Seasonal occurrence and abundance of dabbling ducks across the continental United States: Joint spatio-temporal modelling for the Genus Anas

John M. Humphreys\textsuperscript{1,2} | Jennifer L. Murrow\textsuperscript{3} | Jeffery D. Sullivan\textsuperscript{4} | Diann J. Prosser\textsuperscript{2}

\textsuperscript{1}Michigan State University, East Lansing, Michigan, USA
\textsuperscript{2}U.S. Geological Survey, Patuxent Wildlife Research Center, Laurel, Maryland, USA
\textsuperscript{3}University of Maryland, College Park, Maryland, USA
\textsuperscript{4}Natural Systems Analysts, Winter Park, Florida, USA

Abstract

Aim: Estimating the distribution and abundance of wildlife is an essential task in species conservation, wildlife management and habitat prioritization. Although a host of methods and tools have been proposed to accomplish this undertaking, several challenges remain in accurately forecasting occurrence and abundance for highly mobile species. Exhibiting extensive geographic ranges with seasonally varying local occupancy, migratory ducks are exemplary highly mobile species and are foci for waterfowl conservation and management globally. With the goal of informing species conservation and management, our aim was to leverage citizen science data to estimate occurrence and relative abundance for ten dabbling duck species across the continental United States.

Location: Conterminous United States.

Methods: We applied spatially and temporally explicit Bayesian hierarchical modelling to jointly estimate season-specific occurrence and relative abundance for ten dabbling duck species in the Genus Anas. Our conditionally dependent model design enabled relative abundance estimates to be informed by occurrence probability while accounting for cumulative spatial and temporal errors across shared model components.

Results: Outcomes suggest that although dabbling duck distributions show little inter-annual variability at the continental scale, local relative abundances may differ year to year. Commensurate with being highly mobile migratory species, estimates indicate considerable intra-annual variation with occurrence probability, local abundance and habitat preferences differing by season and species.

Main conclusions: Our approach offers a powerful and flexible framework for quantifying intra-/inter-annual duck occurrence and relative abundance while accounting for spatial, temporal and data collection biases. We believe that model produced maps can be applied to inform waterfowl conservation and management throughout the continental United States.

\textbf{Keywords}

abundance, waterfowl, Bayesian hierarchical, biogeography, distribution, ducks, macroecology, North America, spatio-temporal
1 | INTRODUCTION

Species distribution models are valuable tools in waterfowl ecology and have been broadly applied in assessing assessing waterfowl vulnerability to climate change (Reese & Skagen, 2017; Sofaer et al., 2016; Steen, Skagen, & Noon, 2014), delimiting endangered waterfowl ranges (Pickett et al., 2018; Si et al., 2018; Zeng et al., 2015) and in waterfowl habitat conservation prioritization (Barker, