**Title:** Refining land cover maps in a hierarchical Bayesian framework to improve ecological models

**Authors:** Tim M. Szewczyk, Jenica M. Allen, Mark J. Ducey

**Affiliations:** Department of Natural Resources and the Environment, University of New Hampshire

**Submission for:** *Ecological Applications*

# ABSTRACT

Text.

KEY WORDS

Text.

INTRODUCTION

Our ability to model species’ distributions are constrained both by the availability of distributional data and the availability of data detailing the ecologically relevant drivers (Elith and Leathwick 2009, Kearney and Porter 2009, Allen et al. 2013, Duputié et al. 2013, Merow et al. 2014, 2017, Guillera-Arroita et al. 2015). While technological and methodological advancements have produced high quality datasets for factors such as climate or broad vegetation characteristics, other key variables remain elusive at an appropriate extent and resolution (He et al. 2015, Karger et al. 2017). For example, if a focal species associates strongly with a particular forest type, a reliable model of the species’ regional distribution or population dynamics would likely require knowledge of the distribution of the associated forest type. These more localized relationships and interactions are particularly important for predicting the effects of management at a landscape scale (Burnham and Lee 2010, Allen et al. 2013).

Land cover is a key driver of ecological patterns and processes, and is commonly incorporated into a species’ distribution model in the form of broad categories or binary suitability (Merow et al. 2011, Allen et al. 2013, Fordham et al. 2018). Within the United States, the commonly used National Land Cover Database has the benefit of extending across the entire contiguous United States with a consistent set of categories (Homer et al. 2007). Unfortunately, this consistency across a near continental extent carries the trade off of less specific categories (e.g., *Deciduous Forest*) and regionally variable accuracy (Wickham et al. 2010). Many states or subregions have developed land cover datasets which may provide more detailed categories (e.g., *Oak-Hickory Forest*) and have a higher density of verified locations (e.g., Justice et al. 2002). These, of course, carry the trade off of a reduced extent, limiting their ecological applicability.

The form of land cover data further complicates its use. Detailed categories must either be binarized into suitable and unsuitable habitat (Fordham et al. 2018), or aggregated into a smaller number of the most biologically relevant categories (Merow et al. 2011, 2017, Allen et al. 2013). Such compositional data, i.e., the proportional cover of each land cover category, presents statistical complications due to the constraints that each value must range from 0–1 and that, for a given area, the proportions must sum to one across categories. Further, the distribution of the land cover proportions is often zero-inflated, resulting from the natural distribution of land cover types (CITE). Various methods have been developed to address these issues, typically relying on a baseline category to enforce the sum-to-one constraint, and a transformation on the remaining categories (Aitchison J. 1982, Leininger et al. 2013). A recently proposed approach employs a latent, unrestrained representation of the proportions (Clark et al. 2017). With a model on the latent variable, a link function then enforces the necessary constraints, provides for zero- and one- inflation, and allows the latent values to remain approximately on the same scale as the proportional data to improve interpretability.

The performance of ecological models varies with spatial resolution. This is partly the result of scale-dependence in ecological processes (Turner et al. 1989, He et al. 2005, Chase and Knight 2013, Beaumont et al. 2016). At a regional scale, climate is often a key driver of distributions (Pearson and Dawson 2003), while fine scale distributions rely on much more local variables like light availability, for example (Elith and Leathwick 2009, Burnham and Lee 2010). However, the available resolutions also vary among predictors. For example, climate data is most robust at fairly coarse resolution (CITE), while anthropogenic data such as the length of roads in a pixel is reliable at very fine resolutions. Other covariates likely relevant to a species’ distribution, for example census data, are collected in irregular spatial configurations that change in both size and shape across the landscape (CITE). The validity of the assumptions that are necessarily made in order to use these kinds of data sources may vary with the spatial resolution or extent. Further, any spatially varying quantity exhibits some degree of spatial autocorrelation (CITE). While the covariates themselves may account for some autocorrelation in the response variable, the residuals still often show spatial autocorrelation. This may mark that an important predictor is missing or that geography itself is relevant (Elith and Leathwick 2009). Accounting for residual autocorrelation may be difficult at high resolutions due to computational limitations (Datta et al. 2016a, Finley et al. 2017). Of course, the choice of spatial resolution must also take into account the purpose of the model. A highly mechanistic, individual-based simulation of population growth might require covariates at a high resolution to effectively capture local dynamics. On the other hand, a correlative species distribution model may be informative at a much coarser resolution. As satellite-derived land cover data is available in many areas at ~30m resolution, these datasets can easily be scaled up to coarser resolutions as appropriate.

