administration (NOAA), the 10 warmest years globally since 1880 occurred within the last 15 years. These reports are part of a body of mounting evidence that the Earth's climate is changing. Historically, the fossil record shows that global fluctuations in climate coincided with species extinctions or shifts in the geographic ranges of species distributions.

For example, Figure 1 maps the presence and absence of American beech (*Fagus grandifolia*) fossil tree pollen from cores of accumulated sediments at the bottoms of lakes and bogs in the eastern United States over the last 20,000 years. The northward expansion of beech from 20,000 years ago toward the present follows the retreat of a large glacier, the Laurentide ice sheet, which covered northern North America during this time period, illustrating how species ranges may shift with climatic change.

A key goal of ecology is to understand these very types of relationships between organisms and their environments. In the face of global climatic change, ecologists want to know how species ranges might be influenced by changes in environmental conditions. Understanding species range shifts is important to society because it has implications for the conservation of rare and endangered species and for species that provide important ecosystem services. For instance, there is strong interest among ecologists in understanding the fates of tree species in response to climate change, since trees provide humanity with timber, reduction of storm water run-off, energy savings in home cooling due to shading, and carbon sequestration to counteract carbon dioxide emissions.

To begin to understand how current and future ranges of species may differ, ecologists often fit species distribution models (SDMs). Standard inputs for SDMs include geographically referenced species occurrence (presence/absence) or abundance data and environmental data (e.g., climatic or land-use data) extracted from geographic information systems layers for the locations of species records (Fig. 2). Species distribution models fit correlative relationships between occurrence and environmental data. Fits of the models to current environmental conditions may be used to predict the range of the species of interest in unsampled locations (Fig. 2).

Alternatively, fits of the models to current conditions may be used to project the range of the species under historic (e.g., paleoclimatic) or potential future environments (e.g., based on different carbon dioxide emissions scenarios). Although SDMs are the most commonly used tool for understanding how species ranges may shift in response to climatic change, they have a number of biological and statistical shortcomings that present a great opportunity for ecologists and statisticians to collaborate.

Ecologists increasingly criticize standard SDMs for their lack of biological realism. Simple correlative relationships between species occurrence or abundance records and environmental variables ignore much of the interesting natural history of species for which ecologists have amassed evidence and theories.

Ecologists have long known that biotic interactions (e.g., species eating each other or competing with each other for food) influence distributions of species at local scales. Recent studies such as that by Belmaker, et al. (2015) illustrate that biotic interactions may also be important determinants of species ranges at large spatial scales (i.e., up to 50 km). Where the edge of a species' range falls and whether it is stable ultimately depends upon whether a species can colonize new areas beyond the edge (range expansion) and whether populations at the edge go extinct (range contraction).

Colonization of new areas is often a rare and unpredictable event, limited in organisms such as plants by how far a seed is carried from the parent by wind, water, or animals. Extinction, on the other hand, is often driven by stochastic events such as floods, fires, and disease outbreaks.

Standard SDMs assume that species have only two options in the face of global change: migrate or go extinct. However, evolutionary ecologists would argue that species, especially those with short generation times, have an additional course of action: adapt.

Populations of many species exhibit local adaptation to the environmental conditions in which they occur. Typical SDMs treat a species as a uniform population that does not exhibit local adaptation and whose tolerances cannot change with time, which limits the ability of projecting changes in local adaptation with changing environments. These are just a handful of examples of how conventional SDMs ignore biological realism.

In addition to being criticized for a lack of ecology, fitting of SDMs presents a multitude of statistical

