**Running Head:** Refining land cover maps for ecological models

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

**Authors:** Tim M. Szewczyk\*1,2, Mark J. Ducey1, Valerie Pasquarella3,4, Jenica M. Allen1

**Affiliations:**

1Department of Natural Resources and the Environment, University of New Hampshire

2Department of Computer Science, University of New Hampshire

3Department of Environmental Conservation, Univerisity of Massachusetts Amherst

4Northeast Climate Science Center, University of Massachusetts Amherst

**\*Corresponding author address:**

114 James Hall, Durham, NH 03823, USA

timothy.szewczyk@unh.edu

**Submission for:** *Ecological Applications*

**Tables:** 2

**Figures:** 4

**Appendices:** 3

# **ABSTRACT**

Ecological models are constrained by the availability of high quality data at biologically appropriate extents. Modeling a species' affinity or aversion with a particular type of land cover relies on data detailing that land cover across the full study area. Precise datasets must often sacrifice geographic breadth, while geographically broad datasets often compromise on precision and accuracy. Ecologists face the choice of restricting the study extent to match the precise dataset, or ignoring potentially key associations to study a larger area. We introduce a hierarchical Bayesian model to reconcile two partly overlapping proportional land cover datasets, where one more accurately details all necessary categories but is less extensive, and the other covers the full extent but with lower accuracy and two categories aggregated into one. The model estimates systematic discrepancies between the data sets and divides the aggregated category into the two more specific categories across the full extent. We illustrate with the invasive shrub glossy buckthorn (*Frangula alnus*) which associates with Eastern white pine (*Pinus strobus*) forests in New England. We use the National Land Cover Database, which extends across the full study area but with generalized forest categories, and the New Hampshire GRANIT land cover dataset, which only exists within the state of New Hampshire but with high accuracy and *White Pine Forest* as a category. We evaluate the model at coarse (9 km2) and fine (20 acre) resolutions, with and without spatial random effects at the coarse resolution. Each model successfully produced a cohesive map of proportional land cover for the full study extent, reducing the inaccuracy of the geographically extensive dataset while partitioning *White Pine Forest* out of *Evergreen Forest*. Accuracy was higher at the coarser 9 km2 resolution and with spatial random effects included. Each model improved estimates overall compared to simply partitioning the *Evergreen Forest* category in the less precise dataset based on the predicted distribution of white pine. This flexible statistical method helps ecologists capitalize on localized mapping efforts and expand models of species distributions, population dynamics, and management approaches beyond the artificial political boundaries that delineate land cover datasets.

# **KEY WORDS**

Compositional; Conditional autoregressive model; Spatially explicit; Species distribution model

# **INTRODUCTION**

Species distribution models are constrained 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, Merow et al. 2014a, 2017, Duputié et al. 2014, Guillera-Arroita et al. 2015). While technological and methodological advancements have produced high quality datasets describing climate or broad vegetation characteristics, other key variables remain elusive at an appropriate extent and resolution (Sulla-Menashe et al. 2011, He et al. 2015, Karger et al. 2017). For a species that 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. Localized relationships and interactions are particularly important for predicting the effects of management at a landscape scale (Guisan and Thuiller 2005, Burnham and Lee 2010, Allen et al. 2013, Guisan et al. 2013). Highly detailed datasets are often limited in extent, and consequently may only be availble for a subset of the ecologically relevant region.

Ecological processes are scale-dependent, and accordingly, the performance of ecological models varies with spatial resolution (Turner et al. 1989, Rahbek and Graves 2001, He et al. 2005, Chase and Knight 2013, Beaumont et al. 2016). Ideally, the purpose of the model drives the choice of spatial resolution (Wiens 1989, Belmaker et al. 2015, Jackson and Fahrig 2015). Fine scale distributions are often shaped by local variables such as light availability (Guisan and Thuiller 2005, Elith and Leathwick 2009, Burnham and Lee 2010), and a highly mechanistic, individual-based simulation of population growth might require a high resolution to effectively capture local dynamics (Grimm and Railsback 2013). In contrast, regional distributions are often driven by climate (Pearson and Dawson 2003), and a correlative model of a species' continental distribution may be informative at much coarser resolutions (Phillips and Dudík 2008, Merow et al. 2013). However, the geographic attributes of predictor data also affect model performance. Climate data are most robust at fairly coarse resolutions (Karger et al. 2017), while anthropocentric variables such as road density may be reliable for a wide range of resolutions. In contrast, human population density and other census data are collected in irregular spatial configurations that change in size and shape across the landscape and through time (U.S. Census Bureau 2017). Reconciling these datasets with a grid requires assumptions about the geographic distribution of the data within each irregular polygon.

The successful prioritization of management efforts must account for uncertainty rather than relying solely on best estimates of species distributions or invasive species risk (Guisan et al. 2013). Uncertainty in a the predicted occurrence probability for a species may arise from the quality or distribution of observed presence data, the correspondence between the included covariates and the actual biological drivers, and the accuracy and precision of the covariate data (Barry and Elith 2006, Merow et al. 2014b, Burke et al. 2015). Bayesian methods allow the propagation of known uncertainty in the covariates or model structure to generate posterior distributions that more faithfully describe the plausible outcomes given current biological understanding (Wikle 2003, Clark 2005, Gelfand et al. 2006, Latimer et al. 2006, Merow et al. 2014b). In models predicting biological responses to future climate change, covariate uncertainty is often accounted for by incorporating multiple climatic projections (Burke et al. 2015, Allen and Bradley 2016, Merow et al. 2017). However, data for current conditions also have uncertainty which may vary geographically. Ecological models and subsequent management decisions are more robust when accounting for this uncertainty where possible (Lydakis et al. 2018).

Spatially varying quantities often exhibit spatial autocorrelation, where correlation among observations is dependent upon spatial proximity (Legendre 1993, Diniz-Filho et al. 2003, Latimer et al. 2006, Warren et al. 2014). Autocorrelation in covariates may account for much of the spatial distribution in the response variable. However, patterning in residual error may indicate that an important predictor is missing or that geography itself impacts the distribution, as when a species’ range is limited by dispersal (Legendre 1993, Elith and Leathwick 2009). Statistically accounting for spatial autocorrelation may be difficult at high resolutions across large areas due to computational limitations (Datta et al. 2016a, Finley et al. 2017).

Land cover is a key driver of ecological patterns and processes, and is commonly incorporated into species distribution models (Merow et al. 2011, Allen et al. 2013, Fordham et al. 2018). Currently available datasets must balance geographic extent against precision and accuracy. For example, the National Land Cover Database extends across the contiguous United States with a consistent set of categories (Homer et al. 2007), but the categories are general (e.g., *Deciduous Forest*) and accuracy is regionally variable (Wickham et al. 2010). Some states or subregions have produced datasets with greater specificity (e.g., *Oak-Hickory Forest*) and a higher density of verified locations to improve accuracy (e.g., Justice et al. 2002), but with reduced geographic extent, potentially restricting applications to artificial political boundaries.

The form of land use/land cover data further complicates its use. Detailed categories must either be dichotomized into suitable and unsuitable habitat (Guisan et al. 2013, Fordham et al. 2018), or aggregated into a smaller number of the most biologically relevant categories (Ibáñez et al. 2009, Merow et al. 2011, 2017, Allen et al. 2013). Biogeographical models typically operate at a coarser resolution than land cover data, and each grid cell is consequently described by the proportional cover across land cover categories (Elith and Leathwick 2009, Ibáñez et al. 2009, Merow et al. 2011, Allen et al. 2013, Bellard et al. 2013). Compositional data, including proportional cover, present statistical complications arising from the constraints that each value is bounded by [0,1] and that, for a given area, the proportions must sum to one across categories. Further, the proportions are often zero-inflated due to the naturally patchy distribution of land cover types (Leininger et al. 2013). Various methods have been developed to account for these issues, typically relying on a baseline category to enforce the sum-to-one constraint, and a transformation on the remaining categories (Aitchison 1982, Aitchison and Egozcue 2005, Leininger et al. 2013). A recently proposed approach employs an unobserved, or latent, representation of the proportions in an unconstrained space (Clark et al. 2017). This latent variable is then used in the model with a link function that enforces the necessary constraints, allows for zero- and one-inflation, and keeps the latent values on the same scale as the proportional data to improve interpretability (Clark et al. 2017).

Regional ecological models incorporating specific land cover types therefore face several challenges. Here, we use the example of glossy buckthorn (*Frangula alnus*), an invasive shrub throughout the northeastern United States (Catling and Porebski 1994, Johnson et al. 2006, Webster et al. 2006). Glossy buckthorn shows particular associations with eastern white pine (*Pinus strobus*) in New England, and there are economic incentives to manage buckthorn populations in commercial white pine forests in New Hampshire and Maine (Frappier et al. 2003, 2004, Burnham and Lee 2010, Lee and Thompson 2012). A model evaluating regional management strategies for glossy buckthorn would consequently be more ecologically realistic and more practically applicable by including white pine forests as a component of the landscape. Currently, there is no land cover map distinguishing eastern white pine forest from other evergreen forests that extends beyond a single state in the region. Instead, ecologists must restrict models to areas where such a dataset is available (Justice et al. 2002), or 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 to reconcile two compositional land cover datasets that are individually incomplete for the desired application. In a hierarchical Bayesian framework, our model rectifies two partly overlapping land cover datasets, where one dataset contains all necessary categories but is less extensive, and the other dataset covers the full extent but with two categories aggregated into one. Using climatic, topographic, and anthropocentric covariates, we estimate systematic discrepancies between the mutual categories and split the aggregated category into the two more specific categories across the full extent. With the option to account for spatial autocorrelation, the model produces a cohesive map of compositional land cover for all required categories, providing estimates of the expected value and uncertainty for each category across the full geographic extent.