Glossy buckthorn (*Frangula alnus*) is an invasive shrub throughout the northeastern United States (Catling and Porebski 1994, Johnson et al. 2006, Webster et al. 2006). It shows particular associations with eastern white pine forests in New England, and there are economic incentives to optimize management in white pine forests from the forestry industry in New Hampshire and Maine (Frappier et al. 2003, 2004, Burnham and Lee 2010, Lee and Thompson 2012). Consequently, a population model with the goal of evaluating regional management strategies would be greatly improved, both in terms of ecological realism and applicability, by including white pine forests as a component of the landscape. Unfortunately, there is no land cover map that distinguishes eastern white pine forest from other evergreen forests that expands beyond a single state in the region. Instead, ecologists face the decision either to restrict the study extent to a single state such as New Hampshire, where a heavily ground-truthed dataset that distinguishes white pine is available (Justice et al. 2002), or to ignore the association with white pine forests and use a less accurate, less precise, but more geographically extensive land cover dataset (Homer et al. 2007, Wickham et al. 2010).

We introduce a statistical model as a solution to this problem. In a hierarchical Bayesian framework, this model rectifies two partly overlapping land cover datasets where the less specific map covers the full extent. Using topographic, climatic, and anthropocentric covariates, we estimate any systematic discrepancy between the mutual categories and split the aggregated category into the two more specific categories required across the study area. With the option to include spatial random effects, the model produces a cohesive map of compositional land cover for all required categories, providing estimates of both the expected value for each category and the uncertainty for the full extent.

METHODS

## Model structure

The desired result is the composition across *D* land cover categories within each cell *i* of the study area. The model requires two datasets at the same spatial resolution, **Y** and **Z**, which each describe the land cover composition within each cell. The first, **Y**, has high accuracy and all *D* categories, but does not cover the full extent. The second, **Z**, covers the full extent, but with lower accuracy and only *d* = *D* – 1 categories, where two of the *D* categories aggregated into one. In the example illustrated here, **Y** differentiates ecologically important *White Pine Forest* from *Other Evergreen Forest*, while in **Z**, it is aggregated within *Evergreen Forest*. For this second dataset, we represent the raw observed composition as **Z**, with **Z’** as the processed dataset with improved accuracy and the aggregate category refined into two.

In the model structure, both **Y** and **Z** are observations with error of the unconstrained latent variable **ν**. To enforce compositional constraints, we model *d* categories, with the *Dth* as a reference category calculated in the final link function. For each cell *i*, therefore, ***Yi***, ***Z’i***, and ***νi***, are vectors of length *d*, while ***Zi*** is a vector of length *d*–1 where *Zd-1,i* = *Z’d-1,i* + *Z’d,i*. Lastly, the compositionally constrained latent variable with *D* categories is calculated as ***ηi*** = *g*(***νi***). This transformation ensures that the elements of ***ηi***, which represent the latent compositional cover of categories in cell *i*, fall between 0–1 and sum to 1 (Clark et al. 2017).