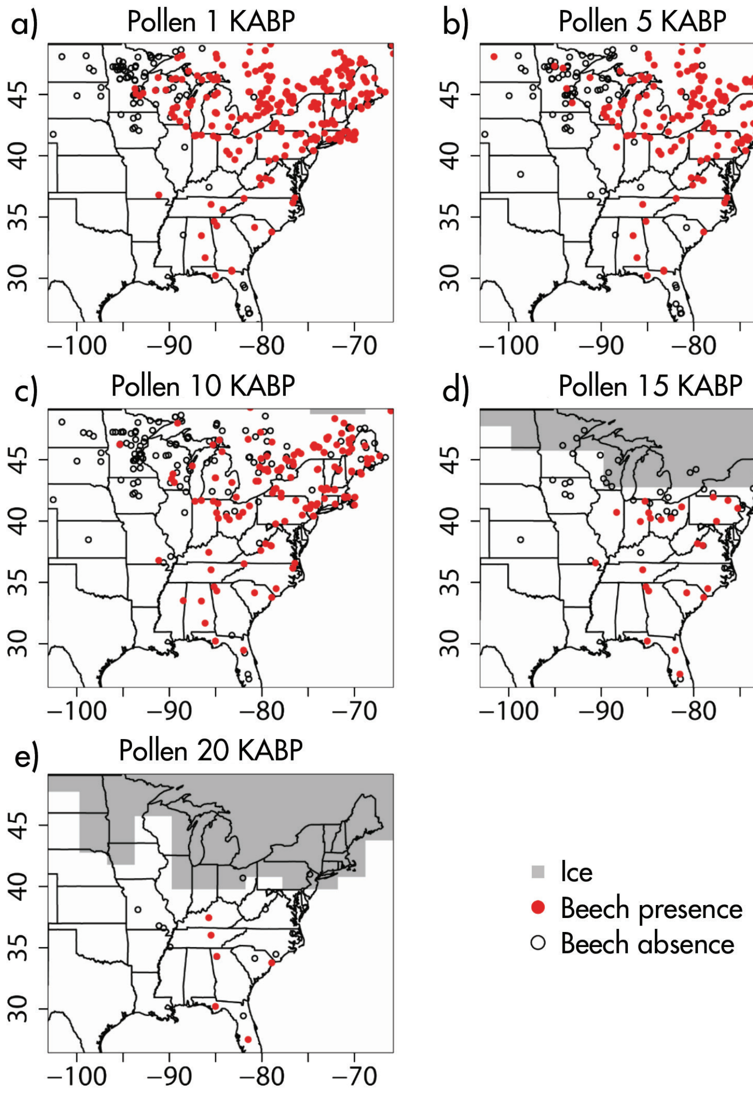

Figure 1. Maps of the presence (red points) and absence (white points) of fossil tree pollen from American beech (*Fagus grandifolia*) in the eastern United States at 1, 5, 10, 15, and 20 kiloannums before present (KABP), shown in panels a–e, respectively. Glacial ice (gray) overlaid onto the maps illustrates the tie between the shifting range of beech and climatic change. Pollen data provided by the Neotoma Paleoecology Database (*www.neotomadb.org*).

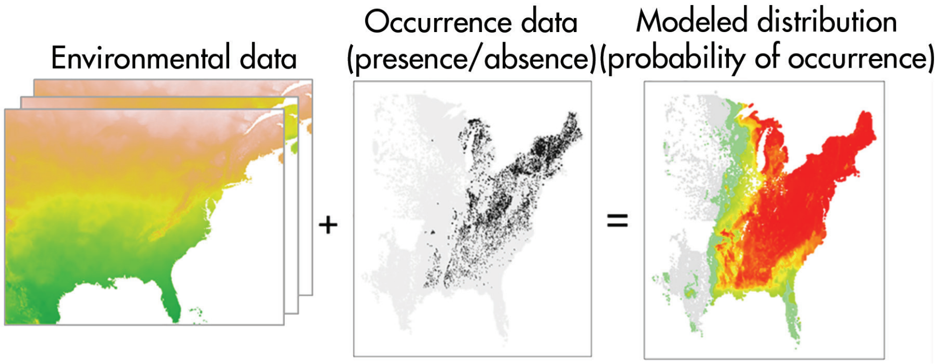

Figure 2. Species distribution models (SDMs) most commonly implemented by ecologists take species occurrence (i.e., presence/absence) or abundance data and environmental data as inputs. Ecologists frequently fit correlative models between species and environmental data to then map species ranges in sampled and unsampled locations across the landscape.

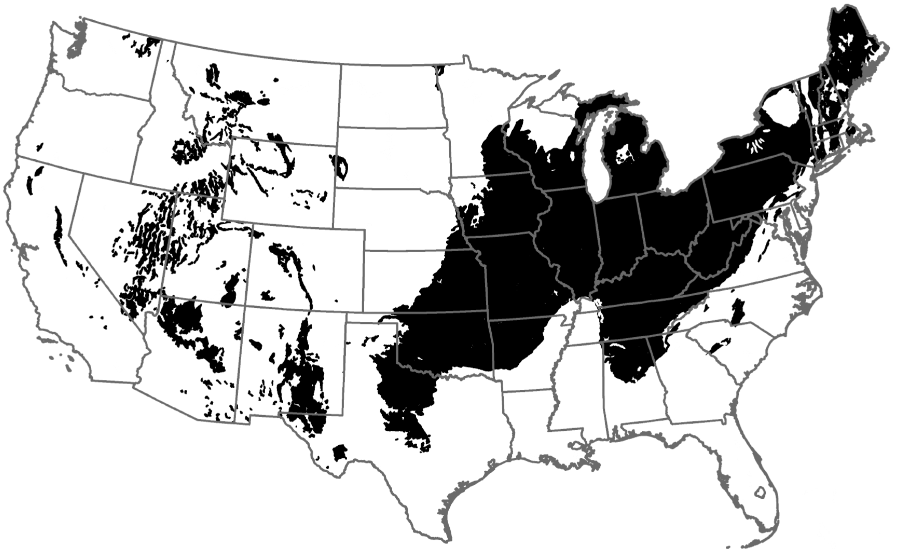

Figure 3. Distribution of Paleozoic sedimentary geologic units (black shading) within the United States (*http://ngmdb.usgs.gov/gmna/*).

challenges. For instance, most available species data are from observations where a species occurs (e.g., museum records) and lack information about localities where a species is absent. A lack of absence data for many analyses necessitates the generation of pseudo-absence localities based on a specification of criteria set by the researcher—Barbet-Massin, et al. (2012) suggest environmentally stratified random selection. This adds a level of uncertainty to the inferences based on the model.

There also often is a scale-mismatch between the environmental data used in SDMs and the processes being modeled. Most environmental data are available from geographic information systems layers with large grid cell sizes that may not capture microenvironmental heterogeneity relevant to the species of study. When considering the growth and survival of an individual seed that has landed on a patch of soil outside its species' range, regional annual precipitation may be less relevant than local soil moisture as controlled by small-scale factors such as topography, substrate, and shadows of nearby objects.

Another issue rarely addressed by conventional SDMs is spatial autocorrelation, where points that are closer in space are more similar to one another than points that are further apart in terms of environmental conditions or species occurrences. If data inputs for SDMs are autocorrelated spatially, the models violate the assumptions of independent and identically distributed residuals, potentially leading to incorrect conclusions. Greater risks of inferring that species responses depend upon unimportant variables and misinterpretations of error rates occur when data used to fit SDMs are spatially dependent.

The location of a species' range is often a historical byproduct of where the species happened to evolve. As illustrated by invasive species, at least some species survive quite well when transported manually to new regions outside their historic range. Can our models parse out the spatial patterns that are historical accidents versus the spatial patterns that are driven by biological constraints? For wide-ranging species, we might try to test an SDM built on one continent by going to another continent and seeing if the climate-species correlations hold up in an independent system. Many species, however, only occur on a single continent. In this sense, despite the volumes of data that go into an SDM, we are usually dealing with a sample size of just one, and at most seven.

All of these shortcomings stem from the fundamental challenge confronting SDMs: They are correlational, not mechanistic, models. Imagine that, instead of modeling the distribution of species, we tried to model the distribution of rocks. Consider the major geologic units underlying the United States, such as the swath of Paleozoic sedimentary deposits covering much of the northern half of the eastern United States (Figure 3).

Given this configuration, we probably could build a decent correlation-based model using a suite of climatic variables to predict the distribution of Paleozoic sedimentary rocks. The conceptual basis of such a model would be nonsensical, since these deposits were formed hundreds of millions of years ago when North America