# **METHODS**

## **Study System**

The motivating species is glossy buckthorn (*Frangula alnus*), a shrub native to Eurasia and invasive throughout the northeastern United States (Catling and Porebski 1994, Johnson et al. 2006, Webster et al. 2006). Glossy buckthorn reduces regeneration of eastern white pine (*Pinus strobus*), an important timber species in New Hampshire and Maine (Frappier et al. 2003, 2004, Burnham and Lee 2010, Lee and Thompson 2012, Lanzer et al. 2017). We focus our example on the southern portions of New Hampshire and Maine where glossy buckthorn is most abundant and eastern white pine is relatively common (Justice et al. 2002). A land cover dataset that partitions eastern white pine is available only for the state of New Hampshire (Justice et al. 2002) rather than for the full extent.

## **Model structure**

The model predicts the land cover composition across *D* categories within each cell *i* of a grid for the full extent. We employ two datasets, **Y** and **Z**, to describe the land cover composition within cells (Table 1). **Y** has high accuracy and all *D* categories, but only exists for a subset of cells. **Z** covers the full extent, but with lower accuracy and only details *d*=*D*–1 categories, where two of the *D* categories are aggregated into one. In the example illustrated here, **Y** differentiates ecologically important *White Pine Forest* from *Other Evergreen Forest*. In **Z**, *White Pine Forest* is aggregated within *Evergreen Forest*. We represent the raw observed composition as **Z**, with **Z’** as the processed composition with improved accuracy and the aggregate category partitioned into the desired two.

Both **Y** and **Z** are represented as 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 by the final link function. For each cell *i*, therefore, ***Yi***, ***Z’i***, and ***νi***, are vectors of length *d*, and ***Zi*** is a vector of length *d*–1 such that *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***). The link function *g*(.) ensures that the elements of ***ηi***, which represent the true compositional cover in cell *i*, fall between 0–1 and sum to unity (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) partitioning 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, **Vj** is a matrix of covariates, ***βj*** is a vector of slopes, and ***φj*** are spatial random effects to account for 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. Note that **Z’** and **ν** are not constrained to be compositional. In the spatially explicit model, we include spatial random effects following an exact sparse Conditional Autoregressive (CAR) model with , where *τj* is a precision parameter, **D** is a diagonal matrix with the number of neighbors for each cell, *αj* varies between 0–1 and determines the degree of spatial dependence, and **W** is an adjacency matrix identifying neighbors (Besag and Moran 1975, Jin et al. 2005, Joseph 2016). The spatially explicit model is provided in Appendix S1.

## **Data**

We ran the model at two resolutions, using grids with 9 km2 cells (3x3 km) and 20-acre cells (~280x280 m). These resolutions fall within a range typical of regional species distribution models (e.g., Guisan et al. 2007, Manzoor et al. 2018). The low-resolution 9 km2 grid contained 1,662 cells with data for both **Y** and **Z**, and 385 cells with only **Z**. The high-resolution 20 ac 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 high-resolution spatially explicit model. We therefore used a non-spatial model for the 20 ac resolution, excluding ***φ*** from the model, and included for comparison a non-spatial model for the 9 km2 grid. Thus, we compare three models: a non-spatial 20 ac resolution model, a non-spatial 9 km2 resolution model, and a spatially explicit 9 km2 resolution CAR model.

For each land cover dataset, we aggregated the categories to *Open Canopy*, *Deciduous Forest*, *Mixed Forest*, *White Pine Forest* (**Y** only), *Other Evergreen Forest*, and *Other* (Appendix S3: Table S1-S2). For the dataset with the smaller extent, **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 ~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 (Appendix S3: Table S3). For the dataset covering the full extent, **Z**, we used the 2001 National Land Cover Database (NLCD; Homer et al. 2004, 2007). Using Landsat data from primarily 2001, NLCD classifies land cover at ~30 m resolution across the conterminous United States, with an accuracy of 82.8% for the aggregated categories within the New England region (Appendix S3: Table S4; Wickham et al. 2010). In QGIS 2.18 (QGIS Development Team 2017), we aggregated the land cover categories for each dataset at the initial 30 m resolution before calculating the percent composition within each cell for the 9 km2 and 20 ac resolution grids.

To estimate the observed discrepancy between the **Y** and **Z**, 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 9 km2 grid and the 20 ac grid.

For topographic variables, we used elevation at a 1 arc-second resolution from the National Elevation Dataset (Gesch et al. 2002). In addition, 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 variables, we calculated the mean value within each cell of the model grids.

For human development variables, 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 population and home densities were uniform within each block. We then overlaid the 9 km2 and 20 ac grids, allocated the intra-block 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 primary and secondary roads within each grid cell from the USGS National Transportation Dataset (U.S. Geological Survey 2014).

Lastly, we included a raster at ~30 m resolution of predicted white pine distribution. A Random Forest classifier (Breiman 2001) was trained using Forest Inventory and Analysis data, and predicted the forest type of each pixel following the methods of Pasquarella (2018). The probability of white pine forest is calculated as the proportion of the 500 decision trees classifying a pixel as white pine forest, given that the pixel was classified as forest by NLCD 2011. The classification of the training data required that eastern white pine was both the dominant tree species in the plot and comprised ≥ 30% of the total tree basal area (Pasquarella et al., unpublished data). Because of the difference in classification criteria for *White Pine Forest* in **Y** (Justice et al. 2002) as well as the contingency on any NLCD 2011 forest category rather than on NLCD 2001 *Evergreen Forest* as in our model, we included the predicted white pine distribution as an additional covariate in **X** to estimate ***p*** rather than as the direct proportion of *White Pine Forest* within the *Evergreen Forest* category in **Z**. We calculated the mean probability within each 9 km2 and 20 ac 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**, defined as the 'fitting region', and then **Z’** is predicted for cells without **Y**, defined as the 'prediction region', using the estimated slopes. Within the fitting region, **η** is calculated from the modeled estimates of **ν**. The model assumes that **Z’** is a representation of **ν** with only sampling error, and so **Z’** is used directly as a best estimate of **ν** in the prediction region. Therefore, **η** is calculated in the prediction region from the predicted **Z’** rather than **ν** (Table 1, Appendix S1). The full Bayesian implementation includes prior distributions on **ν**, **ΣY**, **ΣZ’**, ***β***, ***θ***, ***α***, and ***τ***. Additionally, we used several reparameterizations to increase efficiency based on the algorithms implemented by Stan (Stan Development Team 2017a). The prior distributions and reparameterizations are detailed in Appendix S1-S2.

We performed variable selection for each of the three model versions (9 km2 spatially explicit, 9 km2 non-spatial, 20 ac non-spatial) by splitting the cells with both **Y** and **Z** into a training subset (9 km2: 85% = 1,412 cells; 20 ac: 15% = 29,072 cells) and a testing subset (9 km2: 15% = 250 cells; 20 ac: 85% = 164,739 cells). Though it would be preferable to use 85% of cells for training at 20 ac, the computation time rendered this infeasible; the number of cells nevertheless provides a large sample. Further, each training subset was a stratified random sample such that the quintiles for each covariate were represented proportionally. Each model was parameterized with the training subset using both **Y** and **Z**, and predictions of **η** were generated for the testing subset 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*, *topography*, or *human development*, and created variable sets based on all combinations of these categories for a total of seven candidate variable sets. Within a variable set, we used the same variables for each ***ρ*** term such that **V1** = … = **Vd-1**. For **X**, we used the same variables as for **V**, but also included the estimated distribution of white pine. All variables were centered and scaled to improve model behavior. For each variable set, we ran 3 chains for each model version, including 5,000 iterations for warmup and 5,000 iterations for posterior sampling, and retaining every 12th iteration for the posterior distribution for a posterior sample size of 1,250 per variable set. To identify the optimal variable set for each model, we evaluated the out-of-sample predictive ability for the testing subset using the log predictive density, selecting the variable set with the highest log predictive density (Gelman et al. 2013, 2014, Hooten and Hobbs 2015).

Using the optimal variable sets, we parameterized each full extent 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 extent 9 km2 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 extent 20 ac 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 with Stan 2.16.0 (Carpenter et al. 2017) through *rstan* 2.16.2 (Stan Development Team 2017b) in R 3.4.1 (R Core Team 2017) and *Cmd-Stan* 2.17.0 (Stan Development Team 2017c).

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**, referred to hereafter as **Zs**. To generate **Zs**, we split the *Evergreen Forest* category from **Z** into *White Pine Forest* and *Other Evergreen Forest* by multiplying the *Evergreen Forest* proportion in each cell by the white pine probability raster. 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 variable selection testing subset, we calculated the root mean squared error for **η** and for **Zs** relative to **Y** for each land cover class at each resolution.

# **RESULTS**

The study region is heavily forested with corridors of open habitat, driven by human activity, and patches of unsuitable habitat; these broad patterns are represented in all estimates at both resolutions (Fig. 1). 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 (Fig. 1). Though the desired latent land cover proportions are, by definition, unknown, we use the more accurate **Y** as the nearest comparison. The posterior means, , reduced residual error for nearly all land cover categories for each model compared to the simply modified proportions of **Zs** (Fig. 2). This reduction is more dramatic at the coarse (9 km2) resolution than the finer (20 ac) resolution, and for the spatially explicit model than the non-spatial model.

For each model, the variable sets did not differ dramatically in their out-of-sample predictive ability. At the coarse resolution, the optimal variable set for the spatially explicit model included *climate* and *human development* variables, while the set for the non-spatial model included *climate* and *topography* variables. At the fine resolution, the optimal variable set for the non-spatial model included only *climate* variables. Thus, *climate* variables occurred in the optimal variable set for all three models.

The CAR spatial random effects in the spatially explicit model greatly improved the model fit and predictions. In the variable selection testing subset, the spatially explicit model showed lower root mean squared error (RMSE) than the non-spatial models for every land cover category and dramatically reduced RMSE relative to **Zs** (Table 2). Performance was also higher at 9 km2 resolution than at 20 ac resolution. The percent change in root mean squared error (i.e., (RMSE**η**–RMSE**Zs**)/RMSE**Zs**) ranged from -45.9% – 1.7% (mean ± SE: -17.8% ± 7.4) for the 9 km2 non-spatial model, and from -65.4% – -38.7% (mean ± SE: -54.8% ± 3.9) for the 9 km2 spatially explicit model. For the 20 ac non-spatial model, the percent change in root mean squared error was less dramatic, ranging from -22.8% – 4.4% (mean ± SE: -4.6% ± 3.9).

