Incorporating canopy structure from simulated GEDI lidar into bird species distribution models

Patrick Burns1,7, Matthew Clark2, Leonardo Salas3, Steven Hancock4, David Leland5, Patrick Jantz1, Ralph Dubayah6 and Scott J Goetz1

1 School of Informatics, Computing and Cyber Systems, Northern Arizona University, Flagstaff, AZ, United States of America
2 Geography, Environment and Planning, Sonoma State University, Rohnert Park, CA, United States of America
3 Point Blue Conservation Science, Petaluma, CA, United States of America
4 School of Geosciences, The University of Edinburgh, Edinburgh, United Kingdom
5 Madrone Audubon Society, Santa Rosa, CA, United States of America
6 Department of Geographical Sciences, University of Maryland, College Park, MD, United States of America
7 Author to whom any correspondence should be addressed.
E-mail: Patrick.Burns@nau.edu

Keywords: bird species distribution models, GEDI, lidar, canopy structure, machine learning

Abstract
The Global Ecosystem Dynamics Investigation (GEDI) lidar began data acquisition from the International Space Station in March 2019 and is expected to make over 10 billion measurements of canopy structure and topography over two years. Previously, airborne lidar data with limited spatial coverage have been used to examine relationships between forest canopy structure and faunal diversity, most commonly bird species. GEDI’s latitudinal coverage will permit these types of analyses at larger spatial extents, over the majority of the Earth’s forests, and most importantly in areas where canopy structure is complex and/or poorly understood. In this regional study, we examined the impact that GEDI-derived Canopy Structure variables have on the performance of bird species distribution models (SDMs) in Sonoma County, California. We simulated GEDI waveforms for a two-year period and then interpolated derived Canopy Structure variables to three grid sizes of analysis. In addition to these variables, we also included Phenology, Climate, and other Auxiliary variables to predict the probability of occurrence of 25 common bird species. We used a weighted average ensemble of seven individual machine learning models to make predictions for each species and calculated variable importance. We found that Canopy Structure variables were, on average at our finest resolution of 250 m, the second most important group (32.5%) of predictor variables after Climate variables (35.3%). Canopy Structure variables were most important for predicting probability of occurrence of birds associated with Conifer forest habitat. Regarding spatial analysis scale, we found that finer-scale models more frequently performed better than coarser-scale models, and the importance of Canopy Structure variables was greater at finer spatial resolutions. Overall, GEDI Canopy Structure variables improved SDM performance for at least one spatial resolution for 19 of 25 species and thus show promise for improving models of bird species occurrence and mapping potential habitat.

1. Introduction
Due to the widespread impacts of land cover change, pollution, wildlife exploitation, and climate change on species and ecosystems across the world, biodiversity assessment and monitoring is imperative. The United Nations’ latest global assessment report on biodiversity and ecosystem services estimates that the number of species threatened from extinction now reaches an unprecedented 1 million, rates of extinction are accelerating, and current actions are insufficient to reverse these trends (IPBES 2019). Monitoring biodiversity across large areas through rapid, scalable, and accurate remote sensing methods will greatly improve our understanding of species distributions, loss and resilience, drivers of changes,
and impacts at multiple ecosystem levels (Pereira et al 2013, Corbane et al 2015, He et al 2015, Rocchini et al 2015). Proenca and colleagues (2017) reviewed scientific approaches to global biodiversity monitoring, with a focus on GEO BON’s Essential Biodiversity Variables, and highlight the use of remote sensing for characterizing ecosystem structure and function. In terrestrial applications, remote sensors provide measurements of vegetation properties, including chemistry, structure and phenology, which are related to species habitat requirements, specifically shelter and food resources. In this regard, many unexplored opportunities with new technologies remain open to investigation. For example, NASA’s Global Ecosystem Dynamics Investigation (GEDI; Dubayah et al 2020) is a light detection and ranging (lidar) sensor that was installed on the International Space Station (ISS) in December 2018. This sensor makes fine-scale measurements of 3D vegetation structure, such as overall canopy height and vertical distributions of plant material. The acquisition of these structural measurements over the majority of the land surface is expected to lead to significant advances in biodiversity monitoring and modeling (Bergen et al 2009, Goetz et al 2014, Bakx et al 2019). In this study, our primary goal is to examine the importance and added value of 3D canopy structure data derived from simulated GEDI lidar waveforms in regional-extent bird species distribution models (SDM).

Bird species have been well-monitored relative to other taxa (Troudet et al 2017), especially in North America, and recent work incorporating long-term surveys and radar remote sensing suggests a drastic decline of 2.9 billion birds since 1970 (Rosenberg et al 2019). Similar to other taxa, habitat loss is a major driver of reduction in bird species abundance and range (Jetz et al 2007). Suitable bird habitat includes necessities like food, water, shelter, and nesting sites. Habitats are characterized at various scales, from broad associations to more descriptive microhabitats. Habitat associations are typically related to land-cover classes, frequently broad vegetation types (e.g. coniferous forest, grassland), which can be extracted from existing national or global land-cover maps (e.g. MODIS C6; Sulla-Menashe et al 2019). Habitat associations can be further subdivided into microhabitats (e.g. shaded ground beneath mature canopy) which are characterized by fine-scale topography, climate, and vegetation structure. Measurements of vegetation structure are not widely available at fine scales, but previous work suggests that structural complexity is associated with a greater variety of microhabitats which can lead to higher diversity and abundance due to more opportunities for foraging, shelter, and nesting (Cody 1985, Hunter 1999, Whittaker et al 2001, Hill et al 2004).

Previous SDM applications have mainly used remote sensing to derive spatially-explicit environmental predictors, such as topography and bioclimatic conditions (Wilson and Jetz 2016), or land-cover classes (He et al 2015, Bradie and Leung 2017). When considering bird diversity, continental-extent distributions are largely modeled with bioclimatic variables (e.g. WorldClim; Fick and Hijmans 2017), which can predict overall physiological constraints, and correlate with broad vegetation patterns that influence habitat (Stralberg et al 2009, Lawler et al 2009). However, at regional extents and finer spatial resolution (grain), information on vegetation structure is useful for characterizing microhabitats, particularly in areas with heterogeneous vegetation patterns, for example from topographic variation (Stralberg et al 2009) or land-use/land-cover change (Sohl 2014). Bird species diversity is known to respond to both vegetation structure and floristic diversity (Macarthur and Macarthur 1961, Rotenberry 1985, Adams and Matthews 2019), both factors that have been missing in models operating beyond the plot scale. Recent developments in airborne remote sensing have provided missing links at regional to local extents, implying possibilities for broader-extent applications when these technologies are elevated to space.

Airborne lidar surveys of topography and vegetation structure are becoming more frequent, but are still limited in areal extent due to operational costs. While most airborne lidar sensors typically capture discrete returns from small footprints (<1 m diameter), airborne and spaceborne waveform sensors capture the full vertical profile of vegetation structure, typically using relatively larger footprints (>20 m diameter) collected over a wider area (Lim et al 2003; Anderson et al 2016; Hancock et al 2019). Lidar data are already improving our understanding of how vegetation structure influences bird species distributions at multiple spatial scales (Tattoni et al 2012, Rechsteiner et al 2017, Carrasco et al 2019) and several review papers have examined lidar-based studies that consider avian distributions (Vierling et al 2008, Davies and Asner 2014, Bakx et al 2019). In general, current research tends to use airborne lidar and derived metrics related to canopy height, vertical and horizontal variability, and cover, as well as understory density. Most studies report a positive effect of lidar metrics in explaining bird distributions and/or patterns of richness, however the extent and scale of analysis are important factors. Species sense and respond to a range of temporal and spatial scales (Wiens 1989, Levin 1992), and this has important implications for parameterizing and interpreting SDMs. (Mayor 2009) and McGarigal et al (2016) review how hundreds of studies address the issue of scale dependence, finding no consensus. At a minimum, they recommend testing SDMs at a variety of spatial scales, ideally optimizing the scale of each variable (multi-scale optimization).