The observed compositions for each cell *i* are modeled by *d*-variate normal distributions about ***νi***:

where **ΣY** and **ΣZ** are *d* x *d* covariance matrices for the error in **Y** and **Z’** respectively. The generation of **Z’** involves two components: (1) rectifying the systematic inaccuracy of **Z** relative to **Y**, and (2) refining the aggregate ***Zd-1*** category into the more specific ***Z’d-1*** and ***Z’d*** categories. This is performed for each cell *i* as:

where, for each land cover category *j* = 1, …, *d*–1, ***ρj*** is the expected inaccuracy in each cell *i*, **Vj** is a matrix of covariates, ***βj*** is a vector of slopes, and ***φj*** are spatial random effects, if included, to account for residual spatial autocorrelation. The vector ***p*** is predicted by covariates **X** and slopes ***θ***, and estimates the proportion of the aggregate ***Zd-1*** category expected to be ***Z’d-1***, with ***Z’d*** estimated as its complement. Thus, the slopes ***β*** and ***θ*** are indirectly informed by **Y** through its impact on ***ν*** within the fitting region, where data for both **Y** and **Z** are available. The full spatially explicit model is available (SuppStanMod).

## Data

We ran the model at two resolutions, using grids with 9km cells (3x3km) and 20-acre cells (~280x280m). The low-resolution grid contained 1662 cells with data for both **Y** and **Z**, and 385 cells with only **Z**. This allowed the inclusion of Conditional Autoregressive (CAR) spatial random effects as described above (Besag and Moran 1975). The high-resolution grid contained 193,811 cells with data for both **Y** and **Z**, and 46,845 cells with only **Z**. With several land cover categories modeled for each cell, run-time constraints precluded a spatially explicit model. We therefore used a nonspatial model for the 20a resolution, and included for comparison a nonspatial model for the 3km grid. Thus, we compare three models: a nonspatial 20a resolution model, a nonspatial 3km resolution model, and a spatially explicit 3km resolution CAR model.

We used two land cover datasets where the spatial extent of one was a subset of the extent of the other. For each, we aggregated the categories to *Open*, *Deciduous Forest*, *Mixed Forest*, *White Pine Forest* (**Y** only), *Other Evergreen Forest*, and *Other* (SuppAggScheme). For the subset, **Y**, we used the NH GRANIT 2001 Land Cover Assessment (Justice et al. 2002). Derived from satellite images acquired between 1990 and 1999, it classifies land cover at a 30m resolution across the state of New Hampshire, with an accuracy of 89.3% for the aggregated categories based on ground verification at nearly 1,000 sites (SuppConfMx). For the dataset covering the full extent, **Z**, we used the 2001 National Land Cover Database (Homer et al. 2007). Using Landsat data from primarily 2001, NLCD classifies land cover at a 30m resolution across the counterminous United States, with an accuracy of 82.8% for the aggregated categories within the New England region (SuppConfMx; Wickham et al. 2010). In QGIS 2.18 (CITE), we aggregated the land cover categories for each dataset at the initial 30m resolution before calculating the percent composition within each grid cell at the 3km and 20a resolutions.

To estimate the observed discrepancy between the land cover datasets, we explored covariates relating to climate, topography, and human development. We accessed climate data from CHELSA, which provides downscaled climatic variables at a 30 arc-second resolution based on data from 1979–2013 (Karger et al. 2017). To reduce collinearity while representing a diversity of climatic influences, we selected annual mean temperature, temperature annual range, and annual precipitation. We calculated the mean values within each cell for both the 3km grid and the 20a grid.

For topographic variables, we used elevation at a 1 arc-second resolution from the National Elevation Dataset (CITE USGS). In addition to elevation, we calculated a Terrain Ruggedness Index using the QGIS GDAL algorithm, where ruggedness is defined as the mean difference in elevation between a pixel and its neighbors. As with the climate variable, we calculated the mean value within each cell of the model grids.

For variables describing human development, we accessed census block data from the IPUMS National Historical Geographic Information System (Manson et al. 2017). To maximize temporal coherence with the land cover datasets we used data from the 2000 census. We used total population as well as the number of housing units for seasonal, recreational, or occasional use, which correlates with regional patterns of land use (CITE). Because census block shapes are irregular, reflecting patterns of human settlement, we assumed that the population and home densities were uniform within a given block. We then overlaid the 3km and 20a grids, allocated the intrablock densities proportionally among grid cells based on block area, and summed the allotted population and number of homes within each grid cell. In addition, we calculated the total length of roads within each grid cell using the USGS National Transportation Dataset (CITE USGS).