The difference between **η** and **Zs** represents the degree of bias correction. With informative covariates, the distribution of these differences should be similar in the fitting region and the prediction region. For the 9 km2 spatially explicit model, the distribution of absolute difference between and **Zs** was qualitatively similar between fitted and predicted cells (Fig. 3). For the non-spatial models, the distributions were somewhat more variable between the fitting and prediction regions. All models showed a tendency toward non-zero differences between and **Zs** for the majority of land cover categories, though the bias correction in the prediction region was stronger in the 9 km2 models (Fig. 3).

At the full extent, the spatially explicit model likewise showed very low residual error, measured as – **Y** in cells with data for both **Y** and **Z**, compared to both the non-spatial model and **Zs** (Fig. 2). However, the geographic pattern of uncertainty in **η** was quite different between the spatially explicit and non-spatial models (Fig. 4). For the non-spatial models uncertainty was largest in the fitting region and low in the prediction region. In contrast, the spatially explicit model showed much higher uncertainty in the prediction region, with very low uncertainty in the fitting region where the estimates were informed by both **Y** and **Z**.

# **DISCUSSION**

The hierarchical Bayesian model successfully produced a cohesive map of proportional land cover for the full study extent. The model estimated the discrepancy between the more accurate, but geographically constrained dataset **Y** and the less accurate but geographically more extensive dataset **Z**, while reconciling categorization differences by partitioning *White Pine Forest* out of *Evergreen Forest* in **Z**. The accuracy was higher at the coarser 9 km2 resolution than the finer 20 ac resolution, and the inclusion of spatial random effects further improved the model predictions. Each model outperformed the simple approach of multiplying the *Evergreen Forest* proportions in **Z** by the probability of white pine forest.

The impact of spatial resolution on model performance likely arises from two primary sources. First, the discrepancy between the land cover composition datasets becomes increasingly extreme as resolution increases (Appendix S3: Fig. S1). At high resolutions, the proportional cover within each grid cell is calculated from fewer pixels. Each pixel with conflicting classifications strongly influences the proportional cover. At coarser resolutions, each discrepancy carries less weight, and differences between the proportions are less extreme as broader landscape trends dominate. Second, the covariates differ in their spatial patterning. It is critical to carefully consider the informative capacity of covariates at various resolutions. For example, climate, which varies across the landscape at a coarse scale (Pearson and Dawson 2003, Karger et al. 2017), may best inform broad regional differences such as the underclassification of *Mixed Forest* by **Z** in the southeastern portion of the study region (Fig. 1), rather than discrepancies between neighboring pixels. In contrast, elevation, human density, and road density exhibit variation at a much finer scale, and may modulate local estimates amidst the regional effects of climate (Huston 1999, Keil et al. 2012, Fraterrigo et al. 2014).

However, only *climate* variables were retained in the optimal high resolution model; the fine-scale disagreement between **Y** and **Z** was poorly explained by the *topographic* and *human development* variables despite their localized variation, and residual error remained high compared to the coarse resolution models (Fig. 2). Given the improved performance of the spatially explicit model over the non-spatial model at the coarse resolution, the covariates did not fully capture the discrepancy between **Y** and **Z** at either resolution. Climate, topography, and human development are demonstrated drivers of species distributions and land cover (Munroe and Müller 2007, Elith and Leathwick 2009, Redo and Millington 2011, Allen et al. 2013). These biologically relevant covariates would be expected to perform well in the estimation of the proportion of *White Pine Forest* within evergreen-dominated forests. In contrast, the estimated bias does not necessarily reflect biological processes, but instead the non-random disagreement between **Y** and **Z** in the proportion of each land cover type. In the bias estimation component of the model, the aim of the covariates is to predict systematic miscategorization of particular land cover types rather than to describe the actual distribution of land cover types. Categorization differences may arise from the source imagery used to produce each dataset, the algorithms used to process and classify pixels, or the definitions of the land cover categories. Each dataset used different categories (Justice et al. 2002, Homer et al. 2004, 2007), which we aggregated according to the expected biological preferences of glossy buckthorn (Appendix S3: Table S1-S2). It is possible that the aggregation scheme could obscure relationships with covariates. Land cover patterns are also affected by historical land use. In our study region, the distribution of pine forests is heavily influenced by the distribution of abandoned cropland (Cogbill et al. 2002). Unfortunately, data detailing historic land use patterns do not exist at an appropriately broad extent. Further, some species, such as hemlock, are difficult to classify (CITE). The supervised classification methods for **Y** likely account for some additional accuracy. While the model does predict the discrepancy between **Y** and **Z**, additional variables may further improve the bias estimation.

A finer resolution requires predicting a larger number of cells, and we were consequently unable to include computationally expensive spatial random effects for the high resolution model (Finley et al. 2009, Leininger et al. 2013, Datta et al. 2016a). Recent advances have produced several options to account for spatial autocorrelation without the intensive matrix operations traditionally required (Banerjee 2017). Low-rank or reduced-rank models estimate spatial correlation by overlaying a coarser grid of ‘knots’ rather than estimating effects for each location separately and can describe broad spatial patterns (Banerjee et al. 2008, Finley et al. 2009, Eidsvik et al. 2012, Banerjee 2017). Alternatively, Nearest Neighbor Gaussian Processes estimate spatial correlation within a limited neighborhood for each observation (Datta et al. 2016b, 2016a, Finley et al. 2017). Theoretically, the computation costs grow linearly with the number of cells rather than as the cube as with CAR models (Datta et al. 2016a). 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. However, the non-spatial models performed better than the simple approach at both resolutions despite residual spatial autocorrelation, though less well than the spatially explicit model (Fig. 2).

In addition to mean estimates of land cover proportions for each cell, the Bayesian model produces estimates of uncertainty, which varies geographically and among land cover categories. In applications of the model output, the uncertainty can help inform decisions and appropriately temper inferences. The spatially explicit and non-spatial models show dramatic differences in uncertainty. For cells in the prediction region for the non-spatial models, **η** is calculated directly from **Z’** using only **Z**, the predicted bias, and the predicted proportion of *White Pine Forest* within the *Evergreen Forest* category. While there is uncertainty in these estimates, it is reduced because **Z’** is used in place of the latent variable **ν** in the predicted cells. In the fitting region, the uncertainty is greater because **ν** is estimated hierarchically from both **Z’** and **Y**. In contrast, the spatially explicit model shows larger uncertainty in the prediction region. In the fitting region, where **η** is directly informed by both **Y** and **Z**, spatial random effects dramatically reduce uncertainty as the posterior converges about **Y** with narrow credible intervals. In the prediction region, ***φ*** is neither informed nor constrained by **Y**, resulting in the spatial propagation of uncertainty and correspondingly broad credible intervals. This uncertainty overwhelms the reduced complexity of using **Z’** in place of **ν**. In the full extent models, the predicted cells are extended geographically from the fitted region (Fig. 1), rather than interspersed among the fitted cells. Because spatial random effects are informed by neighboring observations, the prediction region is effectively a data gap, leading to larger spatial uncertainty (Datta et al. 2016a).

This model allows ecologists to capitalize on the efforts of state and local governments to expand more accurate and detailed maps beyond their political borders. This is particularly helpful for species with particular habitat affinities or aversions that often cannot be included in regional distribution models. If more than two land cover datasets are available in the study region, the model structure can easily accommodate that additional information. We aim to maximize the utility of localized mapping efforts and to improve models of species distributions, population dynamics, and management efforts. Political borders often delineate the boundaries of our datasets, artificially restricting the geography of ecological models. The statistical method described here helps to alleviate this problem, allowing for the combination and rectification of multiple data sources.

# **REFERENCES**

Aitchison, J. 1982. The Statistical Analysis of Compositional data. Journal of the Royal Statistical Society Series B (Statistical Methodology) 44:139–177.

Aitchison, J., and J. J. Egozcue. 2005. Compositional Data Analysis: Where Are We and Where Should We Be Heading? Mathematical Geology 37:829–850.

Allen, J. M., and B. A. Bradley. 2016. Out of the weeds? Reduced plant invasion risk with climate change in the continental United States. Biological Conservation 203:306–312.

Allen, J. M., T. J. Leininger, J. D. Hurd, D. L. Civco, A. E. Gelfand, and J. A. Silander. 2013. Socioeconomics drive woody invasive plant richness in New England, USA through forest fragmentation. Landscape Ecology 28:1671–1686.

Banerjee, S. 2017. High-Dimensional Bayesian Geostatistics. Bayesian Analysis 12:583–614.

Banerjee, S., A. E. Gelfand, A. O. Finley, and H. Sang. 2008. Gaussian predictive process models for large spatial data sets. Journal of the Royal Statistical Society Series B (Statistical Methodology) 70:825–848.

Barry, S., and J. Elith. 2006. Error and uncertainty in habitat models. Journal of Applied Ecology 43:413–423.

Beaumont, L. J., E. Graham, D. E. Duursma, P. D. Wilson, A. Cabrelli, J. B. Baumgartner, W. Hallgren, M. Esperón-Rodríguez, D. A. Nipperess, D. L. Warren, S. W. Laffan, and J. VanDerWal. 2016. Which species distribution models are more (or less) likely to project broad-scale, climate-induced shifts in species ranges? Ecological Modelling 342.

Bellard, C., W. Thuiller, B. Leroy, P. Genovesi, M. Bakkenes, and F. Courchamp. 2013. Will climate change promote future invasions? Global Change Biology 19:3740–3748.

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

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

Breiman, L. 2001. Random forests. Machine Learning 45:5–32.

Burke, M., J. Dykema, D. B. Lobell, E. Miguel, and S. Satyanath. 2015. Incorporating Climate Uncertainty into Estimates of Climate Change Impacts. Review of Economics and Statistics 97:461–471.

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

Carpenter, B., A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell. 2017. *Stan* : A Probabilistic Programming Language. Journal of Statistical Software 76.

Catling, P. M., and Z. S. Porebski. 1994. The history of invasion and current status of glossy buckthorn, Rhamnus frangula, in southern Ontario. Canadian Field-Naturalist 108:305–310.

Chase, J. M., and T. M. Knight. 2013. Scale-dependent effect sizes of ecological drivers on biodiversity: Why standardised sampling is not enough. Ecology Letters 16:17–26.

Clark, J. S. 2005. Why environmental scientists are becoming Bayesians. Ecology Letters 8:2–14.

Clark, J. S., D. Nemergut, B. Seyednasrollah, P. J. Turner, and S. Zhang. 2017. Generalized joint attribute modeling for biodiversity analysis: Median-zero, multivariate, multifarious data. Ecological Monographs 87:34–56.

Cogbill, C. V., J. Burk, and G. Motzkin. 2002. The forests of presettlement New England, USA: Spatial and compositional patterns based on town proprietor surveys. Journal of Biogeography 29:1279–1304.

Datta, A., S. Banerjee, A. O. Finley, and A. E. Gelfand. 2016a. Hierarchical Nearest-Neighbor Gaussian Process Models for Large Geostatistical Datasets. Journal of the American Statistical Association 111:800–812.

Datta, A., S. Banerjee, A. O. Finley, and A. E. Gelfand. 2016b. On nearest-neighbor Gaussian process models for massive spatial data. Wiley Interdisciplinary Reviews: Computational Statistics 8:162–171.

Diniz-Filho, J. J. A. F., L. Mauricio Bini, B. A. Hawkins, and L. M. Bini. 2003. Spatial Autocorrelation and red herrings in geographical ecology. Global Ecology and Biogeography 12:53–64.

Duputié, A., N. E. Zimmermann, and I. Chuine. 2014. Where are the wild things? Why we need better data on species distribution. Global Ecology and Biogeography 23:457–467.

Eidsvik, J., A. O. Finley, S. Banerjee, and H. Rue. 2012. Approximate Bayesian Inference for Large Spatial Datasets Using Predictive Process Models. Computational Statistics and Data Analysis 56:1362–1380.

Elith, J., and J. R. Leathwick. 2009. Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. Annual Review of Ecology, Evolution, and Systematics 40:677–697.

Finley, A. O., A. Datta, B. C. Cook, D. C. Morton, H. E. Andersen, S. Banerjee, U. States, and F. Service. 2017. Applying Nearest Neighbor Gaussian Processes to Massive Spatial Data Sets: Forest Canopy Height Prediction Across Tanana Valley Alaska. arXiv preprint:arXiv:1702.00434.

Finley, A. O., H. Sang, S. Banerjee, and A. E. Gelfand. 2009. Improving the performance of predictive process modeling for large datasets. Computational Statistics and Data Analysis 53:2873–2884.

Fordham, D. A., C. Bertelsmeier, B. W. Brook, R. Early, D. Neto, S. C. Brown, S. Ollier, and M. B. Araújo. 2018. How complex should models be? Comparing correlative and mechanistic range dynamics models. Global Change Biology 24:1357–1370.

Frappier, B., R. T. Eckert, and T. D. Lee. 2003. Potential Impacts of the Invasive Exotic Shrub Rhamnus Frangula L. (Glossy Buckthorn) on Forests of Southern New Hampshire. Northeastern Naturalist 10:277–296.

Frappier, B., R. T. Eckert, and T. D. Lee. 2004. Experimental Removal of the Non-indigenous Shrub Rhamnus frangula (Glossy Buckthorn): Effects on Native Herbs and Woody Seedlings. Northeastern Naturalist 11:333–342.

Fraterrigo, J. M., S. Wagner, and R. J. Warren. 2014. Local-scale biotic interactions embedded in macroscale climate drivers suggest Eltonian noise hypothesis distribution patterns for an invasive grass. Ecology letters.

Gelfand, A. E., J. A. J. Silander, S. Wu, A. Latimer, J. a Silander, S. Wuz, A. Latimerx, P. O. Lewis, A. G. Rebelok, and M. Holder. 2006. Explaining species distribution patterns through hierarchical modeling. Bayesian Analysis 1:41–92.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. 2013. Bayesian Data Analysis. 3rd edition. Chapman & Hall, London.

Gelman, A., J. Hwang, and A. Vehtari. 2014. Understanding predictive information criteria for Bayesian models. Statistics and Computing 24:997–1016.

Gesch, B. D., M. Oimoen, S. Greenlee, C. Nelson, M. Steuck, and D. Tyler. 2002. The National Elevation Dataset. Photogrammetric Engineering & Remote Sensing 68:1–9.

Grimm, V., and S. F. Railsback. 2013. Individual based modeling and ecology. Princeton University Press.

Guillera-Arroita, G., J. J. Lahoz-Monfort, J. Elith, A. Gordon, H. Kujala, P. E. Lentini, M. A. McCarthy, R. Tingley, and B. A. Wintle. 2015. Is my species distribution model fit for purpose? Matching data and models to applications. Global Ecology and Biogeography 24:276–292.

Guisan, A., C. H. Graham, J. Elith, F. Huettmann, M. Dudik, S. Ferrier, R. Hijmans, A. Lehmann, J. Li, L. G. Lohmann, B. Loiselle, G. Manion, C. Moritz, M. Nakamura, Y. Nakazawa, J. M. C. Overton, A. T. Peterson, S. J. Phillips, K. Richardson, R. Scachetti-Pereira, R. E. Schapire, S. E. Williams, M. S. Wisz, and N. E. Zimmermann. 2007. Sensitivity of predictive species distribution models to change in grain size. Diversity and Distributions 13:332–340.

Guisan, A., and W. Thuiller. 2005. Predicting species distribution: Offering more than simple habitat models. Ecology Letters 8:993–1009.

Guisan, A., R. Tingley, J. B. Baumgartner, I. Naujokaitis-Lewis, P. R. Sutcliffe, A. I. T. Tulloch, T. J. Regan, L. Brotons, E. Mcdonald-Madden, C. Mantyka-Pringle, T. G. Martin, J. R. Rhodes, R. Maggini, S. A. Setterfield, J. Elith, M. W. Schwartz, B. A. Wintle, O. Broennimann, M. Austin, S. Ferrier, M. R. Kearney, H. P. Possingham, and Y. M. Buckley. 2013. Predicting species distributions for conservation decisions. Ecology Letters 16:1424–1435.

He, F., K. J. Gaston, E. F. Connor, and D. S. Srivastava. 2005. The local–regional relationship: Immigration, extinction, and scale. Ecology 86:360–365.

He, K. S., B. A. Bradley, A. F. Cord, D. Rocchini, M. N. Tuanmu, S. Schmidtlein, W. Turner, M. Wegmann, and N. Pettorelli. 2015. Will remote sensing shape the next generation of species distribution models? Remote Sensing in Ecology and Conservation 1:4–18.

Homer, C., J. Dewitz, J. Fry, M. Coan, N. Hossain, C. Larson, N. Herold, A. McKerrow, J. N. VanDriel, and J. D. Wickham. 2007. Completion of the 2001 National Land Cover Database for the Conterminous United States. Photogrammetric Engineering & Remote Sensing 73:337–341.

Homer, C., C. Huang, L. Yang, B. Wylie, and M. Coan. 2004. Development of a circa 2000 landcover database for the United States. Photogrammetric Engineering & Remote Sensing 70:829–840.

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

Huston, M. a. 1999. Local Processes and Regional Patterns: Appropriate Scales for Understanding Variation in the Diversity of Plants and Animals. Oikos 86:393–401.

Ibáñez, I., J. A. Silander, J. M. Allen, S. A. Treanor, and A. Wilson. 2009. Identifying hotspots for plant invasions and forecasting focal points of further spread. Journal of Applied Ecology 46:1219–1228.

Jackson, H. B., and L. Fahrig. 2015. Are ecologists conducting research at the optimal scale? Global Ecology and Biogeography 24:52–63.

Jin, X., B. P. Carlin, and S. Banerjee. 2005. Generalized hierarchical multivariate CAR models for areal data. Biometrics 61:950–961.

Johnson, V. S., J. A. Litvaitis, T. D. Lee, and S. D. Frey. 2006. The role of spatial and temporal scale in colonization and spread of invasive shrubs in early successional habitats. Forest Ecology and Management 228:124–134.

Joseph, M. B. 2016. Exact sparse CAR models in Stan. http://mc-stan.org/users/documentation/case-studies/mbjoseph-CARStan.html.

Justice, D., A. Deely, and F. Rubin. 2002. New Hampshire land cover assesment.

Karger, D. N., O. Conrad, J. Böhner, T. Kawohl, H. Kreft, R. W. Soria-Auza, N. E. Zimmermann, H. P. Linder, and M. Kessler. 2017. Climatologies at high resolution for the earth’s land surface areas. Scientific Data 4:1–20.

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

Keil, P., O. Schweiger, I. Kühn, W. E. Kunin, M. Kuussaari, J. Settele, K. Henle, L. Brotons, G. Pe’er, S. Lengyel, A. Moustakas, H. Steinicke, and D. Storch. 2012. Patterns of beta diversity in Europe: the role of climate, land cover and distance across scales. Journal of Biogeography 39:1473–1486.

Lanzer, N. B., T. D. Lee, M. J. Ducey, and S. E. Eisenhaure. 2017. Sapling white pine (Pinus strobus L.) exhibits growth response following selective release from competition with glossy buckthorn (Frangula alnus P. Mill) and associated vegetation. Forest Ecology and Management 404:280–288.

Latimer, A. M., S. Wu, A. E. Gelfand, and J. A. Silander. 2006. Building statistical models to analyze species distributions. Ecological Applications 16:33–50.

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

Legendre, P. 1993. Spatial Autocorralation: Trouble or New Paradigm? Ecology 74:1659–1673.

Leininger, T. J., A. E. Gelfand, J. M. Allen, and J. A. Silander. 2013. Spatial Regression Modeling for Compositional Data With Many Zeros. Journal of Agricultural, Biological, and Environmental Statistics 18:314–334.