In this study, we focus on implementing simulated GEDI lidar—the first spaceborne lidar mission
designed specifically to measure vegetation structure. The instrument is scheduled to collect data for a minimum of two years, from 2019 to 2020. Initial GEDI data were released on January 22, 2020, and so our analysis uses 2 years of GEDI data simulated from airborne lidar data. Previous research using NASA’s airborne full waveform Land Vegetation and Ice Sensor (LVIS), a precursor to GEDI, showed that lidar metrics were useful for modeling bird species richness, abundance, and habitat quality at local extents (Goetz et al. 2007, 2010). More recent research has used space-based ICESat-1 waveform lidar at broader scales (Goetz et al. 2014), corroborating other studies showing vegetation properties derived from multispectral satellites (e.g. MODIS), such as percent tree cover and life form type, along with bioclimatic variables were generally sufficient predictors of breeding bird species richness.

We created SDMs for 25 species of birds in Sonoma County (figure 1) using canopy structure predictor variables as well as additional climate, phenology, and other auxiliary geospatial predictors. Our three objectives are to: (1) generate bird SDMs and determine the importance of individual predictor variables and variable groups for 25 bird species; (2) determine whether including Canopy Structure variables improve SDM performance, while also exploring the impacts of spatial scale; and, (3) map probability of bird species occurrence and associated uncertainty of these predictions.

2. Methods

2.1. Spatial and temporal domain of SDMs

We focus on Sonoma County because there are a relatively large number of geolocated bird species observations and the entire county is covered by airborne lidar data (figure 1) which we used for the GEDI simulation. A previous effort, the Sonoma County Breeding Bird Atlas (U.S. Geological Survey 2019), was based entirely on field observations and mapped occurrence of bird species in Sonoma County at 5-km spatial resolution. We sought to create finer spatial resolution maps showing probability of occurrence by generating three regularly-spaced square grids covering the county at 250-, 500- and 1000 m spatial resolution.

The airborne lidar collection occurred from September to November 2013 and is the primary temporal constraint on our analysis. We chose water years (starting in October and ending in September) 2013 to 2015 to bracket the lidar data collection. Additional remote sensing datasets used for SDMs were temporally-aggregated to the entire three-year period. Due to the limited spatial coverage of species observations from this three-year period, we extended the species observation window to ten years (2006 to 2015).

2.2. Bird species data

2.2.1. Selecting species of interest

We gathered field observations from point count surveys collected by Point Blue Conservation Science (figure 1; appendix A) from the California Avian Data Center (http://data.prbo.org/cadc2/) between 2006 and 2015, along with data from the citizen science observation network eBird (Sullivan et al. 2009), and the Breeding Bird Survey (Sauer et al. 2017). eBird data comprised the majority of records. We used data from all detailed survey events that recorded all species present, and thus from which we could discern detection or non-detection. If presence detections occurred in a given grid cell at a particular resolution then that cell response was classified as presence (value = 1). Grid cell surveys in which the observer did not encounter the species of interest were classified as absence (value = 0). We selected the top 25 species based on number of grid cell presences at 250 m spatial resolution (see table 1). Main species habitat associations (Conifer, Oak, Shrub, Riparian, Grass, Urban, and Variable; table 1) were determined from the Sonoma County Breeding Bird Atlas, Audubon Guide to North American Birds (https://www.audubon.org/bird-guide) and The Cornell Lab of Ornithology Birds of North America guide (https://birdsna.org; Rodewald 2015). A variable habitat association means that the species does not spend a majority of time in a given habitat.

2.3. GEDI lidar and canopy structure predictor variables

2.3.1. GEDI Instrument Description

The GEDI lidar sensor measures vertical canopy structure of forests between ±51.6 degrees latitude. The instrument emits pulses of energy (1064-nm) which reflect off of various canopy layers and/or the Earth’s surface, and the returned energy of each pulse is recorded as a short duration time series. Using the speed of light, ISS orbital geometry, and precise positioning, this time series is geolocated on the Earth’s surface and transformed into a waveform—a profile of returned energy discretized into 15-cm vertical height bins for each ~22- to 25 m diameter footprint (Dubayah et al. 2020). Footprints are separated by approximately 60 m along each of the 8 ground tracks and by 600 m between tracks. GEDI, like other lidar systems, is not able to make measurements through clouds.

2.3.2. Canopy Structure predictor variables from simulated GEDI

The primary GEDI data product (L1B) is a geolocated lidar waveform. Variations of the main waveform and waveform-derived metrics (L2; Tang et al. 2012) are output for each footprint.

We simulated two years of GEDI lidar observations from airborne lidar data by using the methods
Figure 1. Sonoma County, California, USA dominant vegetation and land cover classes (a.) and vegetation height map (b.) derived from airborne lidar survey in 2013 (sonomavegmap.org). All species observation locations from 2006 to 2015 (blue dots in b.) and corresponding high density observation areas (orange outlines in b.) are shown as well.

Table 1. Species selected for distribution modeling and their habitat association.

| Species Code | Common Name                | Pres. Grid Cells at 250 m | Abs. Grid Cells at 250 m | Habitat Association |
|--------------|----------------------------|--------------------------|--------------------------|---------------------|
| ACWO         | Acorn Woodpecker           | 189                      | 718                      | Oak                 |
| AMGO         | American Goldfinch         | 114                      | 486                      | Grass               |
| BEWWR        | Bewick’s Wren              | 98                       | 615                      | Variable            |
| BHGR         | Black-headed Grosbeak      | 106                      | 517                      | Variable            |
| BLPH         | Black Phoebe               | 193                      | 478                      | Riparian            |
| BRBL         | Brewer’s Blackbird         | 172                      | 499                      | Urban               |
| BUSH         | Bushtit                    | 165                      | 579                      | Oak                 |
| CALT         | California Towhee          | 306                      | 438                      | Oak                 |
| CAQU         | California Quail           | 193                      | 520                      | Shrub               |
| CBCH         | Chestnut-backed Chickadee  | 127                      | 544                      | Conifer             |
| DEJU         | Dark-eyed Junco            | 155                      | 589                      | Conifer             |
| HOFI         | House Finch                | 261                      | 362                      | Urban               |
| LEGO         | Lesser Goldfinch           | 164                      | 459                      | Grass               |
| MODO         | Mourning Dove              | 281                      | 741                      | Urban               |
| NOFL         | Northern Flicker           | 89                       | 624                      | Conifer             |
| NOMO         | Northern Mockingbird       | 214                      | 808                      | Oak                 |
| NUWO         | Nuttall’s Woodpecker       | 152                      | 471                      | Oak                 |
| OATI         | Oak Titmouse               | 181                      | 490                      | Oak                 |
| RWBL         | Red-winged Blackbird       | 175                      | 448                      | Riparian            |
| SOSP         | Song Sparrow               | 213                      | 610                      | Riparian            |
| SPTO         | Spotted Towhee             | 137                      | 486                      | Conifer             |
| STTA         | Steller’s Jay              | 119                      | 625                      | Conifer             |
| WSCP         | White-crowned Sparrow      | 168                      | 432                      | Shrub               |
| WEBL         | Western Bluebird           | 132                      | 491                      | Grass               |
| WESJ         | Western Scrub-jay          | 257                      | 414                      | Oak                 |

|                | Mean                       | 174                      | 538                      |                     |
|                | Std. Deviation             | 57                       | 108                      |                     |

described by (Hancock et al 2019). All simulated L2 variables are based on a gaussian ground-finding algorithm (Hofon et al 2000). The ‘Canopy Structure’ predictor variables used in this study are described in appendix Table C1. Since GEDI is a sampling instrument we used the ‘Moving Window Kriging’ tool from ArcGIS Pro to produce predicted Canopy Structure metric values at 50 m intervals. We generated the geostatistical model for each Canopy Structure metric using the ArcGIS Pro Geostatistical Wizard and the ‘Simple’ Kriging method. We resampled the 50 m prediction raster to our three different analysis scales using an averaging method. Additional details regarding airborne lidar data processing, GEDI simulation processing and interpolation can be found in appendix C.a.
2.4. Additional predictor variables

2.4.1. Climate.
We selected six ‘Climate’ predictor variables from the California Basin Characterization Model (BCM; Flint et al. 2013) to include in SDM: precipitation (Ppt), average minimum temperature (TMn), average maximum temperature (TMx), potential evapotranspiration (PET), actual evapotranspiration (AET), and climatic water deficit (CWD). The monthly data for water years 2013, 2014 and 2015 were resampled from a native spatial resolution of 270 m to our three analytical spatial scales using an averaging method. Variables for each water year were then grouped and averaged into four three-month periods: October-December (Q1), January-March (Q2), April-June (Q3), and July-September (Q4). Final SDM climate variables were averages of the three water years (2013–2015) for each three-month BCM variable, resulting in 24 total ‘Climate’ variables for each spatial analysis scale.