Lastly, we included a raster at 30m resolution of predicted probabilities that white pine was both the dominant tree species and comprised at least 30% of the total tree basal area, given that a pixel was classified as forest by the National Land Cover Database (CITE). Because the classification of *White Pine Forest* in the GRANIT dataset (**Y**) used slightly different criteria, we included this raster as an additional covariate for the estimation of ***p*** rather than as the sole estimate of the proportion of *White Pine Forest* within the *Evergreen Forest* category in **Z**. As with the climatic and topographic variables, we calculated the mean probability within each 3km and 20a grid cell.

## Model parameterization and comparison

During the Markov Chain Monte Carlo fitting proccess, the model is parameterized using cells that contain data for both **Y** and **Z**, and **Z’** is then predicted for cells without **Y** using the estimated slopes. Within the fitting region, **η** is calculated from the modeled estimates of **ν**. In the predicted region where **Y** is unavailable, **Z’** is used as a best estimate of **ν** since the model assumes that **Z’** is a representation of **ν** with only sampling error. Therefore, **η** is calculated in the predicted region from the predicted **Z’** rather than **ν**. The full Bayesian implementation includes appropriate prior distributions on **ν**, **ΣY**, **ΣZ**, ***β*** and ***θ*** (SuppStanMod).

We performed variable selection for each of the three model types (3km CAR, 3km Nonspatial, 20a Nonspatial) by splitting the cells with both **Y** and **Z** into a training subset (3km: 85% = 1412 cells; 20a: 15% = 29,072 cells) and a testing subset (3km: 15% = 250 cells; 20a: 85% = 164,739 cells), evaluating the out-of-sample predictive ability for the testing subset using the log predictive density (Gelman et al. 2013, Hooten and Hobbs 2015). The subsets were stratified random samples such that the quintiles for each covariate in the full dataset were represented proportionally. The training subset was used to parameterize the model using both **Y** and **Z**, while with the testing subset, predictions of **η** were generated using the estimated slopes and only **Z**. These predictions were then compared to the known values of **Y**. Due to the large number of possible combinations of individual variables, we categorized variables as *climate*, *census*, or *topographic*, and created variable sets based on all combinations of these categories (SuppVarSets). Within a variable set, we used the same variables for each ***ρ*** term such that **V1** = … = **Vd**. For **X**, we used the same variables but included the estimate of white pine cover as well. All variables were centered and scaled to improve model behavior. For each variable set, we ran 3 chains for each model, including 5,000 iterations for warmup and 5,000 iterations for posterior sampling, and retaining every 12th iteration for the posterior distribution.

After identifying the optimal variable set for each model, we parameterized the full model using all cells with data for both **Y** and **Z**, predicting the land cover proportions, **η**, for all cells as described above. We ran 3 chains for each full 3km model, including 10,000 iterations for warmup and 8,000 iterations for posterior sampling, retaining every 12th iteration to generate posterior distributions with 2,000 samples. We ran 24 chains for the full 20a model, including 1,000 iterations for warmup and 1,000 iterations for posterior sampling, retaining every 12th iteration to generate posterior distributions with 2,000 samples. For all models, we confirmed that the chains had converged and were well-behaved with an adequate warmup by inspecting R-hats, Geweke diagnostics, and ensuring that no divergences occurred (Gelman et al. 2013, Carpenter et al. 2017). Models were run in Stan 2.16.0 (Carpenter et al. 2017) through *rstan* 2.16.2 (Stan Development Team 2017a) in R 3.4.1 (CITE) and *Cmd-Stan* 2.17.0 (Stan Development Team 2017b).