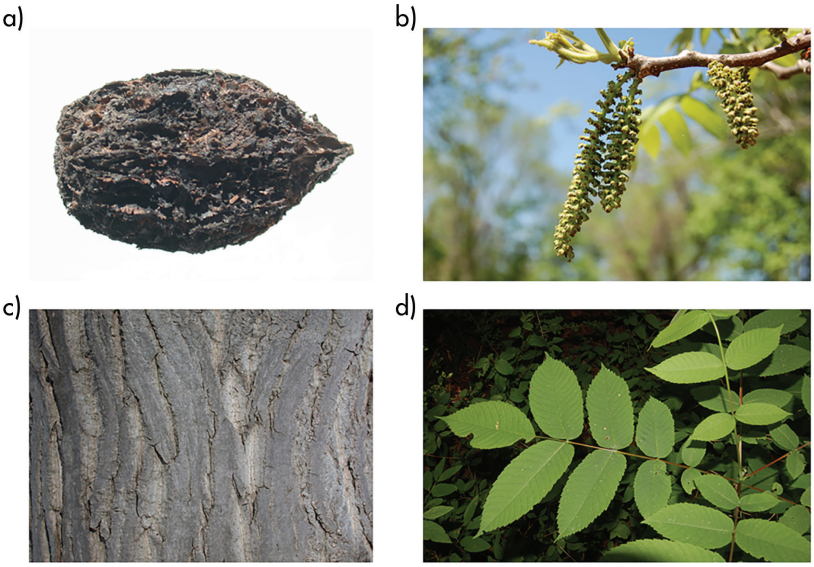

Figure 4. Butternut (*Juglans cinerea*) a) nut, b) male flowers, c) bark, and d) leaf.

Photos courtesy of C.S. Eiseman.

was near the equator, and their distribution is the result of plate tectonics, not the modern climate. Even though there is little causal link between climate and bedrock type, our model may still hold up fine within the United States, as long as the continent does not move and the climate does not change. If the climate changes, though, the correlation structure will break down and our model will begin to fail.

When we build an SDM, we are optimistic that our model will tell us something about the environmental limitations of a species, such as, given a species' biology, where can it survive? This information will allow us to make good predictions about the future, but the pessimists among us are afraid that our model will merely tell us that "the species is where it is." We can use such a model to draw a circle around the points in our data where the species was observed, but any future projections will be fraught. To tackle this problem, we can construct models with both environmental predictor variables and purely spatial variables. These spatial variables account for spatial autocorrelation, and can be thought of as null terms with no biological information other than location in space. To be useful in forecasting to different conditions, the biologically relevant variables should add significant power beyond what can be done with the spatial terms alone.

The Butternut Example

The butternut tree (*Juglans cinerea*), also called white walnut or oilnut, provides an example of modeling range. The butternut is an eastern North American tree species, and can illustrate differences between nonspatial models and models that account for spatial autocorrelation (Figure 4).

Lumber from butternut is occasionally used in woodworking and furniture-making, but the tree is most valued by humans, as well as wildlife, for its delicious and oily nuts. Historically, the bark of butternut was also used to dye cloth a light-tan shade, and Confederate soldiers during the Civil War were called Butternuts because some of the troops' uniforms were dyed with the bark of the tree or faded to a light-tan color (Brosi 2010). The bark and inner wood of butternut has also been used medicinally to treat smallpox, dysentery, toothaches, and digestive woes.

Butternut is native to the eastern United States and southeastern Canada, and tends to grow in low densities on the banks of streams and in well-drained soils. In terms of conservation status, although butternut is listed as globally secure, over half of the states in which butternut occurs list it as a species of conservation concern, with statuses ranging from "Endangered" in Tennessee and North Carolina to a "Species of Special Concern" in Minnesota and South Carolina. Declines of butternut populations throughout the species' range coincide with butternut canker disease linked to infection by the fungus *Sirococcus clavigignenti-juglandacearum*, which forms ulcers on the branches and stems that eventually kill the tree.

To model the distribution of butternut, and many other tree species across the United States, the Forest Inventory and Analysis (FIA) database, maintained by the United States Forest Service (USFS) as part of the nation's forest census (*http://www.fia.fs.fed.us/*), contains a plethora of data. Before field sampling, the USFS identifies "forested" land from satellite imagery. The USFS then sets up plots comprising four subplots with radii of 7.2 m each in every 2,428 hectares of forested land across the country. Within these plots, foresters take field measurements (e.g., species presence/absence, stem diameters). In the eastern half of the United States alone, there are more than 72,000 FIA plots, making this a very rich data set for inference.