2.4.2. Auxiliary variables
We used four ‘Auxiliary’ predictor variables as inputs to SDMs: elevation, distance to ocean coast, distance to streets, and distance to streams. Sonoma County has a diverse topography, ranging from Pacific coastlines to mountains greater than 1000 m in elevation. Elevation can be a proximate determinant of a species’ presence (Hof et al. 2012). Elevation rasters were derived from the USGS National Elevation Dataset (NED; 1/3 arcsecond native spatial resolution). In Sonoma County, distance to coast is correlated with a longitudinal pattern of habitats. Closer to the coast are mixed hardwood and conifer forests, such as Coastal redwoods and Douglas fir, with interspersed patches of chaparral, and further inland occur conifer patches, oak woodlands, chaparral and grasslands. Distance to coast was also calculated initially as a 30 m raster from a coastal vector layer using ArcGIS. Distance to streets was incorporated as a proxy gradient for human influence. We used the Sonoma County streets vector layer to calculate distances in a 30 m raster. Distances to streams were incorporated as an additional predictor related to moisture and water availability. Numerous species of birds prefer habitats along or near streams (Rottenborn 1999, Mcclure et al. 2015). We used the lidar-derived streams vector layer from the Sonoma County Vegetation Mapping and Lidar program (https://sonomavegmap.org) to calculate distances in a 30 m raster. All auxiliary variables were resampled to the spatial analysis scales (250-, 500-, and 1000 m) using an averaging method.

2.4.3. Phenology
Multispectral satellites like the Moderate Resolution Imaging Spectroradiometer (MODIS) are particularly useful for monitoring vegetation presence, vitality, and phenology. We calculated Dynamic Habitat Indices (DHI) Phenological predictor variables from MODIS, most recently described by Hobi et al. (2017) and Radeloff et al. (2019). These indices are measures of vegetation productivity over the course of a year. We were limited to using MOD13 NDVI since it is the only index measured at (approximately) our finest analysis scale. The DHIs we calculated include: (1) DHI sum—the area under the phenological curve of a year, (2) DHI min—the minimum value of the phenological curve of a year, (3) DHI var—the coefficient of variation of the phenological curve of a year, (4) DHI median—the median value of the composite time series, (5) DHI 95p—the 95th percentile value of the composite time series, and (6) DHI seasonal difference—the difference between mean June and mean November NDVI, a potential proxy for deciduousness in this area. The six DHI metrics were calculated at their nominal spatial resolution (232-, 463-, 927 m) and then resampled to our analysis scales. See appendix C.c for additional details.

2.5. Species distribution model approach

2.5.1. Individual SDMs
We combined Climate (24), Canopy Structure (20), Phenology (7), and Auxiliary (4) variables for a total of 55 predictors covering Sonoma County at each spatial resolution. We rescaled each predictor variable by subtracting the mean and dividing by the standard deviation. To account for potential multicollinearity we used a variance inflation factor (VIF) analysis to remove the most correlated variables. We used our highest spatial scale scenario (250 m) variable values and set a VIF threshold value greater than or equal to 10—equivalent to excluding variables that have an R-square of 0.9 or higher when regressed against all other variables. We selected the same variables for the other two model scales that use Canopy Structure predictors. This model scenario is referred to as ‘All with Canopy Structure’. For comparison we ran a scenario referred to as ‘All without Canopy Structure’ which excluded Canopy Structure variables.

We ran seven individual SDMs (Random Forests, Support Vector Machine, Boosting, Extreme Gradient Boosting, Neural Network, Net Regularized Generalized Linear Models, and K-Nearest Neighbors) 500 times (bootstraps) for each species, resolution, and model scenario combination. Each bootstrap corresponded to a random sample (80% training, 20% testing) of spatially-thinned species observations (Aiello-Lammens 2015) and associated predictor variables. Cross-validation was used to tune models and prevent overfitting. Additional details related to individual model tuning and bootstrapping can be found in appendix D. We focused on the area under the receiver operator characteristic curve (AUC; Fielding and Bell 1997) for individual model evaluation. AUC values equal to 0.5 mean that the predicted values are equivalent to random guesses. AUC
values greater than 0.5 and up to 1 (perfect) indicate that the model has some level of predictive power.

2.5.2. Variable importance
We used a sensitivity analysis method from the R package rminer (Cortez and Embrechts 2013, Cortez 2016) to assess variable importance. We selected the data-based sensitivity analysis (DSA) method and associated average absolute deviation from the median (AAD) for measuring importance. We focused on the model scenario ‘All with Canopy Structure’ and assessed variable importance for species both in terms of individual variables and variable groups (Climate, Canopy Structure, Phenology, and Auxiliary). We also assessed variable importance aggregated by habitat association.

2.5.3. Ensemble SDMs
We generated two different types of ensembles, an individual iteration ensemble (IIE) and combined iteration ensemble (CIE). The IIE for each species calculates a weighted average prediction (Marmion et al 2009) value from up to seven individual model predictions for a single bootstrap (figure 2). To create this ensemble we only selected individual models with AUC greater than 0.5 and then calculated a weighted average prediction for each grid cell, where the weights were based on adjusted AUC (AUC score minus 0.5). All prediction averaging was done following a logit transform of the model predictions. Averaged values were then transformed back to probability (0 to 1). Our second ensemble method, CIE, aggregates species model predictions from individual models across all 500 bootstraps, again weighting the final prediction value by adjusted AUC. The maximum number of individual prediction values was 3500 (7 models * 500 iterations), assuming all individual models had AUC greater than 0.5. We used the weighted average prediction value and associated weighted standard error to display probability of occurrence and associated uncertainty for each species (figure 2). All SDM analyses were performed in R (R Core Team 2019). Appendix H lists specific R packages used and associated references.

2.6. Impact of Canopy Structure variables on model performance
2.6.1. Ensemble SDMs
Differences in predictive accuracy when contrasting model constructs can be highly informative with respect to the importance of a particular predictor for understanding the niche of a species, and for management and decision-making (Austin and Van Niel 2011 Mod et al 2016, Bell and Schlaepfer 2016). However, consistency in identification of important environmental variables varies with the algorithm used and other factors, so caution must be taken to ensure models do not over-fit when tested with and without a variable, as they may include variable interactions that increase model complexity and may not correctly reflect the ecology of the species (Aguirre-Gutiérrez et al 2013, Bell and Schlaepfer 2016). This was noted by (Araújo and Guisan 2006), who recommend the use of carefully chosen predictors (see also Araújo et al 2019). We postulated that GEDI variables provide ecologically-meaningful information about vegetation structural complexity that partly determines the niche of several of our study species. We visually compared the performance of SDMs that included and omitted Canopy Structure variables. We expected the algorithms used would adjust model complexity to optimally predict with and without the Canopy Structure variables. However, we also expected Canopy Structure variable importance would still be evident as an increase in performance when they are included in the model.

Though information from the full range of the species has been noted as very important for SDM performance (Kadmon et al 2003, Syphard and Franklin 2010, Martínez-Freiría et al 2016; see review in Engler et al 2017), the geographic coverage of the airborne lidar dataset precluded us from running models outside Sonoma County. Relatively less important than geographic bias and sample size (Thibaud et al 2014), spatial autocorrelation will also affect SDM performance and was not included in our models. Since our goal was to evaluate the relative importance of the Canopy Structure variables in SDMs, we focused on building relatively accurate models and comparing among constructs that vary only with respect to the variables used.