Lydakis, A., J. M. Allen, M. Petrik, and T. M. Szewczyk. 2018. Robust strategies for managing invasive plants. Page IJCAI Artificial Intelligence for Wildlife Conservation.

Manson, S., J. Schroeder, D. Van Riper, and S. Ruggles. 2017. IPUMS National Historical Geographic Information System: Version 12.0 [Database]. Minneapolis.

Manzoor, S. A., G. Griffiths, and M. Lukac. 2018. Species distribution model transferability and model grain size – finer may not always be better. Scientific Reports 8:7168.

Merow, C., S. T. Bois, J. M. Allen, Y. Xie, and J. A. Silander. 2017. Climate change both facilitates and inhibits invasive plant ranges in New England. Proceedings of the National Academy of Sciences 114:E3276–E3284.

Merow, C., J. P. Dahlgren, C. J. E. Metcalf, D. Z. Childs, M. E. K. Evans, E. Jongejans, S. Record, M. Rees, R. Salguero-Gómez, and S. M. Mcmahon. 2014a. Advancing population ecology with integral projection models: A practical guide. Methods in Ecology and Evolution 5:99–110.

Merow, C., N. LaFleur, J. A. Silander Jr., A. M. Wilson, and M. Rubega. 2011. Developing Dynamic Mechanistic Species Distribution Models: Predicting Bird-Mediated Spread of Invasive Plants across Northeastern North America. The American Naturalist 178:30–43.

Merow, C., A. M. Latimer, A. M. Wilson, S. M. Mcmahon, A. G. Rebelo, and J. A. Silander. 2014b. On using integral projection models to generate demographically driven predictions of species’ distributions: Development and validation using sparse data. Ecography 37:1167–1183.

Merow, C., M. J. Smith, and J. A. Silander. 2013. A practical guide to MaxEnt for modeling species’ distributions: What it does, and why inputs and settings matter. Ecography 36:1058–1069.

Munroe, D. K., and D. Müller. 2007. Issues in spatially explicit statistical land-use/cover change (LUCC) models: Examples from western Honduras and the Central Highlands of Vietnam. Land Use Policy 24:521–530.

Pasquarella, V. J., C. E. Holden, and C. E. Woodcock. 2018. Improved mapping of forest type using spectral-temporal Landsat features. Remote Sensing of Environment 210:193–207.

Pearson, R. G., and T. Dawson. 2003. Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? Global Ecology & Biogeography 12:361–371.

Phillips, S. J., and M. Dudík. 2008. Modeling of species distribution with Maxent: new extensions and a comprehensive evalutation. Ecography 31:161–175.

QGIS Development Team. 2017. QGIS Geographic Information System. Open Source Geospatial Foundation Project.

R Core Team. 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing.

Rahbek, C., and G. R. Graves. 2001. Multiscale assessment of patterns of avian species richness. Proceedings of the National Academy of Sciences of the United States of America 98:4534–9.

Redo, D. J., and A. C. Millington. 2011. A hybrid approach to mapping land-use modification and land-cover transition from MODIS time-series data: A case study from the Bolivian seasonal tropics. Remote Sensing of Environment 115:353–372.

Stan Development Team. 2017a. Stan modeling language: User’s guide and reference manual. Stan version 2.16.0.

Stan Development Team. 2017b. RStan: the R interface to Stan.

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

Sulla-Menashe, D., M. A. Friedl, O. N. Krankina, A. Baccini, C. E. Woodcock, A. Sibley, G. Sun, V. Kharuk, and V. Elsakov. 2011. Hierarchical mapping of Northern Eurasian land cover using MODIS data. Remote Sensing of Environment 115:392–403.

Turner, M. G., R. V O’Neill, R. H. Gardner, and B. T. Milne. 1989. Effects of changing spatial scale on the analysis of landscape pattern. Landscape Ecology 3:153–162.

U.S. Census Bureau. 2017. 2017 TIGER/Line Shapefiles Technical Documentation.

U.S. Geological Survey. 2014. National Transportation Dataset. https://nationalmap.gov/transport.html.

Warren, D. L., M. Cardillo, D. F. Rosauer, and D. I. Bolnick. 2014. Mistaking geography for biology: inferring processes from species distributions. Trends in Ecology and Evolution 29:572–580.

Webster, C. R., M. A. Jenkins, and S. Jose. 2006. Woody invaders and the challenges they pose to forest ecosystems in the eastern United States. Journal of Forestry 104:366–374.

Wickham, J. D., S. V. Stehman, J. A. Fry, J. H. Smith, and C. G. Homer. 2010. Thematic accuracy of the NLCD 2001 land cover for the conterminous United States. Remote Sensing of Environment 114:1286–1296.

Wiens, J. A. 1989. Spatial Scaling in Ecology Spatial scaling in ecology. Source: Functional Ecology 3:385–397.

Wikle, C. 2003. Hierarchical Bayesian models for predicting the spread of ecological processes. Ecology 84:1382–1394.

# **TABLES**

**Table 1.** Description of land cover objects.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Object** | **Type** | **Description** | **Compositinal constraints** | **Extent** | **Number of land cover categories** |
| **Y** | Observed | higher accuracy | Yes | subset | *D* |
| **Z** | Observed | lower accuracy | Yes | full | *d=D*-1 |
| **Z’** | Modeled | **Z** with bias correction,  ***Zd-1*** split into ***Z’d-1*** , ***Z’d*** | No | full | *D* |
| **ν** | Modeled | latent truth | No | subset | *D* |
| **η** | Modeled | latent truth | Yes | full | *D* |
| **Zs** | Calculated | ***Zd-1*** split into ***Zsd-1*** , ***Zsd*** | Yes | full | *D* |

**Table 2**. Root mean squared error (RMSE) for the variable selection testing subset. The posterior mean, , from the optimal models predicted **Y** better for nearly every land cover category than did **Zs**. Note that the training subset at 9 km2 resolution consisted of 85% of cells, but only 15% at 20 ac resolution. Bolded percent differences indicate improvements of compared to **Zs**.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Resolution** | **Spatial random effects** | **Land cover category** | **RMSE  (Zs : Y)** | **RMSE ( : Y)** | **Percent difference** |
| 9 km2 | CAR | *Open Canopy* | 0.0716 | 0.0287 | **-59.9** |
| 9 km2 | CAR | *Deciduous Forest* | 0.1248 | 0.0606 | **-51.5** |
| 9 km2 | CAR | *White Pine Forest* | 0.0603 | 0.0286 | **-52.5** |
| 9 km2 | CAR | *Other Evergreen Forest* | 0.0852 | 0.0295 | **-65.4** |
| 9 km2 | CAR | *Mixed Forest* | 0.1744 | 0.0686 | **-60.6** |
| 9 km2 | CAR | *Other* | 0.0584 | 0.0358 | **-38.7** |
| 9 km2 | None | *Open Canopy* | 0.0716 | 0.0697 | **-2.6** |
| 9 km2 | None | *Deciduous Forest* | 0.1248 | 0.1113 | **-10.8** |
| 9 km2 | None | *White Pine Forest* | 0.0603 | 0.0416 | **-31.0** |
| 9 km2 | None | *Other Evergreen Forest* | 0.0852 | 0.0461 | **-45.9** |
| 9 km2 | None | *Mixed Forest* | 0.1744 | 0.1430 | **-18.0** |
| 9 km2 | None | *Other* | 0.0584 | 0.0594 | 1.7 |
| 20 ac | None | *Open Canopy* | 0.1492 | 0.1480 | **-0.8** |
| 20 ac | None | *Deciduous Forest* | 0.2283 | 0.2161 | **-5.4** |
| 20 ac | None | *White Pine Forest* | 0.1181 | 0.1233 | 4.4 |
| 20 ac | None | *Other Evergreen Forest* | 0.1484 | 0.1146 | **-22.8** |
| 20 ac | None | *Mixed Forest* | 0.2863 | 0.2770 | **-3.3** |
| 20 ac | None | *Other* | 0.1507 | 0.1515 | 0.5 |