To evaluate the ability of the model to rectify differences between **Z** and **Y**, we compared the posterior estimates of **η** with a simplified processing of **Z**, which we refer to as **Zs**. Specifically, we split the *Evergreen Forest* category in **Z** into *White Pine Forest* and *Other Evergreen Forest* by multiplying *Evergreen Forest* proportion in each cell by the covariate estimating white pine basal area. Compared to the model, this method does not attempt to correct for the lower accuracy of **Z** relative to **Y** and does not incorporate any additional covariates in the estimation of ***p***. Using the subset of cells reserved for testing during variable selection, we calculate the root mean squared error in predicting **Y** for each land cover class for **η** and for **Zs**.

# RESULTS

The study region is heavily forested with corridors of open habitat, driven by human activity, and patches of unsuitable habitat; these broad patterns occur in all estimates at both resolutions (FigPropMaps). The most notable disagreement between **Y** and **Zs** is in the *Mixed Forest* and *Deciduous Forest* categories, and all three models help to rectify this difference (FigPropMaps). Though the desired latent land cover proportions are, by definition, unknown, we use the more accurate **Y** as the nearest comparison. The posterior means, , are much nearer to **Y** than are the simply modified proportions of **Zs**, resulting in reduced residuals for nearly all land cover categories for each model (FigResidMaps). This reduction is more dramatic at the coarse (3km) resolution than the finer (20a) resolution.

For each model, the variable sets did not differ dramatically in their out-of-sample predictive ability (SuppVarSelection?). Nevertheless, we selected the variable set with the best performance as optimal for each full model. At the coarse resolution, the optimal variable set for spatially explicit model included *climate* and *census* variables, while the set for the nonspatial model included *climate* and *topographic* variables. At the fine resolution, the optimal variable set for the nonspatial model included only *climate* variables. Thus, *climate* variables occurred in the optimal variable set for all three models.

Including CAR spatial random effects greatly improved the model fit. The spatially explicit model showed lower root mean squared error (RMSE) than the nonspatial model for every land cover category in the out-of-sample testing cell subset during variable selection (TabVarSelRMSE). In the full model, the coarse resolution CAR model likewise showed very low residual error, measured as – **Y** in cells with data for both **Y** and **Z**, compared to both the nonspatial model and **Zs** (FigResidMaps). However, the uncertainty in **η** was quite different between the spatially explicit and nonspatial models (FigCI95). The nonspatial models showed larger uncertainty in the fitting region than in the predicted region. In contrast, the spatially explicit model showed greater uncertainty in the predicted region, with very narrow credible intervals in the fitting region where the estimates are informed by both **Y** and **Z**.

The model predictions for the testing subset of cells in the variable selection runs improve upon the simplistic **Zs** predictions for nearly every land cover category (TabVarSelRMSE). The percent change in root mean squared error (i.e., (RMSE**η**–RMSE**Zs**)/RMSE**Zs**) for the coarse resolution nonspatial model ranges from -45.9% – 1.7% (mean ± SE: -17.8% ± 7.4), and from -65.4% – -38.7% (mean ± SE: -54.8% ± 3.9) for the coarse resolution spatially explicit model, indicating that **η** predicts **Y** much more successfully than does **Zs**. At the fine resolution, the percent change in root mean squared error is less dramatic, ranging from -22.8% – 4.4% (mean ± SE: -4.6% ± 3.9) in the nonspatial model.

The difference between **η** and **Zs** represents the degree of bias correction. Ideally, the distribution of these differences should be comparable between the fitting region and the prediction region. The distribution of absolute difference between and **Zs** is in fact qualitatively similar between fitted and predicted cells for the coarse spatially explicit model (FigEtaZ), though the mean absolute difference significantly varied by region for each model type (P < 0.001). Nevertheless, all models show a tendency toward non-zero differences between and **Zs** for the majority of land cover categories, and the distribution of these differences for the fitted cells is comparable among models.

DISCUSSION

The hierarchical Bayesian model successfully produces a cohesive map covering the full extent, improving upon the simple alternative. It estimates the discrepancy between the more accurate but geographically constrained dataset **Y** and the less accurate but geographically more extensive dataset **Z**, and it splits the more specific category detailed in **Y** from the aggregated category in **Z**. The model performs more successfully at the coarser 3km resolution than the finer 20a resolution. The inclusion of spatial random effects improves the model estimates further.