2.6.2. Logistic regression analyses
In order to provide additional evidence of the impact of Canopy Structure variables, we used all the variables remaining after VIF to construct an additive logistic regression model (the ‘All with Canopy Structure’ model scenario), and also fitted the ‘All without Canopy Structure’ model scenario. For each bootstrap sample, we then calculated the likelihood ratio test statistic of the full model vs the restricted model (i.e. with vs without Canopy Structure), and report the range of values of the statistic in relation to the value at which it is expected to occur fewer than 5% of the times (the ‘statistically significant’ departure value). Results of this logistic regression analysis are provided for comparison in appendix E.c. since our main focus is the use of Canopy Structure variables in ensemble SDMs.

3. Results
3.1. Predictor variables and importance
The VIF method used to remove highly-correlated predictor variables at 250 m spatial resolution
reduced the number of predictor variables from 55 to 23. This reduction left 38% of Climate, 35% of Canopy Structure, 100% of Auxiliary and 43% of Phenology predictor variables. The following Climate variables remained: Actual Evapotranspiration (AET) from Q1, Q2, Q3, and Q4, Potential Evapotranspiration (PET) from Q1 and Q2, Precipitation from Q1 (PptQ1), Maximum Temperature from Q1 (TMxQ1), and Minimum Temperature from Q1 (TMnQ1). The following Canopy Structure variables remained: Biomass (BM; referred to as BM_2_1 in appendix Table C1), Leaf Area Index 0- to 10 m, 10- to 20 m, 20- to 30 m, and 30- to 40 m, Vertical Distribution Ratio Middle (VDRM) and Vertical Distribution Ratio Bottom (VDRB). All Auxiliary variables remained after VIF. The following Phenology variables remained: NDVI Annual 95th Percentile (NDVI_Ann_95p), NDVI Seasonal Difference (NDVI_Seas_Diff), and NDVI Variance (NDVI_Var). The correlation matrix for the remaining variables is shown in appendix figure C3.

For all 500 bootstraps we summarized variable importance by predictor variable group for habitat associations (figure 3) and individual species (appendix E). We ranked DSA variable importance for each individual bootstrap and then selected the top-5 most important variables from each bootstrap. Figure 3 shows the relative distribution of all top-5 variables by variable group and habitat association. From this perspective at 250 m, Canopy Structure variables most frequently ranked in the top-5 for Conifer, Shrub, and Urban specialists. The importance of Canopy Structure diminishes at coarser spatial resolutions, such that Climate variables most frequently ranked in the top-5 when considering Shrub and Urban specialists at 1000 m. Across all spatial resolutions and habitat associations, Climate variables most frequently ranked in the top-5, but the relative frequencies generally matched the proportion of Climate variables available relative to the total number of variables. However, Oak, Riparian, and Grass specialists had a higher frequency of Climate variables ranked in the top-5. For all habitat associations, Phenology variables ranked in the top-5 more frequently than would be expected based on proportion of Phenology variables available relative to the total number of variables. Across all habitat associations at 250 m, Auxiliary variables comprised 12.9% of the top-5 (vs. 17.4% expected), Phenology 19.3% (vs. 13.0% expected), Canopy Structure 32.5% (vs. 30.4% expected), and Climate 35.3% (vs. 39.1% expected). These importance patterns are corroborated by another perspective in which we summed total importance for each resolution and predictor variable group (supplementary figure E2 (stacks.iop.org/ERL/15/095002/mmedia)).

The DSA importance method also provided insight into which individual predictor variables were most important (appendix figure E3). Individual Phenology variables were relatively more important than median importance of variables from all models, regardless of habitat association. Canopy Structure variables were usually about as important as the median variable importance. The two vertical distribution ratios (VDRB and VDRM) were more important than other canopy structure variables, with VDRB showing noticeably higher relative importance for all habitat associations. Canopy Structure variables BM, LAI 0- to 10 m, and LAI 10- to 20 m were relatively more important for Conifer specialists, particularly at finer spatial resolution. Climate variables were relatively homogenous in their importance, but TMxQ1 showed noticeable deviation from the median importance value. Precipitation from the
Figure 3. The percent of time that individual variables from different groups were included in the top-5 most important (DSA method) variables across all 500 bootstrap iterations. Percent Available is percentage of variables available from a group (Canopy Structure, Climate, Phenology, Auxiliary) relative to the total number of variables remaining after VIF (n = 23) and can be thought of as a baseline for comparison if all variables were equally important. The dashed vertical line corresponds to Canopy Structure percent available and is used for assessing the relative importance of this variable group for different habitat associations. We included all models (except XGBoost) which had AUC > 0.5 and used at least 5 variables.

first quarter (PptQ1) was another relatively important Climate variable. Importance of Climate variables generally increased going from fine- to coarse-scale.

3.2. Model performance
3.2.1. Comparing performance of different model scenarios
Ensemble models were most frequently the best performing models over all bootstraps (appendix F.a.). Figure 4 shows notched boxplots corresponding to Ensemble Weighted Average (EWA) AUC values from 500 bootstraps for each species and resolution. These boxplots are useful for visualizing differences among bootstrap population medians. The ‘All with Canopy Structure’ scenario showed median AUC improvements (i.e. non-overlapping notches) for at least one resolution for 19 species: CBCH (all), DEJU (250- and 500 m), NOFL (250- and 500 m), SPTO (250- and 500 m), STJA (all), AMGO (250- and 500 m), BUSH (250 m), NOMO (500 m), OATI (250 m), WESJ (250 m), RWBL (500 m), SOSP (250 m), CALT (all), CAQU (250 m), WSCP (500 m), BRBL (all), BEWR (250- and 500 m), and BHGR (250 m). Every Conifer specialist showed a median AUC improvement for at least one spatial resolution. Five of the six Oak specialists showed an improvement in performance for at least one spatial resolution. Few of the Grass or Riparian specialists showed improvement after incorporating Canopy Structure variables. All three Shrub specialists showed a median AUC improvement for at least one spatial resolution. Lastly, two of four Urban species showed a median AUC improvement for at least one spatial resolution. Thresholded differences in AUC are shown in Supplementary Table F1. For example, at 250 m, the median AUC value increased by at least 0.02 AUC for 10 of 25 species, while the 95th percentile AUC value increased for 14 of 25 species. The maximum increase in median AUC after incorporating Canopy Structure variables was 0.04 for CBCH.

3.2.2. Influence of spatial resolution
We summarize ‘All with Canopy Structure’ median model performance by spatial resolution in Supplementary Table F2. Considering all 25 SDMs, most (11) have the highest median AUC value at 250 m. Grass and Riparian habitat specialists are notable exceptions, with all of their highest median AUC values at 500 m and 1000 m, respectively. Conifer, Oak, and Shrub specialists generally have the highest median AUC at 250 m spatial resolution. Supplementary Table F2 also shows the difference between the maximum and minimum AUC medians across all spatial resolutions, providing insight into the scale variability of each SDM. The average scale variability is 0.03 AUC, while the maximum variability is 0.067 AUC. Therefore, the effect of scale on median model performance is similar to the effect of incorporating Canopy Structure variables.

3.3. Ensemble maps
We calculated CIE weighted average predictions from individual models which had AUC values greater than 0.5. Figure 5 shows species distribution maps for birds from a range of habitat associations as predicted by the ensemble weighted average as well as the corresponding uncertainty. The uncertainty estimates include variance due to the model used, the bootstrap sampling, and unexplained variance by the model. Because of the small sample sizes of the bird observation data, the total predicted error is relatively large. All species maps are presented in appendix G.