While a number of different models could be employed to show the presence/absence of butternut (e.g., neural networks, classification and regression trees, Maximum Entropy [MaxEnt] models), here we model the current distribution of butternut parameterized with FIA occurrence data by fitting Bayesian generalized linear models (GLMs). Although GLMs are limited in their ability to infer limiting effects of range predictor variables (e.g., temperature, precipitation), the Bayesian framework enables estimation of the full posterior distribution at sampled and unsampled geographic locations.

Initial exploration of the data in which we fit nonspatial GLMs to the data and examined the residuals of the models indicated that there was likely spatial dependence between sampling locations. To explore whether accounting for spatial autocorrelation improved the fit of the model to the data, we fit both nonspatial and spatial versions of the Bayesian GLM. The main difference between the two types of models was that the spatial model included an additional term for a spatial random effect specified by a spatially varying intercept (SVI). The nonspatial model included climatic data as predictor variables, whereas we fit two spatial models: one with the SVI only and no climatic predictors and another with both climatic predictors and the SVI.

The purpose of fitting the spatial models, both with and without climatic predictor variables, was to see how much variability in the occurrence response the climatic variables accounted for, since such broad-scale environmental data are the most commonly used types of predictor variables in SDMs.

The response variable in all models was the presence or absence of butternut. We pulled climatic data at each FIA site for fitting the model from gridded climatic data generated from average monthly weather station data spanning the years 1950–2000 (*www.worldclim.org*). We began with a set of 21 bioclimatic variables, with 19 variables representing various summary statistics of temperature and precipitation, potential evapotranspiration (i.e., the amount of evaporation that would occur given plentiful water in an area; calculated according to Lugo, et al., 1999), and water balance (i.e., the sum of the monthly differences between precipitation and potential evapotranspiration).

From these 21 initial variables, we selected two that were not highly correlated and for which exploratory graphical analyses suggested a relationship with butternut occurrences. These variables were potential evapotranspiration and temperature seasonality (i.e., the standard deviation of monthly temperature values). Before modeling butternut's geographic distribution, we aggregated the plot-level data to a spatial scale of 50 km to reduce the potential for spatial misalignment between the occurrence data and the climatic data. We fit nonspatial and spatial GLMs with R statistical software (v. 3.2.3) using the spBayes package (v. 0.3-9) described in detail by Finley, et al. (2015).

The two spatial models provided better fits to the butternut occurrence data than the nonspatial model as evidenced visually (Fig. 5) and by various measures of model fit (i.e., Deviance Information Criterion, posterior predictive loss, and strictly proper scoring rules). The 2.5% and 97.5% percentiles for the parameter-credible intervals of the climatic variables did not overlap zero for the nonspatial model, but did overlap zero for the spatial model including coefficients for potential evapotranspiration, temperature seasonality, and the SVI. For both spatial models, the 2.5% and 97.5% percentiles for the parameter-credible intervals of the SVI did not overlap zero.

These results suggest that when spatial autocorrelation is not accounted for, incorrect inferences may be drawn about the relationships between broad-scale climatic drivers and butternut occurrences. Furthermore, the spatial random effect modeled by the SVI accounts for much more variation in butternut occurrences than potential evapotranspiration or temperature seasonality.

If we were to continue with this modeling exercise, next steps might include fitting the models to subsets of training data to test and compare their predictive performance (i.e., k-fold cross-validation) and/or projecting the models using historic or future climatic data.

Final Thoughts

Species distribution models provide a great opportunity for collaborations between statisticians and ecologists. The butternut example illustrates how ignoring assumptions of statistical models—independent and identically distributed residuals—may lead to incorrect inferences. We see that accounting for spatial autocorrelation in the spatial models with the spatial random effect improves the fit to the butternut occurrence

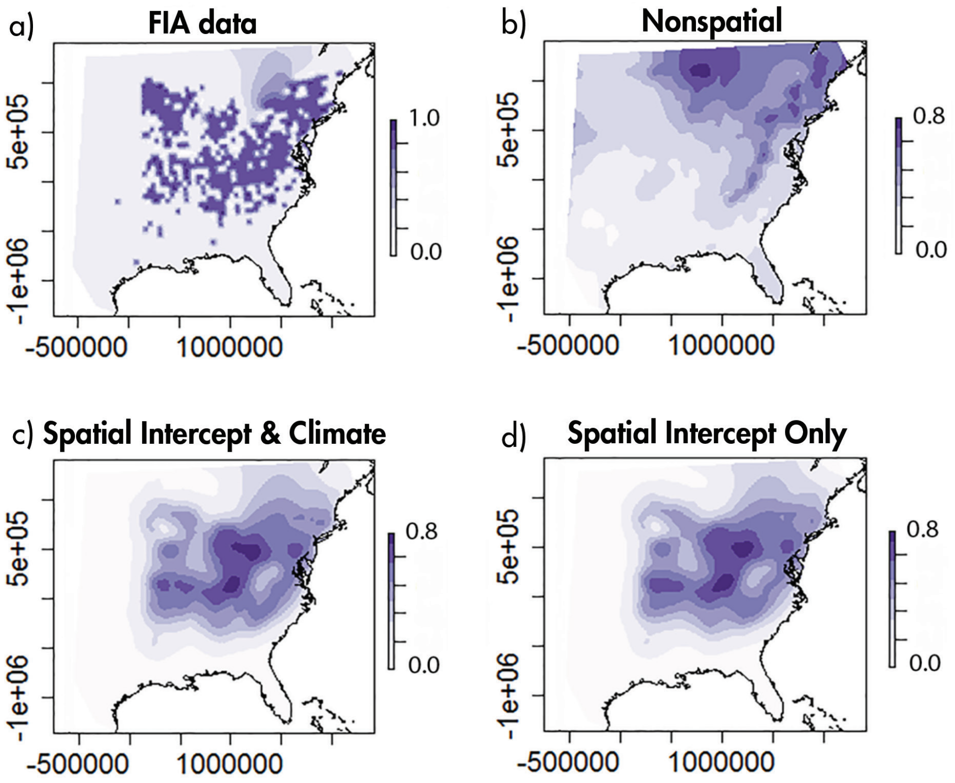

Figure 5. Maps of (a) a surface approximation of the probability of occurrence of butternut produced by fitting a multi-level B-spline to the raw Forest and Inventory Analysis (FIA) data and the predicted probability of presence using the R package MBA, (b) nonspatial, (c) spatially varying intercept plus climate, and (d) spatially varying intercept-only species distribution models fit to observed butternut FIA data. For panels a–d, the purple shading reflects the probability of occurrence of butternut.

data, while the environmental variables do very little to improve the model.

Record, et al. (2013) also found that GLMs—including spatially varying intercepts—outperformed nonspatial GLMs for two other tree species in the eastern United States, American beech (*F. grandifolia*) and eastern hemlock (*Tsuga canadensis*). Furthermore, when models for beech and hemlock were projected 8,000 years back in time with paleoclimatic data and validated with the fossil pollen record, the same study found that spatial GLMs resulted in lower false-positive rates than nonspatial GLMs.

For future projections, we might not expect static spatial structure modeled as a spatial random effect to hold for many parts of the world given rates of anthropogenic land-use change.

One explanation for why the SVI in the butternut example accounts for such a great amount of variation in the occurrence response relative to the climatic predictor variables is that the spatial random effect may be accounting for spatial structure in the response that is due to a missing covariate. As a tree that tends to prefer nutrient-rich sites, perhaps the butternut's range could be explained in part by the distribution of calcium-rich soils such as those derived from limestone. Limestone is one of the constituents in the Paleozoic sedimentary deposits mentioned earlier, the distribution of which seems to align with much of the butternut's range (Figure 3).

Although this may seem to be a plausible variable to consider, data on geology is difficult to fold into SDMs. The difficulty stems in part from the complexity of surficial geologic processes and in part from the format of geologic data, which consists of a myriad of names that do not readily translate into quantitative measures such as calcium concentration in derived soils. "Paleozoic," for instance, refers to the age of the deposits, not the mineral composition.

Other potential missing covariates in our SDM include fine-scale environmental variables (e.g., soil moisture, microclimatic variables) and biotic variables (e.g., species interactions, dispersal limitation, and population dynamics).

Incorporating spatial random effects in SDMs may help address spatial autocorrelation by accounting for missing covariates, however, this does not necessarily mean that SDMs fit with spatial terms will have good forecasting ability. If the spatial structure of the missing covariate(s) stays fixed under a changing climate, as in the case of a geologic formation, then forecasts that couple the spatial term with significant nonspatial terms might be quite good, but forecasts may be unreliable if the missing covariate shifts in the future, as in the case of predators, competitors, or pollinators, whose ranges are also climatically constrained.