Our results demonstrate the impact of spatial resolution on model performance, despite being a statistical model with limited direct biology. The choice of spatial resolution must weigh logistical constraints against the purpose for the map, e.g., population simulations, geographic distribution models, or local management decisions. The covariates available across an appropriately broad spatial extent vary in their resolution and their uncertainty at different resolutions. It is critical to carefully consider the informative abilities of covariates at various resolutions. Further, a finer resolution requires predicting a larger number of cells. For large datasets, spatial random effects may be included as Gaussian Point Processes (Banerjee et al. 2008, Finley et al. 2009, Eidsvik and Finley 2010). Rather than estimate each cell separately, a coarser grid of ‘knots’ is overlaid, with regional spatial correlation among knots. Alternatively, Nearest Neighbor Gaussian Processes estimate spatial correlation within a limited neighborhood for each cell (Datta et al. 2016a, b, Finley et al. 2017). Theoretically, the computation grows linearly with the number of cells rather than as the square or cube as with CAR models. However, because multiple spatial random effects are estimated for each cell in our model and operations cannot currently be parallelized within a Markov Chain, even these approaches are logistically untenable for very large numbers of cells. Despite residual spatial autocorrelation, the nonspatial model nevertheless performs better than the simple approach, though less well than the spatially explicit model.

The improved performance of the spatially explicit model over the nonspatial model suggests that the covariates did not fully describe the discrepancy between **Y** and **Z**. The covariates represent quantities that have shown to drive both species distributions and land cover (CITE). The estimation of the proportion of *White Pine Forest*, ***p***, essentially predicts the distribution of white pine conditional on the presence of evergreen-dominated forest, and so such biologically relevant covariates would be expected to perform well. However, the bias, ***ρ***, does not necessarily reflect biological processes. That is, ***ρ*** is the non-random disagreement between **Y** and **Z** in the proportion of each land cover type. Thus, the aim of the covariates is to capture systematic miscategorization of particular land cover types rather than to describe their actual distributions. While the covariates are in fact able to predict this to a degree, the difference between the nonspatial and spatially explicit models indicates that additional variables may further improve the bias estimation.

The spatially explicit and nonspatial models show dramatic differences in uncertainty. For cells in the prediction region, **η** is calculated directly from **Z’** using only Z, the predicted bias ***ρ***, and the predicted proportion of *White Pine Forest* within the *Evergreen Forest* category, ***p***. While there is indeed uncertainty in these estimates, it is reduced because **Z’** is used as a direct estimate of the latent variable **ν** in the predicted cells. In the fitted region, the uncertainty is greater because **ν** is hierarchically from both **Z’** and **Y**. In contrast, the spatially explicit model shows larger uncertainty in the predicted region. In the fitting region, where **η** is directly informed by both **Y** and **Z**, spatial random effects dramatically reduced uncertainty as the posterior converged tightly about **Y** with narrow 95% credible intervals. In the predicted region, ***φ*** was not informed or constrained by **Y**, resulting in the spatial propagation of uncertainty and correspondingly broad credible intervals. This uncertainty overwhelms the reduced hierarchical complexity.

This model allows ecologists to capitalize on the efforts of state and local governments to expand more accurate and detailed maps to reach beyond their political borders. This is particularly helpful for species that have affinities or aversions to a particular habitat type that we often cannot effectively include in distribution models that span regional extents. We aim to maximize the utility of localized mapping efforts and to improve models of species distributions, population dynamics, and management efforts. Though species do not recognize them, political borders often delineate the boundaries of our datasets, artificially restricting the geography of ecological models. This statistical method alleviates this problem, allowing for the combination and rectification of multiple data sources.

FIGURES

**FigPropMaps**: Maps of land cover proportions. Panel rows are LCs, panel columns are 3kmY, 3kmZ, 3kmCAR, 3kmNon, 20aY, 20aZ, 20aNon (6x7).