4. Discussion
Our ability to describe, parameterize, and model species’ habitat continues to improve with the addition and refinement of relevant geospatial datasets. Our central objective was to determine the added benefit of Canopy Structure variables from simulated GEDI lidar in SDMs that also include relevant Climate, Phenology, and Auxiliary variables. We
Figure 4. Model performance comparison between two model scenarios—‘All with Canopy Structure’ vs. ‘All without Canopy Structure’. Each boxplot corresponds to EWA AUC values from up to 500 bootstraps. Notches that do not overlap can be interpreted as having different median values. ‘X’ points on each boxplot correspond to the 95th percentile value.

quantified the added benefit of these Canopy Structure variables by calculating variable importance of the ‘All with Canopy Structure’ model scenario and by comparing the performance of this model scenario against the ‘All without Canopy Structure’ model scenario using 500 bootstraps. The spatial scale at which SDMs are generated is a very important consideration and impacts our estimates of variable importance and comparative assessment of model performance. For this analysis we selected three practical SDM spatial scales which coincided with readily available and ecologically-meaningful Climate, Phenology, and Auxiliary predictor variables. Below we discuss the variables that remained after the VIF process, their importance, the change in SDM performance after incorporating Canopy Structure variables, current constraints on the analysis, and ideas for building on this work in the future.
Figure 5. Ensemble maps of weighted average probability of occurrence and associated uncertainty of the mean for one species from each habitat association at 250 m spatial resolution. Uncertainty of the mean prediction is shown as the range of the standard error confidence interval, that is the upper 1 SE confidence interval minus the lower 1 SE confidence interval. Other maps are presented in Supplementary Material Appendix G.
4.1. Model variables and importance

Methods to reduce redundant information across predictor variables, such as VIF, are useful for creating models that are parsimonious and interpretable in terms of variable importance. Furthermore, the elimination of certain variables can provide insight regarding unique information content relative to the entire stack of environmental predictors. Considering all variable groups, we found Climate variables to be the most important overall. Maximum temperature was the most important Climate variable for Grass and Oak species, which agrees with previous studies (Howard et al 2015, Bradie and Leung 2017). Six of nine Climate variables were associated with either potential or actual evapotranspiration (ET), which is related to a combination of temperature, moisture availability, and plant productivity. Hawkins et al (2003) and Coops et al (2018) found ET variables to be important for predicting species richness at broad spatial scales. Barbet-Massin and Jetz (2014) showed that annual PET was among the most relevant variables for predicting bird species distributions in the conterminous U.S. The ET variables do not appear to be relatively as important when grouped by habitat association, but still contain some unique information. For example, PETQ2 is one of the most important variables for Bushtit and Mourning Dove.

Seven Canopy Structure variables remained following VIF reduction, and as a group were usually the second most important variable group depending on habitat association. Canopy structure is an important determinant of avian diversity, including the partitioning of niches for species coexistence, changes in microclimate (Zellweger et al 2019) and refuge from predation (Davies and Asner 2014). Müller et al (2010) found that 3D structure measured with airborne lidar was the main statistical determinant of bird assemblages in a German temperate mixed forest. Macarthur and Macarthur (1961) first reported the link between bird species diversity and foliage height diversity (FHD). However, neither version of the FHD metrics remained following VIF reduction. This is not necessarily because the variable is without utility, but rather some of its information content varied linearly with other variables that also include information about the vertical distribution of plants, like vertical distribution ratios (VDRB and VDRM), LAI profiles, and NDVI. VDRB was a particularly important variable for all habitat associations. This variable was initially included to characterize how evenly plant material is distributed, as it compares the overall canopy height to the height of 50% of the cumulative waveform signal. In theory, values close to 0.5 indicate an even distribution, while lower values should indicate a higher concentration of plant material in the understory, and vice versa. In actuality, when examining the kriged raster we observe that the variable shows spatial correspondence with tall vegetation or built-up structures, both of which could be useful for perching or nesting. Some species SDMs may find this variable useful for discriminating between canopy structure characteristics. For instance, this was one of the most important variables for California Towhee, which commonly uses dense shrub/scrub areas that are low to the ground.

Canopy Structure variables were most important for Conifer specialists (figure 3), suggesting that these species seem to prefer locations with certain Canopy Structure characteristics (i.e. based on importance metrics, more biomass at the canopy level and denser foliage, or LAI, in the 0- to 10 m and 10- to 20 m strata). This agrees with our observations of these conifer-associated species in Sonoma County. For example, Dark-eyed Junco is often found in areas with complex vegetation structure less than 5 m from the ground, but also forages in mid-canopy. During the breeding season, males often sing from exposed perches near the tops of conifers and snags. Chestnut-backed Chickadee prefers structurally-complex mature conifer forests, as corroborated by the importance of the variable LAI 20- to 30 m, and is often found gleanig small insects and other arthropods from bark and twigs.

Phenology variables mainly capture temporal cycling of top-of-canopy foliage status, but in some cases (low to moderate forest cover) may contain information about foliage density associated with the near-infrared component of NDVI which is related to the phenomenon of leaf additive reflectance (Jensen 1996). The Phenology variables had the highest top-5 importance difference relative to their expected importance (+6.3% overall at 250 m) suggesting that these variables contain particularly useful information for bird SDMs. Visually, the three Phenology variables show some correspondence with broad land cover and vegetation classes (figure 2), which may provide sufficient predictive information for some SDMs. Canopy Structure variables from GEDI also had a positive top-5 importance difference (+2.1% overall at 250 m) for all habitat associations. Canopy Structure variables likely complement Climate and Phenology variables by adding 3D information related to microhabitat structure that is not captured by the other variables we used. This 3D perspective has the greatest benefit for Conifer specialists (+9.5% overall at 250 m).

4.2. SDM performance

4.2.1. Ensemble model performance perspective

Fewer than half of the ensemble SDMs had median AUC greater than 0.7, a threshold for assessing SDM quality (Pearce and Ferrier 2000, Duan et al 2014). The overall performance of these 25 SDMs was likely related to the bird observation data and the large distribution range of the individual species (Mcpherson and Jetz 2007). Although Sonoma County is relatively diverse in terms of climate, topography, and vegetation type, more bird observations
spanning a wider range of the environmental predictor variables could lead to better presence/absence separation and thus improved SDM performance. Furthermore, the best performing SDMs were generally associated with Conifer and Oak species, which are relatively more selective and less likely to occur across multiple habitats in Sonoma County. Fairly common species SDMs, such as Mourning Doves, Song Sparrows, and House Finches did not benefit much from the incorporation of Canopy Structure information due to their high prevalence in multiple habitats (Kadmon et al. 2003, Brotons et al. 2004, but see Gavish et al. 2017 for a counter-example).

Larger observation datasets and more certainty in presence/absence records both help models to more accurately depict the niche of common species (Kadmon et al. 2003). While there are a large number of survey events in Sonoma County (n = 10606), these occurred largely in the same locations—when aggregated within cells, only a few hundred cells were surveyed (see table 1). However, because most cells had many survey events, our determination of presence/absence within each cell is probably very accurate, the more so for these common and easy to identify generalist species. It is therefore likely that the relatively poor performance of most models is due to low sample sizes and associated lack of high-density observations from different land-cover classes. These are common limitations associated with citizen science observations, such as those used in this study (Boakes et al. 2010, Beck et al. 2014).

Although Canopy Structure variables frequently improved median model performance, especially at finer spatial resolution, the improvement in performance was relatively modest considering the number of new variables (and potential amount of new information) added to the models. The various machine learning models used in the ensemble performed well at discerning patterns and optimizing predictions with relatively few predictor variables. Another reason for modest improvements in performance was related to correlation and information-overlap between variables remaining following VIF. Even after using VIF, there was still correlation among the predictor variables (see appendix figure C3) suggesting that some canopy structure information could be explained by other variables. For examples we found Pearson correlation coefficients greater than 0.6 when regressing NDV195p against three Canopy Structure variables (Biomass, LAI 0- to 10 m, and LAI 10- to 20 m).

Lastly, spatial scale dependence cannot be ignored in SDM analyses (Wiens 1989, Jackson and Fahrig 2015). In this study, we found an average scale variability of 0.03 AUC (Supplementary Table F2) which is similar to the effect of incorporating Canopy Structure variables in some SDMs. When going from fine to coarse spatial scale, we found Canopy Structure variable importance decreased and incorporation of Canopy Structure variables less frequently improved model performance. Mertes and Jetz (2018) found resampling environmental predictor variables with fine spatial structure decreased the performance of SDMs. Curry et al. (2018) found that spatially-explicit ensemble models with finer-scale predictor variables consistently performed better at predicting occurrence for 3 of 11 grassland bird species. Our results, and those from others, are related to the spatial structure and autocorrelation of Canopy Structure variables (Mertes and Jetz 2018). We did not explicitly factor in spatial autocorrelation of environmental predictor variables into SDMs, but semivariograms generated for each Canopy Structure variable confirmed the spatial structure (semivariogram range) of each variable was generally within 250- to 1000 m (Supplementary Table C2). The majority of semivariograms generated from GEDI simulation subsets within Conifer and Oak forests have ranges below 500 m. Additional SDM performance improvements may be possible through exploration of multi-scale optimization (e.g. Wan et al. 2017, Stevens and Conway 2019), especially when considering even finer-scale vertical (Seavy et al. 2009, Gastón et al. 2017) and horizontal (Huang et al. 2014, Carrasco et al. 2019) Canopy Structure variables.