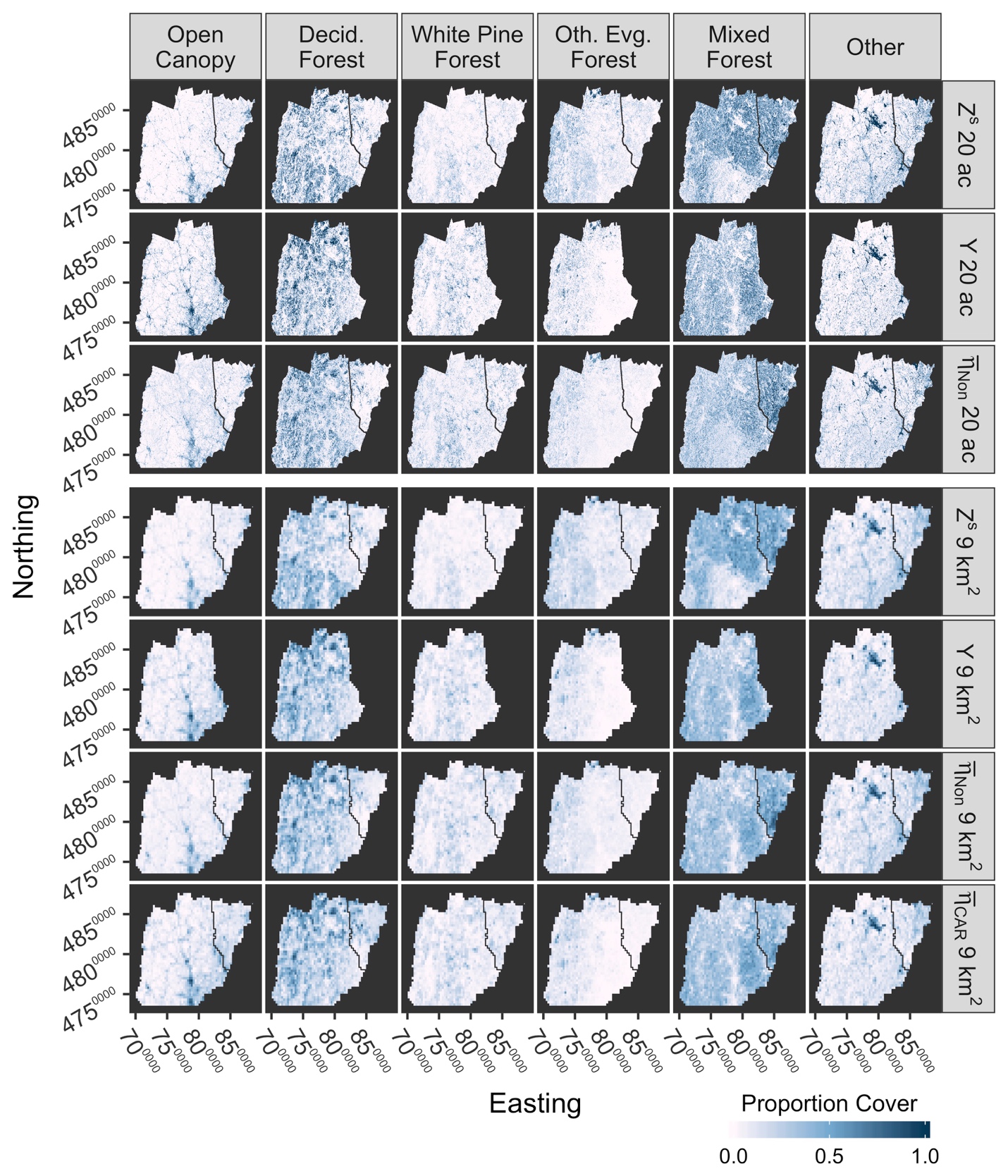
# **FIGURE LEGENDS**

**Figure 1.** Land cover proportions. Columns show aggregate land cover categories, and rows show the observed dataset **Y**, simply processed dataset **Zs**, and posterior means for the non-spatial and spatially explicit CAR models at fine (20 ac) and coarse (9 km2) resolutions.

**Figure 2.** Residual error in land cover proportions using **Y** as reference. Residuals are only shown for the fitting region in the full extent model, as **Y** is unavailable for the predicted region. Overestimates of **Y** are shown in red, with underestimates in blue. The posterior means show lower residual error than the simply processed **Zs**, and with less pronounced spatial patterning. The spatially explicit CAR model decreases residual error more than the non-spatial models, and the model performs better at the coarse (9 km2) resolution than the fine (20 ac) resolution.

**Figure 3.** Density plots of the absolute difference between the posterior mean and **Zs** for the full extent models. The difference between and **Zs** represents the bias estimated by each model. The distributions are expected to be similar in the fitting (purple) and prediction (green) regions. For the spatially explicit CAR model, distributions are comparable for most land cover categories. The non-spatial models are more variable, with posterior means more similar to **Zs** in the prediction region than in the fitting region, particularly for the fine resolution (20 ac) model.

**Figure 4.** Boxplots of 95% credible intervals in the fitting vs. prediction region for the full extent models. The spatially explicit CAR model and non-spatial models show different patterns of uncertainty in land cover proportion estimates. At both the coarse (9 km2) and fine (20 ac) resolutions, the non-spatial model has higher uncertainty in the fitted region (purple) and lower uncertainty in the prediction region (green). In contrast, the CAR model has very low uncertainty in the fitting region and high uncertainty in the prediction region.

Figure 1.

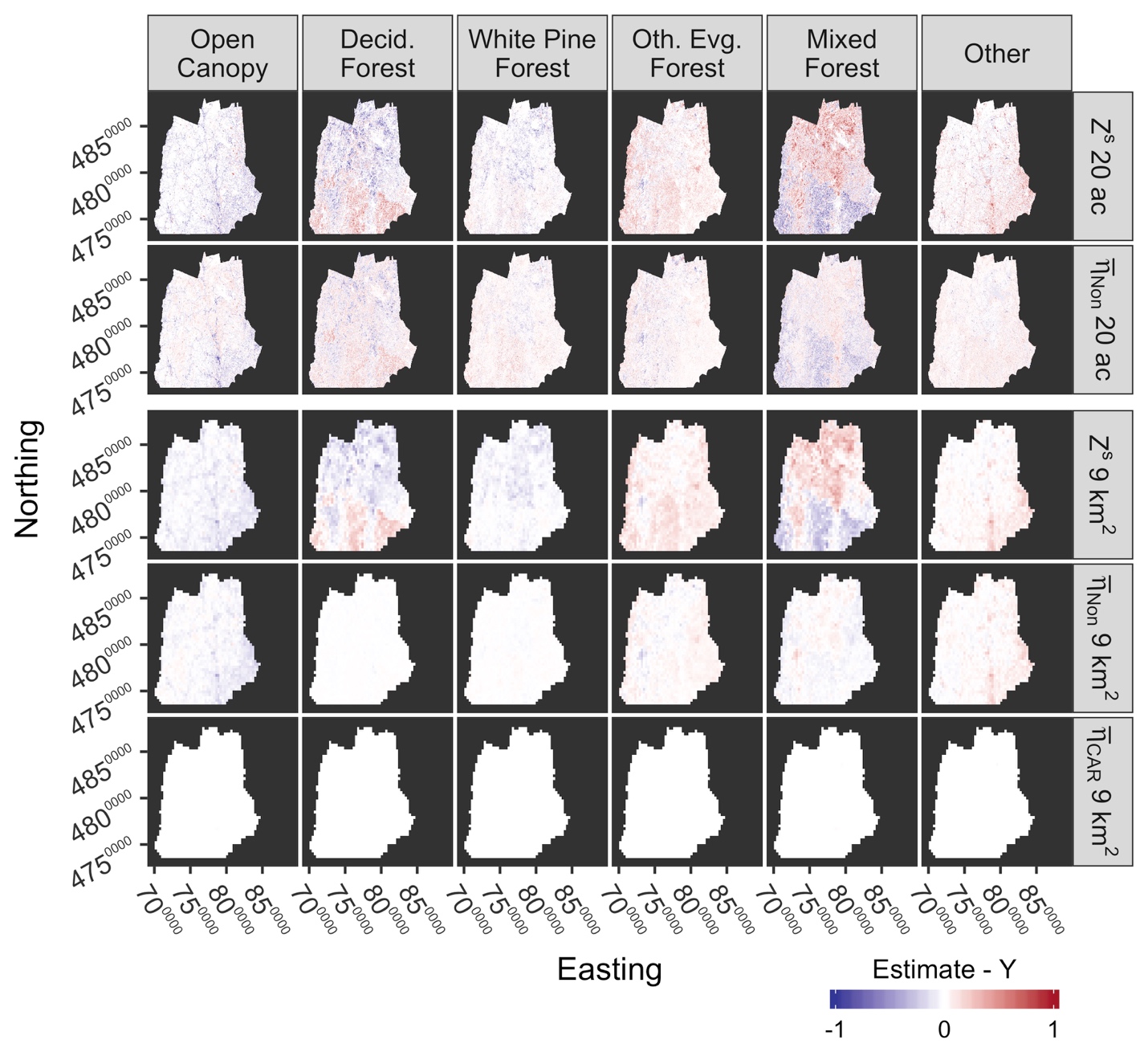
Figure 2.

Figure 3.

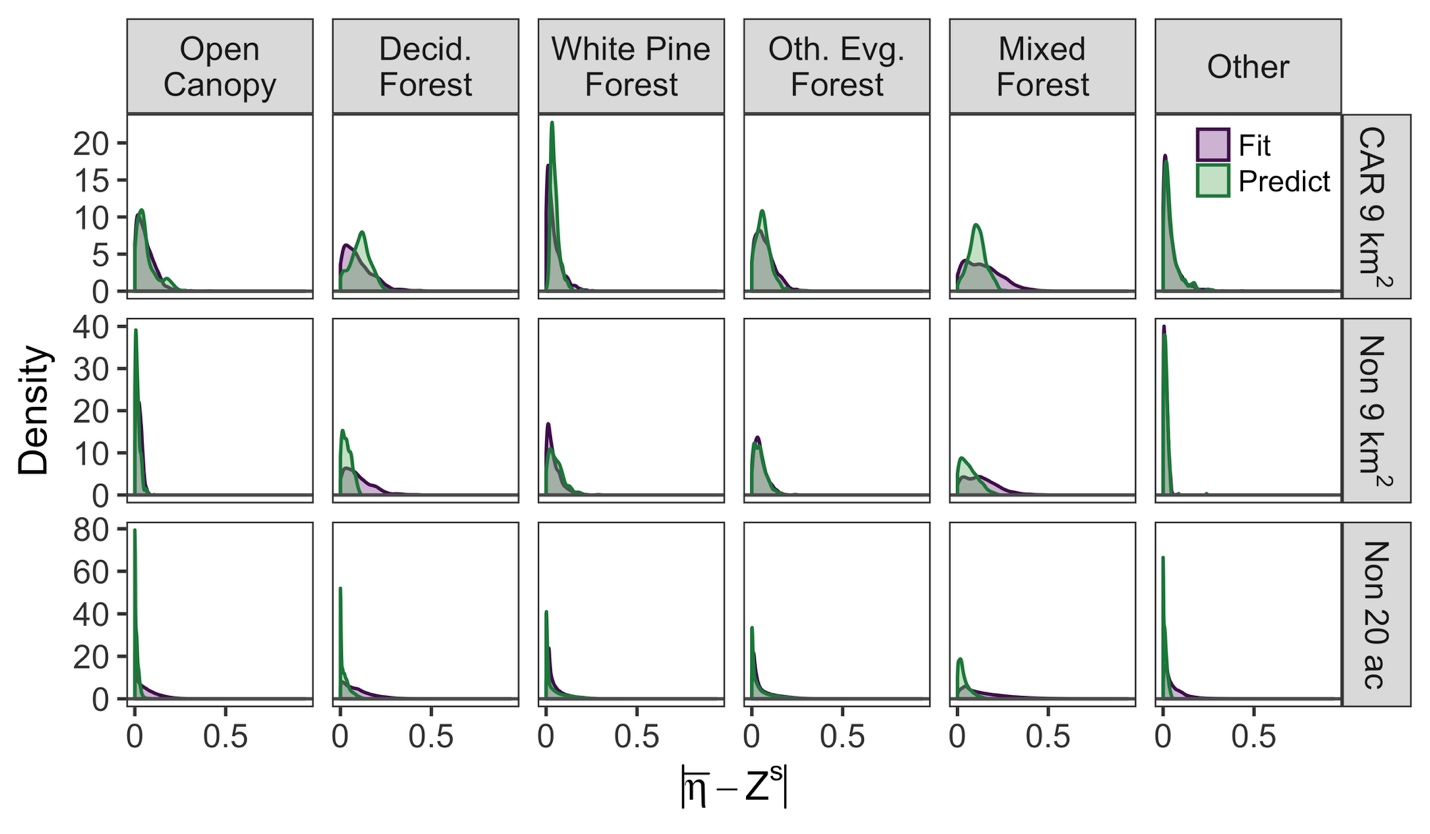
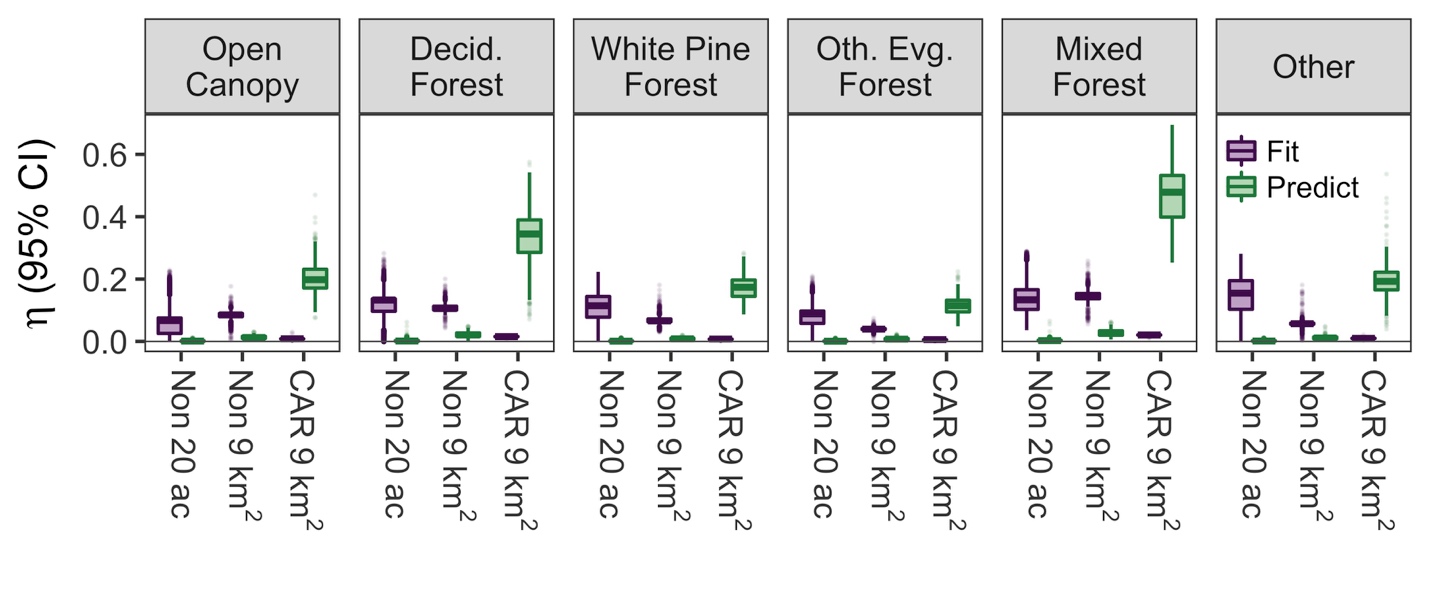


Figure 4.



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

Tim M. Szewczyk, Mark J. Ducey, Valerie Pasquarella, Jenica M. Allen

Ecological Applications

**Appendix S1.** Stan code for the full extent model with conditional autoregressive spatial random effects.

functions {

//sparse CAR prior for spatial random effects

real sparse\_car\_lpdf(vector phi, real tau, real alpha,

int[,] W\_sp, vector dist\_sp, vector lambda, int n, int W\_n) {

row\_vector[n] phit\_dist; // phi' \* dist

row\_vector[n] phit\_W; // phi' \* W

vector[n] ldet\_terms;

phit\_dist = (phi .\* dist\_sp)';

phit\_W = rep\_row\_vector(0, n);

for (i in 1:W\_n) {

phit\_W[W\_sp[i, 1]] = phit\_W[W\_sp[i, 1]] + phi[W\_sp[i, 2]];

phit\_W[W\_sp[i, 2]] = phit\_W[W\_sp[i, 2]] + phi[W\_sp[i, 1]];

}

for (i in 1:n) ldet\_terms[i] = log1m(alpha \* lambda[i]);

return 0.5 \* (n \* log(tau)

+ sum(ldet\_terms)

- tau \* (phit\_dist \* phi - alpha \* (phit\_W \* phi)));

}

//GJAM link function to enforce compositional restrictions

vector tr\_gjam\_inv(vector nu\_i, int D, int d) {

vector[D] eta\_i;

vector[d] nu\_i\_p\_;

real nu\_i\_p;

real T\_i;

eta\_i[1:d] = nu\_i;

for(j in 1:d) {

if(eta\_i[j] < 0) eta\_i[j] = 0;

if(eta\_i[j] < 1) nu\_i\_p\_[j] = eta\_i[j];

else nu\_i\_p\_[j] = 1;

}

nu\_i\_p = sum(nu\_i\_p\_);

if(nu\_i\_p >= 0.99) {

T\_i = (nu\_i\_p^(-1)) \* (1 - (0.01)^(nu\_i\_p/0.99));

while(sum(eta\_i[1:d]) > 0.99) {

vector[5] tmp;

tmp = T\_i \* eta\_i[1:d];

eta\_i[1:d] = tmp;

}

}

eta\_i[D] = 1 - sum(eta\_i[1:5]);

return eta\_i;

}

}

data {

int n1; //number of cells for Y

int n2; //n1 + 1 (for indexing)

int n3; //number of cells for Z + covariates

int D; //number of land cover classes

int n\_b; //number of covariates (beta) for rho for each LC

int ri[2\*(D-2)]; //beta indexes identifying which LC

int n\_t; //number of covariates (theta) for p

vector<lower=0, upper=1>[D-1] Y[n1]; //Y proportions

matrix<lower=0, upper=1>[n3,D-2] Z; //Z proportions

matrix[n1,n\_t] X; //covariates: fitting; [,1] = WhitePineDistribution

matrix[n3-n1,n\_t] X\_new; //covariates: prediction

matrix<lower=0, upper=1>[n3,n3] W; //adjacency matrix

int W\_n; // number of adjacent pairs

}

transformed data {

int d = D-1;

int tot\_b = n\_b\*(d-1); //total number of betas

//QR transformation for X

real qr\_n1 = n1-1; //avoids cmdstan ambiguity with sqrt(int)

matrix[n1,n\_t] Q = qr\_Q(X)[,1:n\_t] \* sqrt(qr\_n1);

matrix[n\_t,n\_t] R = qr\_R(X)[1:n\_t,] / sqrt(qr\_n1);

matrix[n\_t,n\_t] R\_inv = inverse(R);

//QR transformation for X\_new

real qr\_n3 = (n3-n1)-1; //avoids cmdstan ambiguity with sqrt(int)

matrix[n3-n1,n\_t] Q\_new = qr\_Q(X\_new)[,1:n\_t] \* sqrt(qr\_n3);

matrix[n\_t,n\_t] R\_new = qr\_R(X\_new)[1:n\_t,] / sqrt(qr\_n3);

matrix[n\_t,n\_t] R\_new\_inv = inverse(R\_new);

//CAR

int W\_sp[W\_n, 2]; // sparse representation of djacency pairs

vector[n3] dist\_sp; // diagonal of D (number of neigbors for each site)

vector[n3] lambda; // eigenvalues of invsqrtD \* W \* invsqrtD

{

int counter = 1;

for (i in 1:(n3 - 1)) {// identify neighbor pairs with upper triangle of W

for (j in (i + 1):n3) {

if (W[i, j] == 1) {

W\_sp[counter, 1] = i;

W\_sp[counter, 2] = j;

counter = counter + 1;

}

}

}

}

for (i in 1:n3) dist\_sp[i] = sum(W[i]);

{

vector[n3] invsqrtD;

for (i in 1:n3) invsqrtD[i] = 1 / sqrt(dist\_sp[i]);

lambda = eigenvalues\_sym(quad\_form(W, diag\_matrix(invsqrtD)));

}

}

parameters {

cholesky\_factor\_corr[d] L\_Omega[2]; //covariance

vector<lower=0>[d] L\_sigma[2]; //covariance

vector[d] nu[n1]; //landcover: latent, unconstrained

vector[n\_t] theta\_qr; //thetas (QR decomposition)

vector[tot\_b] beta\_z; //betas (QR decomposition, non-centered)

real<lower=0> beta\_scale; //betas (QR decomposition, non-centered)

matrix[n3, d] phi;

real<lower=0> tau[d];

real<lower=0, upper=1> alpha[d];

}

transformed parameters {

matrix[d,d] Sigma[2]; //covariance

vector[d] Z\_[n1]; //Z split & unbiased

vector[tot\_b] beta\_qr = beta\_z \* beta\_scale; //implies beta ~ N(0,scale)

for(k in 1:2) Sigma[k] = diag\_pre\_multiply(L\_sigma[k], L\_Omega[k]);

{

vector[n1] p = inv\_logit(Q[1:n1,] \* theta\_qr);

matrix[n1,d-1] rho;

for(j in 1:(d-2)) {

rho[,j] = Q[1:n1,2:n\_t] \* beta\_qr[ri[j+j-1]:ri[j+j]];