**FigResidMaps:** Maps of land cover residuals (estimate – Y). Panel rows are LCs, panel columns are 3kmZ, 3kmCAR, 3kmNon, 20aZ, 20aNon (6x5)

**FigCI95:** Boxplot of 95% CIs for 3kmCAR, 3kmNon, 20aNon. Panels are LCs

**FigEtaZ**: Boxplot of abs(posterior mean – Z) in predicting known Y proportions in optimal full models. Ideally, distribution of differences would be identical for fitted vs. predicted cells. CAR model generally does this, nonspatial models to a lesser extent. Difference from Z is typically 5-20% (raw composition).

TABLES

**TabRMSE**: Root mean squared error for each optimal variable selection model for predicted cells vs. GRANIT. Shows RMSE(posterior mean), RMSE(NLCD), % difference for each land cover category for each model (18 rows). Percent difference is largest for 3km CAR, smallest for 20a Non, though all show improvements for (nearly) all land cover categories. Best performance is with spatial random effects and a coarser resolution.

SUPPLEMENTS

**SuppStanMod:** Stan code for the model with CAR spatial random effects.

**SuppAggScheme:** Lookup table for original NLCD and GRANIT categories with the aggregated categories used for the model.

**SuppConfMx:** Confusion matrices for the aggregated NLCD and GRANIT categories.

REFERENCES

Aitchison J. 1982. The Statistical Analysis of Compositional data. - J. R. Stat. Soc. Ser. B (Statistical Methodol. 44: 139–177.

Allen, J. M. et al. 2013. Socioeconomics drive woody invasive plant richness in New England, USA through forest fragmentation. - Landsc. Ecol. 28: 1671–1686.

Banerjee, S. et al. 2008. Gaussian predictive process models for large spatial data sets. - J. R. {…} 70: 825–848.

Beaumont, L. J. et al. 2016. Which species distribution models are more (or less) likely to project broad-scale, climate-induced shifts in species ranges? - Ecol. Modell. in press.

Besag, J. E. and Moran, P. A. P. 1975. On the estimation and testing of spatial interaction in Gaussian Lattice Processes. - Biometrika 62: 555–562.

Burnham, K. M. and Lee, T. D. 2010. Canopy gaps facilitate establishment, growth, and reproduction of invasive Frangula alnus in a Tsuga canadensis dominated forest. - Biol. Invasions 12: 1509–1520.

Carpenter, B. et al. 2017. *Stan* : A Probabilistic Programming Language. - J. Stat. Softw. in press.

Catling, P. M. and Porebski, Z. S. 1994. The history of invasion and current status of glossy buckthorn, Rhamnus frangula, in southern Ontario. - Can. Field-Naturalist in press.

Chase, J. M. and Knight, T. M. 2013. Scale-dependent effect sizes of ecological drivers on biodiversity: why standardised sampling is not enough. - Ecol. Lett. 16 Suppl 1: 17–26.

Clark, J. S. et al. 2017. Generalized joint attribute modeling for biodiversity analysis: Median-zero, multivariate, multifarious data. - Ecol. Monogr. 87: 34–56.

Datta, A. et al. 2016a. Hierarchical Nearest-Neighbor Gaussian Process Models for Large Geostatistical Datasets. - J. Am. Stat. Assoc. 111: 800–812.

Datta, A. et al. 2016b. On nearest-neighbor Gaussian process models for massive spatial data. - Wiley Interdiscip. Rev. Comput. Stat. 8: 162–171.

Duputié, A. et al. 2013. Where are the wild things? Why we need better data on species distribution. - Glob. Ecol. Biogeogr.: n/a-n/a.

Eidsvik, J. and Finley, A. O. 2010. Approximate Bayesian Inference for Large Spatial Datasets Using Predictive Process Models. in press.

Elith, J. and Leathwick, J. R. 2009. Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. - Annu. Rev. Ecol. Evol. Syst. 40: 677–697.

Finley, A. O. et al. 2009. Improving the performance of predictive process modeling for large datasets. - Comput. Stat. Data Anal. 53: 2873–2884.