4.3. Constraints on the current analysis

The primary limitations of this study are the quantity and geographic distribution of bird observations, as well as our ability to represent canopy structure at optimal scales using simulated spaceborne lidar data. The biggest limitation associated with the bird species observation data is a lack of high-density, evenly-distributed spatial observations. As shown in figure 1, most observations occurred in urban/suburban areas. Since we are primarily interested in how forest Canopy Structure variables improve SDMs, this was not the ideal network of observations because relatively few observations were distant from roads and inside of intact forests. The impact of these Canopy Structure variables is likely more difficult to resolve when urban features are mixed with tree canopies at the grid cell resolutions used in this analysis. We would expect to receive a more consistent signal from a ‘natural’ forest canopy compared with a suburban or urban-wildland interface canopy interspersed with built-up structures.

Other limitations are associated with our ability to represent canopy structure using simulated spaceborne lidar data. First, the simulated lidar data are hypothetical geospatial observations—actual mission footprint locations will vary as a function of orbital geometry, which cannot be predicted exactly, and local cloud cover patterns. Another major limitation concerns interpolation from footprints (L2 products) to grids (L3 products). There are numerous options for interpolating point data to continuous grids.
depending on the nature of the dataset. As a starting point, we used a kriging methodology similar to the one outlined in the GEDI L3 Algorithm Theoretical Basis Document (Dubayah et al 2020). Over the course of two years, GEDI is expected to provide sufficient coverage for generating L3 products at 1000 m spatial resolution. However, coverage gaps will still be present in 1000 m spatial resolution grids and would be even more prevalent at the finer spatial resolutions used in this study (see appendix figure C2). For natural and continuous forest types, we expect this kriging method to be a good option for interpolation as neighboring grid cells tend to have similar geostatistical characteristics. For fragmented or urban/suburban forests, this method may not be as effective. More advanced interpolation methods such as co-kriging (Tsui et al 2013) or fusion (Swatantran et al 2012) with auxiliary data with continuous spatial coverage, like Sentinel-1, Sentinel-2 or Landsat satellite imagery, may prove to be better options for interpolating large areas or filling in sparsely-sampled grid cells.

4.4. Future directions

We expect canopy structure measurements from GEDI will be informative for modelling habitat for a variety of forest-dependent taxa. Our modeling approach is applicable to larger spatial extents and taxa, but one important consideration would be the source of Climate data, as BCM is a California-wide climate product with an unusually fine spatial resolution (270 m). Global climate variables from WorldClim (Fick and Hijmans 2017), which have approximately 1000 m spatial resolution, are more commonly used for larger SDM extents outside of the conterminous U.S. We note that this resolution matches the planned GEDI L3 products. It will be necessary to test the utility of these global climate variables and GEDI L3 metrics in SDMs at their native resolution. However, for some species the tradeoff in model performance as a function of spatial scale may preclude the use of relatively coarse vegetation structure and/or climate data (Manzoor et al 2018). Alternatively, shifting to more of a microhabitat focus, recent studies have proposed frameworks for dynamically modeling microclimate using fine-scale topography and vegetation structure datasets derived from remote sensing (Kearney et al 2019, Lembrechts et al 2019). Using these finer spatial and temporal scale climate datasets in conjunction with vegetation structure measurements will improve our understanding of the mechanistic interplay between the two, as well as how species and ecosystems respond to changes in climate (Zellweger et al 2019).

It will also be beneficial to explore data and methods for increasing the quantity and coverage of canopy structure measurements. GEDI is scheduled for two years of data collection on the ISS but a different type of spaceborne lidar data, being collected simultaneously from ICESat-2, may have some utility in filling GEDI canopy structure coverage gaps. The Advanced Topographic Laser Altimeter System (ATLAS) instrument on-board ICESat-2 uses photon counting technology which has limitations in dense vegetation, especially during daytime observations (Popescu et al 2018). Although vegetation measurement is not the primary focus of the ICESat-2 mission, we still expect some useful canopy structure observations in forests of sparse to moderate canopy cover (Neuenschwander and Pitts 2019).

5. Conclusions

We incorporated Canopy Structure variables derived from simulated GEDI lidar into 25 bird SDMs in Sonoma County, CA. We found Canopy Structure variables were generally the second most important group of variables across a variety of habitats. Canopy Structure was most important for Conifer and Shrub habitat associations. The incorporation of Canopy Structure variables into ensemble SDMs resulted in performance improvements for the majority of species across all spatial analysis resolutions. More SDMs showed higher, albeit modest, performance improvements at the finest spatial analysis scale suggesting that Canopy Structure variables are more beneficial at smaller scales which are more in line with microhabitats. Although this study covers a regional spatial extent and limited number of species, it demonstrates that Canopy Structure variables derived from spaceborne lidar have the potential to improve our ability to map species distributions.

Our understanding of global forest canopy structure will grow by orders of magnitude over the next decade as a result of the GEDI and ICESat-2 spaceborne lidar missions. We encourage future SDM efforts to explore the utility of operational lidar products from these sensors across natural and anthropogenic landscapes. Future SDM studies may also benefit from fusion of spaceborne lidar with other remote sensing data as this combination can provide more continuous spatial coverage across a broader range of spatial scales. SDMs generated at even finer spatial scales may benefit more from lidar-derived Canopy Structure variables, resulting in more detailed map products for land management and conservation of species habitat.

Acknowledgments

Airborne lidar data were provided by the University of Maryland and the Sonoma County Vegetation Mapping and Lidar Program under grant NNX13AP69G from NASA's Carbon Monitoring System (Dubayah, PI). We acknowledge Curt Joyce at Quantum Spatial Inc. who helped to acquire the airborne lidar data. Curt, a leader in outdoor, remote sensing, and environmental communities, passed away in 2014. Xiaoli
Sun conducted the link margin analysis we used to set the noise levels in the GEDI simulation. Scott Luthcke produced orbital simulations and gave expert suggestions regarding expected data quality/quantity.

Bird observations were largely collected and organized by the invaluable contribution of a number of citizen scientists, through eBird (Sullivan et al 2009) and the BBS programs.

We used Northern Arizona University’s high-performance computing system (Monsoon) to run all species distribution models. Maxwell Gilbert, Emma Forester, and Mariana Palacio Salazar assisted with pre-processing of predictor variable datasets. Sean Mahoney provided contextual knowledge about bird species of interest. Wendy Schackwitz provided a helpful review.

We acknowledge funding support from NASA’s Citizen Science for Earth Systems Program (Cooperative Agreement 80NSSC18M0107 to MC) and the GEDI Science Definition Team (NNL15AA03C to SJG).

Lastly, the manuscript benefited substantially from two anonymous reviewers.

Code and Data Availability

The data that support the findings of this study are available upon request from the authors.