In such cases, the structure of the correlations between space, climatic predictors, and hidden covariates will differ between current and future time periods. This same concern can be applied to the nonspatial terms in an SDM. For instance, temperature and precipitation are often correlated variables. If precipitation is the mechanistic driver, an SDM fit with only temperature as the predictor might work well for the current climate, because temperature is a good proxy for precipitation. However, climate change may alter the correlation between temperature and precipitation, undercutting the forecasting ability of such a model.

Such caveats will always be a concern for correlation-based SDMs, yet SDMs remain a critical ecological tool in both applied conservation and basic ecological theory, providing unique insights unavailable to other modeling approaches. Our task is to produce ever-more-trustworthy models, such as by integrating mechanistic models of biological processes (e.g., population dynamics, biotic interactions).

As we improve SDMs, it should become standard practice to test whether predictor terms add significant power beyond the null model with spatial terms only. Data availability often limits the ability of ecologists to incorporate relevant covariates into improved SDMs, but there are some treasure troves of data out there. Data from large-scale federally funded projects are of great value, as seen with the FIA data provided by the USFS. Other large-scale government-funded networks for ecological data include the Long Term Ecological Research Network and the National Ecological Observatory Network.

Remotely sensed fine-scale environmental variables from satellite imagery are also promising in terms of exploring potential missing covariates. For example, NASA's Soil Moisture Active Passive (SMAP) Earth satellite mission recently produced six months of soil moisture data for the top 5 cm of soils in areas with low-density vegetation globally.

In addition, ecologists are banding together in collaborative teams to collect data on biotic variables at large spatial scales. For instance, the PLANTPOPNET (*www.plantpopnet.com*) group consists of more than 66 researchers spread across the globe, collecting data on the population dynamics of narrow leaf plantain (*Plantago lanceolata*). Data sets collected by teams like PLANTPOPNET will enable more mechanistic modeling of biotic variables at continental or global scales. Greater dialogue between statisticians and ecologists will help to propel the next generation of SDMs forward with these new and exciting data sources to provide better estimates of species ranges. C

Further Reading

Barbet-Massin, M., F. Jiguet, C.H. Albert, and W. Thuiller. 2012. Selecting pseudo-absences for species distribution models: how, where, and how many? *Methods in Ecology and Evolution* 3:327–338.

Belmaker, J., P. Zarnetske, M.-N. Tuanmu, S. Zonneveld, S. Record, S. Strecker, and L. Beaudrot. 2015. Empirical evidence for the scale dependence of biotic interactions, *Global Ecology and Biogeography* 24:750–761.

Brosi, S.L. 2010. Steps toward butternut (*Juglans cinerea L.*) restoration. PhD dissertation, University of Tennessee Knoxville, *http://trace.tennessee.edu/utk_graddiss/779*.

Finley, A.O., S. Banerjee, and A.E. Gelfand. 2015. spBayes for large univariate and multivariate point-referenced spatio-temporal data models. *Journal of Statistical Software* 63:1-28.

Lugo, A.E., S.L. Brown, R. Dodson, T.S. Smith, and H.H. Shugart. 1999. The Holdridge life zones of the coterminous United States in relation to ecosystem mapping. *Journal of Biogeography* 26:1025-1038.

Record, S., M.C. Fitzpatrick, A.O. Finley, S. Veloz, and A.M. Ellison. 2013. Should species distribution models account for spatial autocorrelation? A test of model projections across eight millennia of climate change. *Global Ecology and Biogeography* 22:760-771.

About the Authors

Sydne Record is assistant professor of computational ecology at Bryn Mawr College and a research associate at Harvard Forest. Her interests include Bayesian modeling, community ecology, and biogeography.

Noah Charney is a postdoctoral researcher at the University of Arizona. His interests include conservation biology, spatial ecology, and natural history. His co-authored book (with Charley Eiseman), *Tracks and Sign of Insects and Other Invertebrates: A Guide to North American Species*, won the National Outdoor Book Aaward and American Library Association Outstanding Title recognition.