Finley, A. O. et al. 2017. Applying Nearest Neighbor Gaussian Processes to Massive Spatial Data Sets: Forest Canopy Height Prediction Across Tanana Valley Alaska.: 1–49.

Fordham, D. A. et al. 2018. How complex should models be? Comparing correlative and mechanistic range dynamics models. - Glob. Chang. Biol. 24: 1357–1370.

Frappier, B. et al. 2003. Potential Impacts of the Invasive Exotic Shrub Rhamnus Frangula L. (Glossy Buckthorn) on Forests of Southern New Hampshire. - Northeast. Nat. 10: 277–296.

Frappier, B. et al. 2004. Experimental Removal of the Non-indigenous Shrub Rhamnus frangula (Glossy Buckthorn): Effects on Native Herbs and Woody Seedlings. - Northeast. Nat. 11: 333–342.

Gelman, A. et al. 2013. Bayesian Data Analysis. - Chapman & Hall.

Guillera-Arroita, G. et al. 2015. Is my species distribution model fit for purpose? Matching data and models to applications. - Glob. Ecol. Biogeogr. 24: 276–292.

He, F. et al. 2005. The local–regional relationship: Immigration, extinction, and scale. - Ecology 86: 360–365.

He, K. S. et al. 2015. Will remote sensing shape the next generation of species distribution models? - Remote Sens. Ecol. Conserv. 1: 4–18.

Homer, C. et al. 2007. Completion of the 2001 National Land Cover Database for the Conterminous United States. - Photogramm. Eng. Remote Sens. 73: 337–341.

Hooten, M. B. and Hobbs, N. T. 2015. A guide to Bayesian model selection for ecologists. - Ecol. Monogr. 85: 3–28.

Johnson, V. S. et al. 2006. The role of spatial and temporal scale in colonization and spread of invasive shrubs in early successional habitats. - For. Ecol. Manage. 228: 124–134.

Justice, D. et al. 2002. New Hampshire land cover assesment.

Karger, D. N. et al. 2017. Climatologies at high resolution for the earth’s land surface areas. - Sci. Data 4: 1–20.

Kearney, M. and Porter, W. 2009. Mechanistic niche modelling: Combining physiological and spatial data to predict species’ ranges. - Ecol. Lett. 12: 334–350.

Lee, T. D. and Thompson, J. H. 2012. Effects of logging history on invasion of eastern white pine forests by exotic glossy buckthorn (Frangula alnus P. Mill.). - For. Ecol. Manage. 265: 201–210.

Leininger, T. J. et al. 2013. Spatial Regression Modeling for Compositional Data With Many Zeros. - J. Agric. Biol. Environ. Stat. 18: 314–334.

Manson, S. et al. 2017. IPUMS National Historical Geographic Information System: Version 12.0 [Database].

Merow, C. et al. 2011. Developing Dynamic Mechanistic Species Distribution Models: Predicting Bird-Mediated Spread of Invasive Plants across Northeastern North America. - Am. Nat. 178: 30–43.

Merow, C. et al. 2014. Advancing population ecology with integral projection models: A practical guide. - Methods Ecol. Evol. 5: 99–110.

Merow, C. et al. 2017. Climate change both facilitates and inhibits invasive plant ranges in New England. - Proc. Natl. Acad. Sci. 114: E3276–E3284.

Pearson, R. G. and Dawson, T. 2003. Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? - Glob. Ecol. Biogeogr. 12: 361–371.

Stan Development Team 2017a. RStan: the R interface to Stan. in press.

Stan Development Team 2017b. CmdStan: the command-line interface to Stan, Version 2.17.0. in press.

Turner, M. G. et al. 1989. Effects of changing spatial scale on the analysis of landscape pattern. - Landsc. Ecol. 3: 153–162.

Webster, C. R. et al. 2006. Woody invaders and the challenges they pose to forest ecosystems in the eastern United States. - J. For. 104: 366–374.

Wickham, J. D. et al. 2010. Thematic accuracy of the NLCD 2001 land cover for the conterminous United States. - Remote Sens. Environ. 114: 1286–1296.