## 1 Introduction

## 2 Methods

### 2.1 Study area

We modeled woodcock habitat distribution throughout the state of Pennsylvania, which provides breeding habitat for an estimated 2.3% of the global woodcock population (52 400 birds) and provides stopover habitat during spring and fall migration for woodcock breeding throughout the northeastern United States and eastern Canada, which accounts for nearly 1/3rd of the global woodcock population (30.5% of global woodcock, 684 500 birds; Kelley et al., 2008). Pennsylvania also contains some wintering habitat for woodcock, which is negligible compared to breeding and stopover habitat (Fig. 1). Pennsylvania is composed of 11 U.S. Environmental Protection Agency (EPA) level 3 ecoregions (Omernik and Griffith, 2014), which reflect the topological gradient from Pennsylvania’s ridge-and-valleys in the central portion of the state to the coastal lowlands along the edge of Lake Erie. Less mountainous areas in Pennsylvania tend to be heavily agricultural (ex. Northern Piedmont ecoregion, 38% agricultural; Jin et al., 2019), with development primarily concentrated around the two largest cities, Philadelphia (pop. 1,600,000; U.S. Census Bureau, 2021) and Pittsburgh (pop. 300 000). Mountainous areas, such as the North Central Appalachians ecoregion, remain in mostly contiguous forest cover (84% forest; Jin et al., 2019).

Woodcock in Pennsylvania are managed by the Pennsylvania Game Commission, a state wildlife management agency, which regulates hunting and manages habitat for wildlife. The Pennsylvania Game Commission owns more than 600 000 hectares of land, referred to hereafter as state gamelands, which are managed primarily for wildlife and to provide hunting and trapping opportunities for the public (Pennsylvania Game Commission, 2023). Managing woodcock habitat for both breeding and migratory seasons are priorities for the Pennsylvania Game Commission, which requires tools to prioritize management projects on state gamelands.

### 2.2 Breeding and migratory season data

To model woodcock habitat distribution during the breeding and migratory seasons, we used separate data sources that described woodcock occupancy during each of those time periods. For the breeding season (March–May), we used survey data collected as part of the federally-coordinated American Woodcock Singing Ground Survey (Seamans and Rau, 2020) and additional state-level monitoring conducted by the Pennsylvania Game Commission. Both state and federal surveys consisted of 5.76 km routes with 10 evenly spaced points, where observers listened for woodcock songs during their crepuscular breeding display. Observers recorded counts of all males singing during 2-minute intervals shortly after dusk. We used survey data collected from 2016–2020, and distilled records to presence or likely absence at each point based on detection of at least one male during the 5-year period. Male woodcock are often assumed to display near female nesting habitat, and so male displays are likely an indicator of male and female presence at the scale of application (McAuley et al., 2020).

We delineated woodcock occurrence during the migratory season using GPS-tracking data from the Eastern Woodcock Migration Research Cooperative, a collaboration of 42 federal, state, provincial, non-profit, and university partners throughout the United States and Canada (www.woodcockmigration.org). We captured woodcock at 34 sites in Quebec, Ontario, Nova Scotia, Maine, Vermont, New York, Rhode Island, Pennsylvania, Maryland, West Virginia, Virginia, North Carolina, South Carolina, Georgia, Alabama, and Florida using mist nets during mornings and evening flights (Sheldon, 1960), or using spotlights and dip nets at night (McAuley et al., 1993; Rieffenberger and Kletzly, 1966). We attached 4 g, 5 g, or 6.3 g PinPoint GPS Argos transmitters (Lotek Wireless Inc., Newmarket, Ontario, CA) to captured woodcock. Transmitters recorded GPS locations at 12–60 m accuracy and were programmed to record diurnal locations every 1–3 days. Transmitters, bands, and attachment materials never exceeded 4% of a bird’s body weight, and all capture and handling were conducted with methods approved by the [university redacted for double-blind review] Institutional Animal Care and Use Committee (Protocol # A2020-07-01).

We used woodcock location data to identify stopover locations, defined as any place where a migrant bird can land and survive until the next migratory flight (Mehlman et al., 2005). We consider woodcock to be migrating after they have made their first >16.1 km movement in fall or spring, and to complete their migration after they have made their last >16.1 km movement in the respective season. The >16.1 km threshold was chosen as it roughly divides the bimodal distribution of log-standardized step lengths, presumably distinguishing between local- and long-distance movements (Blomberg et al., In review). Because woodcock migrate at night, we considered all diurnal locations between migratory initiation and termination to be stopovers. We grouped successive locations within 3 km into a single stopover, based on our observations that movements <3 km tended to be recursive rather than directional, and removed all but one location from each stopover from the analysis to reduce pseudoreplication and spatial autocorrelation of closely clustered locations. We also generated 2 000 locations randomly distributed throughout Pennsylvania, which we considered pseudoabsence locations.