Z\_[,j] = to\_array\_1d(Z[1:n1,j] + rho[,j] + phi[1:n1,j]);

}

rho[,d-1] = Q[,2:n\_t] \* beta\_qr[ri[d+d-3]:ri[d+d-2]];

Z\_[,d-1] = to\_array\_1d((Z[1:n1,d-1] + rho[,d-1]) .\* p + phi[1:n1,d-1]);

Z\_[,d] = to\_array\_1d((Z[1:n1,d-1] + rho[,d-1]) .\* (1-p) + phi[1:n1,d]);

}

}

model {

for(k in 1:2) {

L\_Omega[k] ~ lkj\_corr\_cholesky(8);

L\_sigma[k] ~ normal(0, 1);

}

for(j in 1:d) {

phi[,j] ~ sparse\_car(tau[j], alpha[j], W\_sp, dist\_sp, lambda, n3, W\_n);

nu[,j] ~ normal(0.5, 1); //will be near observations [0,1]

}

tau ~ gamma(2,2);

theta\_qr ~ normal(0, 1);

beta\_z ~ normal(0, 1);

beta\_scale ~ normal(0, 1);

Y ~ multi\_normal\_cholesky(nu, Sigma[1]);

Z\_ ~ multi\_normal\_cholesky(nu, Sigma[2]);

}

generated quantities {

vector[n1] log\_lik; //log likelihood for model comparison

simplex[D] eta[n3]; //gjam transformed nu

vector[d] Z\_new\_[n3-n1]; //unbiased, split Z

vector[n\_t] theta = R\_inv \* theta\_qr;

vector[tot\_b] beta;

for(j in 1:(d-1)) {

beta[ri[j+j-1]:ri[j+j]] = R\_inv[2:n\_t,2:n\_t] \* beta\_qr[ri[j+j-1]:ri[j+j]];

}

{

vector[n3-n1] p\_new = inv\_logit(Q\_new \* theta\_qr);

matrix[n3-n1,d-1] rho\_new;

for(j in 1:(d-2)) {

rho\_new[,j] = Q\_new[,2:n\_t] \* beta\_qr[ri[j+j-1]:ri[j+j]];

Z\_new\_[,j] = to\_array\_1d(Z[n2:n3,j] + rho\_new[,j] + phi[n2:n3,j]);

}

rho\_new[,d-1] = Q\_new[,2:n\_t] \* beta\_qr[ri[d+d-3]:ri[d+d-2]];

Z\_new\_[,d-1] = to\_array\_1d((Z[n2:n3,d-1] + rho\_new[,d-1]) .\* p\_new

+ phi[n2:n3,d-1]);

Z\_new\_[,d] = to\_array\_1d((Z[n2:n3,d-1] + rho\_new[,d-1]) .\* (1-p\_new)

+ phi[n2:n3,d]);

}

for(n in 1:n1) {

eta[n] = tr\_gjam\_inv(nu[n], D, d);

log\_lik[n] = multi\_normal\_cholesky\_lpdf(Y[n] | nu[n], Sigma[1]) +

multi\_normal\_cholesky\_lpdf(Z\_[n] | nu[n], Sigma[2]);

}

for(n in n2:n3) eta[n] = tr\_gjam\_inv(Z\_new\_[n-n1], D, d);

}

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

Tim M. Szewczyk, Mark J. Ducey, Valerie Pasquarella, Jenica M. Allen

Ecological Applications

**Appendix S2.** Table of prior distributions

Prior distributions for parameters were set to be uninformative, but to constrain distributions to a reasonable range as suggested for implementation in Stan (Stan Development Team 2017). To increase model efficiency, we used three reparameterizations:

1. We used the Cholesky factor of each covariance matrix **Σ** with a Cholesky LKJ correlation distribution: **Σ** = diag\_matrix(σ) Ω diag\_matrix(σ), where σ is a vector of coefficient scales and Ω is a correlation matrix.
2. We used a QR reparameterization for the covariates.
3. We used a non-centered parameterization for the bias parameters ***β***, such that ***β*** = ***βstd\_Norm***\* ***βscale***, which implies ***β*** ~ Normal(0, ***βscale***). This improved model behavior since the bias terms are on a natural scale, and therefore the magnitude is small. Similarly, because **ν** is expected to predominately fall between 0–1 (Clark et al. 2017), we centered the prior distribution about 0.5.

|  |
| --- |
| **Prior distribution** |
| **ν** ~ Normal(0.5, 1) |
| Ω ~ lkj\_corr\_cholesky(8) |
| σ ~ Normal(0,1) |
| ***βstd\_Norm*** ~ Normal(0, 1) |
| ***βscale*** ~ Normal(0, 1) |
| ***θ*** ~ Normal(0, 1) |
| ***α*** ~ Uniform(0, 1) |
| ***τ*** ~ Gamma(2,2) |

References:

Clark, J. S., D. Nemergut, B. Seyednasrollah, P. J. Turner, and S. Zhang. 2017. Generalized joint attribute modeling for biodiversity analysis: Median-zero, multivariate, multifarious data. Ecological Monographs 87:34–56.

Stan Development Team. 2017. Stan modeling language: User’s guide and reference manual. Stan version 2.16.0. https://mc-stan.org.

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

Tim M. Szewczyk, Mark J. Ducey, Valerie Pasquarella, Jenica M. Allen

Ecological Applications

**Appendix S3.** Land cover dataset information and comparisons.

**Table S1**. Aggregation of original categories in **Y**.

|  |  |
| --- | --- |
| **Original Category** | **Aggregate Category** |
| Resid./Comm./Indus. | Open |
| Transportation | Open |
| Row Crops | Other |
| Hay/Pasture | Other |
| Orchards | Other |
| Beech/Oak | Deciduous Forest |
| Paper Birch/ Aspen | Deciduous Forest |
| Other Hardwood | Deciduous Forest |
| White/Red Pine | White Pine Forest |
| Spruce/Fir | Other Evergreen Forest |
| Hemlock | Other Evergreen Forest |
| Pitch Pine | Other Evergreen Forest |
| Mixed Forest | Mixed Forest |
| Alpine (Krumholz) | Other |
| Water | Other |
| Forested Wetland | Mixed Forest |
| Open Wetland | Other |
| Tidal Wetland | Other |
| Disturbed | Other |
| Bedrock/ Veg. | Other |
| Sand Dunes | Other |
| Other Cleared | Open |
| Tundra | Other |

**Table S2**. Aggregation of original categories in **Z**.

|  |  |
| --- | --- |
| **Original Category** | **Aggregate Category** |
| Water | Other |
| Perennial Ice/Snow | Other |
| Developed, Open Space | Other |
| Developed, Low Intensity | Open Canopy |
| Developed, Medium Intensity | Open Canopy |
| Developed, High Intensity | Other |
| Barren Land (Rock/Sand/Clay) | Other |
| Deciduous Forest | Deciduous Forest |
| Evergreen Forest | Evergreen Forest |
| Mixed Forest | Mixed Forest |
| Dwarf Scrub | Other |
| Shrub/Scrub | Other |
| Grassland/Herbaceous | Other |
| Sedge/Herbaceous | Other |
| Lichens | Other |
| Moss | Other |
| Pasture/Hay | Other |
| Cultivated Crops | Other |
| Woody Wetlands | Mixed Forest |
| Emergent Herbaceous Wetlands | Other |

**Table S3.** Confusion matrix for aggregated categories in **Y**. The values are from the confusion matrix provided in Justice et al. (2002). Overall accuracy is calculated as sum(diag(.))/sum(.) to represent the total proportion of correct cells across all aggregated categories.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | **Ground Truth** | | | | | |  |
|  |  | Open Canopy | Other | Decid. Forest | White Pine Forest | Other Evg. Forest | Mixed Forest | Percent Correct |
| **Predicted** | Open Canopy | **131** | 7 | 2 | 0 | 0 | 0 | 93.6% |
| Other | 12 | **361** | 0 | 0 | 1 | 11 | 93.8% |
| Decid. Forest | 0 | 1 | **127** | 0 | 0 | 6 | 94.8% |
| White Pine Forest | 0 | 0 | 0 | **49** | 1 | 10 | 81.7% |
| Other Evg. Forest | 2 | 0 | 0 | 3 | **124** | 27 | 79.5% |
| Mixed Forest | 1 | 5 | 11 | 2 | 2 | **79** | 79.0% |
|  |  |  |  |  |  |  |  | **89.3%** |

**Table S4.** Confusion matrix for aggregated categories in **Z**. The values are aggregated from the Level II error matrix for Region 10 provided in Table S12 the Appendix in Wickham et al. (2010). Overall accuracy is calculated as sum(diag(.))/sum(.) to represent the total proportion of correct cells across all aggregated categories.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | **Ground Truth** | | | | |  |
|  |  | Open Canopy | Other | Decid. Forest | Evg. Forest | Mixed Forest | Percent Correct |
| **Predicted** | Open Canopy | **2.042** | 0.589 | 0.019 | 0.019 | 0.007 | 76.3% |
| Other | 0.479 | **22.318** | 1.25 | 0.395 | 1.217 | 87.0% |
| Decid. Forest | 0.097 | 2.059 | **28.466** | 2.836 | 1.038 | 82.5% |
| Evg. Forest | 0.058 | 0.709 | 0.418 | **12.781** | 0.616 | 87.6% |
| Mixed Forest | 0.007 | 1.993 | 1.245 | 2.156 | **17.183** | 76.1% |
|  |  |  |  |  |  |  | **82.8%** |

References:

Justice, D., A. Deely, and F. Rubin. 2002. New Hampshire land cover assesment.

Wickham, J. D., S. V. Stehman, J. A. Fry, J. H. Smith, and C. G. Homer. 2010. Thematic accuracy of the NLCD 2001 land cover for the conterminous United States. Remote Sensing of Environment 114:1286–1296.

**Figure S1.** Scatterplot of land cover proportions within the overlapping (i.e., fitting) region for **Zs** and compared to **Y**. Red dotted lines represent perfect correspondence with **Y** proportions.