Relevant model input data and code are stored on GitHub https://github.com/leosalas/Soundscapes2 Landscapes/tree/master/GEDI_Bird_SDG

ORCID iDs

Patrick Burns https://orcid.org/0000-0001-7971-4177
Matthew Clark https://orcid.org/0000-0001-5953-2990
Leonardo Salas https://orcid.org/0000-0002-8801-548X
Steven Hancock https://orcid.org/0000-0001-5659-6964
David Leland https://orcid.org/0000-0001-7931-2223
Patrick Jantz https://orcid.org/0000-0001-5103-2270
Ralph Dubayah https://orcid.org/0000-0003-1440-6346
Scott J Goetz https://orcid.org/0000-0002-6326-4308

References

Adams B T and Matthews S N 2019 Diverse temperate forest bird assemblages demonstrate closer correspondence to plant species composition than vegetation structure Ecol. Geogr. 42 1752–64
Aguirre-Gutiérrez J, Carvalheiro L G, Polce C, van Loon E E, Raes N, Reemer M and Biesmeijer J C 2013 Fit-for-purpose: species distribution model performance depends on evaluation criteria–Dutch hoverflies as a case study PLoS One 8 e63708
Anderson K, Hancock S, Disney M and Gaston K J 2016 Is waveform worth it? A comparison of Li DAR approaches for vegetation and landscape characterization Remote Sens. Ecol. Conserv. 2 5–15
Aiello-Lammens M E, Boria R A, Radosavljevic A, Vilela B and Anderson R P 2015 sPThin: an R package for spatial thinning of species occurrence records for use in ecological niche models Ecol. Geogr. 38 541–5
Araújo M B et al 2019 Standards for distribution models in biodiversity assessments Sci. Adv. 5 eaat4858
Araújo M B and Guisan A 2006 Five (or so) challenges for species distribution modelling J. Biogeogr. 33 1677–88
Austin M P and Van Niel K P 2011 Improving species distribution models for climate change studies: variable selection and scale J. Biogeogr. 38 1–8
Bakx T R, Koma Z, Seijmonsbergen A C and Kissing W D 2019 Use and categorization of light detection and ranging vegetation metrics in avian diversity and species distribution research Divers. Distrib. 25 1045–59
Barbet-Massin M and Jetz W 2014 A 40-year, continent-wide, multispecies assessment of relevant climate predictors for species distribution modelling Divers. Distrib. 20 1265–95
Beck J, Böller M, Erhardt A and Schwanghart W 2014 Spatial bias in the GBIF database and its effect on modeling species’ geographic distributions Ecol. Inform. 19 10–15
Bell D M and Schlaepfer D R 2016 On the dangers of model complexity without ecological justification in species distribution modeling Ecol. Modell. 380 50–59
Bergen K M, Goetz S J, Dubayah R O, Henebry G M, Hunsaker C T, Imhoff M L, Nelson R F, Parker G G and Radloff V C 2009 Remote sensing of vegetation 3-D structure for biodiversity and habitat: review and implications for lidar and radar spaceborne missions J. Geophys. Res. Biogeosciences 114 G00E06
Boakes E H, Mcgowan P J, Fuller R A, Chang-qing D, Clark N E, O’Connor K and Mace G M 2010 Distorted views of biodiversity: spatial and temporal bias in species occurrence data PLoS Biol. 8 e1000385
Bradie J and Leung B 2017 A quantitative synthesis of the importance of variables used in MaxEnt species distribution models J. Biogeogr. 44 1344–61
Brotons L, Thuiller W, Araújo M B and Hirzel A H 2004 Presence-absence versus presence-only modelling methods for predicting bird habitat suitability Geogr. Earth Sci. 27 437–48
Carrasco L, Giam X, Papes M and Sheldon K S 2019 Metrics of species composition than vegetation structure for biodiversity and habitat: review and implications for lidar and radar spaceborne missions J. Geophys. Res. Biogeosciences 114 G00E06
Corbane C, Lang S, Pipkins K, Alleaume S, Deshayes M, Millán V E G, Strasser T, Borre J V, Toon S and Michael F 2015 Remote sensing for mapping natural habitats and their conservation status–New opportunities and challenges Int. J. Appl. Earth Obs. Geoinf. 37 7–16
Cortez P 2016 rmrner: data mining classification and regression methods R Package Version 1.4.2
Cortez P and Embrecht M J 2015 Using sensitivity analysis and visualization techniques to open black box data mining models Inf. Sci. 225 1–17
Curry C M, Ross J D, Contina A J and Bridge E S 2018 Varying dataset resolution alters predictive accuracy of spatially explicit ensemble models for avian species distribution Ecol. Evol. 8 12867–78
Davies A B and Asner G P 2014 Advances in animal ecology from 3D-LiDAR ecosystem mapping Trends Ecol. Evol. 29 681–91
Duan R Y, Kong X Q, Huang M Y, Fan W Y and Wang Z G 2014 The predictive performance and stability of six species distribution models PLoS One 9 e112764
Dubayah R O et al 2020 The global ecosystem dynamics investigation: high-resolution laser ranging of the Earth’s forests and topography Sci. Remote Sens. 1 100002
Engler J O, Siels D, Schidelko K, Strudbe D, Quillfeldt P and Brambilla M 2017 Avian SDMs: current state, challenges, and opportunities J. Avian Biol. 48 1483–504
Fick S E and Hijmans R J 2017 WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas Int. J. Climatol. 37 4302–15
Fielding A H and Bell J F 1997 A review of methods for the assessment of prediction errors in conservation presence/absence models Environ. Conserv. 24 38–49
Flint L E, Flint A L, Thorne J H and Boynton R 2013 Fine-scale hydrologic modeling for regional landscape applications: the California basin characterization model development and performance Ecol. Processes 2 25
Gaston A et al 2017 Species’ habitat use inferred from environmental variables at multiple scales: how much do we gain from high-resolution vegetation data Int. J. Appl. Earth Obs. Geoinf. 55 1–8
Gavish Y, Marsh C J, Kuevermelen M, Stoll S, Haase P and Kunin W E 2017 Accounting for biotic interactions through alpha-diversity constraints in stacked species distribution models Methods Ecol. Evol. 8 1092–102
Goetz S J, Steinberg D, Betts M G, Holmes R T, Doran P J, Dubayah R and Hofton M 2010 Lidar remote sensing variables predict breeding habitat of a Neotropical migrant bird Ecology 91 1569–76
Goetz S J, Steinberg D, Dubayah R and Blair B 2007 Laser remote sensing of canopy habitat heterogeneity as a predictor of bird species richness in an eastern temperate forest, USA Remote Sens. Environ. 108 254–63
Goetz S J, Sun M, Zollos S, Hansen A and Dubayah R 2014 The relative importance of climate and vegetation properties on patterns of North American breeding bird species richness Environ. Res. Lett. 9 034013
Hancock S, Arrmston J, Hofton M, Sun X, Tang H, Duncanson L I, Kellner J R and Dubayah R 2019 The GEDI simulator: a large-footprint waveform lidar simulator for calibration and validation of spaceborne missions Earth Space Sci. 6 294–310
Hawkins B A, Porter E E and Felizia Diniz-Filho J A 2003 Productivity and history as predictors of the latitudinal diversity gradient of terrestrial birds Ecology 84 1608–23
He K S, Bradley B A, Cord A F, Rocchini D, Tuanmu M N, Schmidtlein S, Turner W, Wegmann M and Pettorelli N 2015 Will remote sensing shape the next generation of species distribution models? Remote Sens. Environ. 160 8–23
Hill R A, Hinsley S A, Gaveau D L A and Bellamy P E 2004 The influence of vegetation height heterogeneity on forest and woodland bird species richness across the United States PLoS One 9 e103236
Hunter M L and Hunter Jr M L (eds.) 1999 Maintaining Biodiversity in Forest Ecosystems (Cambridge: Cambridge University Press)
IPBES 2019 Summary for policymakers of the global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services ed S Diaz et al (Bonn, Germany: IPBES Secretariat)
Jackson H B and Fahrig L 2015 Are ecologists conducting research at the optimal scale? Global Ecol. Biogeogr. 24 52–63
Jensen J R 1996 Introductory Digital Image Processing: A Remote Sensing Perspective 2 edn (Englewood Cliffs, NJ: Prentice-Hall Inc)
Jetz W, Wilcove D S and Dobson A P 2007 Projected impacts of climate and land-use change on the global diversity of birds PLoS Biol. 5 1–6
Kadmon R, Faber O and Danin A 2003 A systematic analysis of factors affecting the performance of climatic envelope models Ecol. Appl. 13 853–67
Kearney M R, Gillingham P K, Branner I, Duffy J P and Macken L M 2019 A method for comparing hourly, historical, terrain-corrected microclimate anywhere on Earth Methods Ecol. Evol. 11 38–43
Lawler J J, Shafer S L, White D, Kareiva P, Maurer E P, Blaustein A R and Bartlein P J 2009 Projected climate-induced faunal change in the Western Hemisphere Ecology 90 388–97
Lembrechts J N, Nijs I and Lenoir J 2019 Incorporating microclimate into species distribution models Ecography 42 1267–79
Levin S A 1992 The problem of pattern and scale in ecology: the Robert H MacArthur lecture Ecology 73 1943–67
Lim K, Treitz P, Wulder M, St-Onge B and Flood M 2003 LiDAR remote sensing of forest structure Prog. Phys. Geogr. 27 88–106
Macarthur R H and Macarthur J W 1961 On bird species diversity Ecology 42 594–8
Manzoor S A, Griffiths G and Lukac M 2018 Species distribution model transferability and model grain size–finer may not always be better Sci. Rep. 8 7168
Martínez-Freiría F, Tarroso P, Rebelo H and Brito J C 2016 Multi-scale habitat selection modeling: a review and opportunities Int. J. Appl. Earth Obs. Geoinf. 57 247–59
Mayor S J, Schneider D C, Schaefer A J and Mahoney S P 2009 Habitat selection at multiple scales Ecocene 16 238–47
McClure C J, Korte A C, Heath J A and Barber J R 2015 Pavement and riparian forest shape the bird community along an urban river corridor Global Ecol. Conserv. 4 291–310
McGarigal K, Wan H Y, Zeller K A, Timm B C and Cushman S A 2016 Multi-scale habitat selection modeling: a review and outlook Landsc. Ecol. 31 1161–75
Mcpherson J and Jetz W 2007 Effects of species’ ecology on the accuracy of distribution models Ecography 30 135–51
Mertes K and Jetz W 2018 Disentangling scale dependencies in species environmental niches and distributions Ecography 41 1604–15
Mod H K, Scherrera D, Luoto M and Guisan A 2016 What we use is not what we know: environmental predictors in plant distribution models J. Veg. Sci. 27 1308–22
Müller J, Stadler J and Brandl R 2010 Composition versus productivity and history as predictors of the latitudinal diversity gradient of terrestrial birds Ecology 81 103–12
Nepstad D K, Chistie H, Echeverria D, Jaramillo E and Ramirez E 1994 Tracking carbon: forest and non-forest land-use in the Amazon region Conserv. Biol. 8 130–40
Neuenzweigner A and Pitts K 2019 The ATL08 land and vegetation product for the ICESat-2 mission Remote Sens. Environ. 221 247–59
Pearce J and Ferrier S 2000 Evaluating the predictive performance of habitat models developed using logistic regression Ecol. Modell. 133 225–45