Table tk. Caption tk.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Breeding** | | **Migratory** | |
|  | *Source* | *Sample size* | *Source* | *Sample size* |
| Presence locations | Surveys | tk | GPS locations | tk |
| Absence/pseudoabsence locations | Surveys (absence) | tk | Randomly generated (pseudoabsence) | 2 000 |

### 2.3 Covariates at multiple spatial scales

We chose explanatory variables with presumed relevance to woodcock habitat associations (McAuley et al., 2020), with suites of variables including land use/land cover, forest successional class, topography, region, and soil moisture (Table 1). We also used landscape composition and configuration metrics, which were calculated at multiple spatial scales (500m, 1km, 5km, and 10km) to incorporate uncertainty in the most relevant scales for predicting woodcock habitat associations during each season. We implemented composition metrics by resampling the National Land Cover Dataset to a 90m resolution, and then calculating the percent of each cover type within 500m, 1km, 5km, and 10km radii for each pixel. For configuration metrics, we used the National Land Cover Dataset to create a binary forest/non-forest layer, which we resampled to 90m resolution, and then calculated each configuration metric within 500m, 1km, 5km, and 10km radii of each pixel. Landscape metrics were calculated using the *landscapemetrics* package in statistical software R, version 4.4.1 (Hesselbarth et al., 2019, R Core Team, 2024).

Table 1. Explanatory variables used to model woodcock distributions in Pennsylvania, USA. Suites indicate conceptual grouping of variables into classes relevant to woodcock occurrence.

|  |  |  |
| --- | --- | --- |
| **Suite** | **Covariate** | **Source** |
| Land use/land cover | Land use/land cover | National Land Cover Dataset (Jin et al., 2019) |
| Forest successional class | Forest successional class | LANDFIRE (USGS and USDA, 2020) |
| Topography | Elevation | USGS (2000) |
| Slope | Derived from elevation |
| Region | EPA level 3 ecoregions | Omernik and Griffith (2014) |
| Soil moisture | Soil drainage | Web soil survey (NRCS, 2021) |
| Topographic wetness index | Derived from elevation (Fink, 2013) |
| Landscape composition  (0.5, 1, 5, and 10km scales) | % Forest | Derived from National Land Cover Dataset using *landscapemetrics* (Hesselbarth et al., 2019) |
| % Agricultural |
| % Developed |
| Landscape configuration  (0.5, 1, 5, and 10km scales) | Aggregation index | Derived from National Land Cover Dataset using *landscapemetrics* (Hesselbarth et al., 2019) |
| Cohesion |
| Edge density |

### 2.4 Species distribution modeling

### 2.4.1 Model design and optimization

We constructed separate species distribution models for migratory and breeding seasons to accommodate differences in habitat associations and data sources. After testing various species distribution modeling frameworks on this dataset (Appendix 1) we elected to use random forest machine learning to create both our breeding and migratory season models (cite random forest paper tk. Breiman 2001?). We conducted this modeling separately for breeding and migratory seasons, although we followed the same progression of steps for each model. We implemented these models using the R package *SDMTune* (Vignali et al., 2020). These models were fit using k-fold cross validation with 5 folds. For the breeding season model, we withheld 1/5th of the presence and absence data from each of 5 training folds and used the withheld data as testing folds to evaluate the model. As the migratory season dataset is effectively presence-only, we followed the presence-only k-fold technique implemented by Vignali et al. (2020), which creates random folds only for presence locations and uses the full set of pseudoabsence locations in each testing and training fold. We then followed the methodology for fitting and optimizing species distribution models outlined in Vignali et al. (2020), which is composed of three steps: variable selection, hyperparameter tuning, and variable reduction.

We conducted variable selection to ensure that our models did not include highly correlated variables, as high correlation can complicate the interpretation of variable importance. After fitting an initial random forest model, we measured the correlation between each variable in the model at each of the presence/absence locations. For each set of variables with a correlation of >0.70 (threshold from citation tk), we performed a jackknife test to determine which of these variables, if removed, results in the model with the highest Area Under the Curve metric (AUC; citation tk). After removing this variable from the model, we repeated the Jackknife tests on remaining sets of correlated variables until no variables remained with a correlation of >0.70. We used this pruned set of variables in all subsequent optimization processes.

As hyperparameter values can impact the predictive accuracy of machine learning methods (cite tk), we optimized the hyperparameters in the model at two points: first, after the variable selection step, and second, after the variable reduction step (below). We used the genetic algorithm derived in citation tk to find the values of three hyperparameters which resulted in the highest AUC for each model. These hyperparameters included the number of trees in the random forest model (range: 100–3000), the number of variables sampled at each split (1–8), and the minimum size of terminal nodes (1–10).

Finally, we removed variables from the model which had little importance to the model, with the aim of reducing model complexity while maintaining the model’s predictive capacity. We chose to remove variables with a percent contribution less than 2%, as suggested by Vignali et al. (2020), on the condition that the removal of that variable not result in an AUC reduction in the resulting model. The resulting breeding and migratory season models, including only non-correlated variables with high importance and twice-tuned hyperparameters, were finally tested for predictive accuracy using the AUC and the true skill statistic (TSS; cite tk). Predictive metrics were generated using the previously withheld testing dataset for each of the 5 folds, to ensure that our evaluation metrics were not biased by the model fitting or optimization processes.

### 2.4.2 Creating predictive layers

Using the final, optimized models, we created predictive layers for woodcock habitat suitability throughout Pennsylvania in both the breeding and migratory seasons. We normalized the predictive layer for each season on a percentile scale, indicating whether a given pixel had a greater likelihood of woodcock occupancy than the corresponding percentage of other pixels in the state; for example, a value of 0.65 indicates that the pixel contains habitat that is more suitable than 65% of other pixels statewide. We termed these habitat suitability layers, with habitat suitability defined as a bounded, continuous index representing the degree to which a given site possesses the habitat components required for species occupancy (U.S. Fish and Wildlife Service, 1996). We occasionally refer to these habitat suitability layers as representing habitat distribution, which we define as the geographic distribution of habitat on the landscape, or as a representation of where areas of greater suitability occur or are absent.

### 2.5 Analysis of covariate relationships

Random forest techniques do not provide easily interpretable covariate relationships, leaving the user to determine how covariates might influence the outcomes of the model (Breiman, 2001). While we were not interested in exploring woodcock-habitat relationships per se, we nevertheless wanted to understand how environmental variables contributed to model predictions. We also sought to highlight regional differences in the distribution of breeding and migratory habitat among ecoregions. To depict these differences, we sampled covariate values, ecoregion type, and model-predicted suitability at 10 000 randomly distributed points throughout Pennsylvania. We used hex plots to visualize trends between covariates and predictions for each season, and visualized variation among each EPA level 3 ecoregion in Pennsylvania using box-and-whisker plots.

### 2.6 Comparative distribution of seasonal habitat

We used two metrics to evaluate cross-seasonal transferability between breeding and migratory season species distribution models, to better understand the utility of a multi-season modeling approach. The first was a Pearson correlation coefficient between the breeding and migratory season layers, calculated using the R package *terra* (Cohen et al., 2009; Hijmans, 2022), which measured the correlation between breeding and migratory suitability on the scale of individual pixels (90 m). The second metric measured the Pearson correlation between the sum of breeding and migratory pixel values within each gameland, illustrating the co-occurrence of breeding and migratory habitat at the scale of the average state gameland (1 992 ha).

### 2.7 Shiny application

To allow our results to be interpreted in a management context, we co-developed a Shiny application with the Pennsylvania Game Commission called W-PAST that included our predictive layers and a series of derived metrics for prioritizing management areas. Recognizing that both breeding and migratory season habitat were likely to be relevant to management strategies, we allowed managers to access blended predictive layers based on their comparative prioritization of breeding and migratory habitat (*e.g.,* 70% breeding and 30% migratory). The application is available at https://woodcock.shinyapps.io/W-PAST2/, and detailed descriptions of the derived metrics and blended predictive layers included in the application are in Appendix 2.

## 3 Results

### 3.1 Model fit

At the conclusion of the model optimization process, our models used a reduced subset of variables which were most relevant for predicting woodcock occupancy in the respective season. In the breeding season, the final model included tk, tk, tk, and tk covariates. In the migratory season, the final model included tk, tk, tk, and tk covariates. We evaluated the predictive accuracy of both models using AUC and TSS metrics. The breeding season model had an AUC of tk and a TSS of tk, while the migratory season model had an AUC of tk and a TSS of tk.

### 3.2 Analysis of covariate relationships

Variable importance (also add to methods)

Covariate relationships of most important variables

### 3.3 Comparative distribution of seasonal habitat

Breeding season habitat was not evenly distributed among ecoregions (Fig. 5A), with mean habitat suitability values ranging from 22.9–86.0%. Migratory habitat was more evenly distributed among ecoregions, with mean habitat suitability values ranging from 46.5–87.5%. Most of the difference between the distribution of migratory and breeding season habitat was in the Northern Piedmont, Middle Atlantic Coastal Plain, and Central Appalachians ecoregions, which had mean breeding season habitat suitability values of <30% and mean migratory season habitat suitability values of >60% (Fig. 5B). Breeding and migratory habitat rarely co-occurred at a pixel level, with a Pearson correlation coefficient of 0.15 between the breeding and migratory season predictive layers. Breeding and migratory habitat were slightly more likely to co-occur on gamelands, with a Pearson correlation coefficient of 0.39 between the total breeding habitat and total migratory habitat provided by gamelands.

## 4 Discussion

**Declaration of competing interest**

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

**Supplementary material**

Appendix 1. Pilot evaluation of species distribution modeling frameworks.

Appendix 2. Design of W-PAST.

**Data availability**

The code used in this analysis is publicly available at github.com/EWMRC/AMWO-seasonal-weighted-SDM. Data are available from the corresponding author upon reasonable request.

**References**

TK