Pereira H M et al 2013 Essential biodiversity variables Science 339 277–8

Papenfuss C J, Zhou T, Nelson N, Neuenwander A, Sheridan R, Nairn L and Walsh K M 2018 Photon counting LiDAR: an adaptive ground and canopy height retrieval algorithm for ICESat-2 data Remote Sens. Environ. 208 154–70

Proença V et al 2017 Global biodiversity monitoring: from data sources to essential biodiversity variables Biol. Conserv. 213 256–63

R Core Team 2019 R: A Language and Environment for Statistical Computing (Vienna, Austria: R Foundation for Statistical Computing) https://www.R-project.org/

Radeloff V C et al 2019 The dynamic habitat indices (DHIs) from MODIS and global biodiversity Remote Sens. Environ. 222 204–14

Rechsteiner C, Zellweger F, Gerber A, Breiner F T and Bollmann K 2017 Remotely sensed forest habitat structures improves regional species conservation Remote Sens. Ecol. Conserv. 3 247–58

Rocchini D et al 2015 Potential of remote sensing to predict species invasions: A modelling perspective Prog. Phys. Geogr. 39 283–309

Rodewald P (Editor) 2015 The Birds of North America https://birdsna.org (Ithaca, NY: Cornell Laboratory of Ornithology)

Rosenberg K V et al 2019 Decline of the North American avifauna Science 366 120–4

Rotenberry J T 1985 The role of habitat in avian community composition: physiognomy or floristics? Oecologia 67 213–17

Rottenborn S C 1999 Predicting the impacts of urbanization on riparian bird communities Biol. Conserv. 88 289–99

Sauer J R, Niven D K, Hines J E, Ziolkowski D J Jr, Pardecik K L, Fallon J E and Link W A. 2017 The north american breeding bird survey, results and analysis 1966–2015 Version 2. 07.2017 (Laurel, MD: USGS Patuxent Wildlife Research Center)

Seavy N E, Viess J H and Wood J K 2009 Riparian bird response to vegetation structure: a multiscale analysis using LiDAR measurements of canopy height Ecol. Appl. 19 1848–57

Sohl T L 2014 The relative impacts of climate and land-use change on conterminous United States bird species from 2001 to 2075 PLoS One 9 e112251

Stevens B S and Conway C J 2019 Predicting species distributions: unifying model selection and scale evaluation for multi-scale occupancy models Ecosphere 10 e02748

Stralberg D, Jongomijit D, Howell C A, Snyder M A, Alexander J D, Wiens J A and Root T L 2009 Re-shuffling of species with climate disruption: a no-analog future for California birds? PLoS One 4 e6825

Sulla-Menashe D, Gray J M, Abercrombie S P and Friedl M A 2019 Hierarchical mapping of annual global land cover 2001 to present: the MODIS collection 6 land cover product Remote Sens. Environ. 222 183–94

Sullivan B L, Wood C I, Iliff M J, Bonney R E, Fink D and Kelling S 2009 eBird: A citizen-based bird observation network in the biological sciences Biol. Conserv. 142 2282–92

Swatantran A, Dubayah R, Goetz S, Hofm RN, Betsis M G, Sun M, Simard M and Holmes R 2012 Mapping migratory bird prevalence using remote sensing data fusion PLoS One 7 e28922

Syphard A D and Franklin J 2010 Species traits affect the performance of species distribution models for plants in southern California J. Veg. Sci. 21 177–89

Tang H, Dubayah R, Swatantran A, Hofm RN, Sheldon S, Clark D B and Blair B 2012 Retrieval of vertical LAI profiles over tropical rain forests using waveform lidar at La Selva, Costa Rica Remote Sens. Environ. 124 242–50

Tattoni C, Rizzoli F and Pedrini P 2012 Can LiDAR data improve bird habitat suitability models? Ecol. Modell. 245 103–10

Thibaud E, Petitpierre B, Broennimann O, Davison A C and Giusan A 2014 Measuring the relative effect of factors affecting species distribution model predictions Methods Ecol. Evol. 5 947–55

Troudet J, Grandcolas P, Blin A, Vignes-Lebbe R and Legendre F 2017 Taxonomic bias in biodiversity data and societal preferences Sci. Rep. 7 1–14

Tsai O W, Coops N C, Wulder M A and Marshall P L. 2013 Integrating airborne LiDAR and space-borne radar via multivariate kriging to estimate above-ground biomass Remote Sens. Environ. 139 340–52

U.S. Geological Survey 2019 Breeding bird atlas explorer (online resource) http://www.pwrc.usgs.gov/bba (Accessed 29 May 2019).

Vierling K T, Vierling L A., Gould WA, Martinuzzi S and Clwegs R M 2008 Lidar: shedding new light on habitat characterization and modeling Front. Ecol. Environ. 6 90–98

Wan H Y, Megarigal K, Ganey J L, Laurent V, Timm B C and Cashman S A 2017 Meta-replication reveals nonstationarity in multi-scale habitat selection of Mexican spotted owl The Condor: Ornithol. Appl. 119 641–58

Whittaker R J, Willis K J and Field R 2001 Scale and species richness: towards a general, hierarchical theory of species diversity J. Biogeogr. 28 453–70

Wiens J A 1989 Spatial scaling in ecology Funct. Ecol. 3 385–97

Wilson A M and Jetz W 2016 Remotely sensed high-resolution global cloud dynamics for predicting ecosystem and biodiversity distributions PLoS Biol. 14 e1002415

Zellweger F, De Frenne P, Lenoir J, Rocchini D and Coomes D 2019 Advances in microclimate ecology arising from remote sensing Trends Ecol. Evol. 34 327–41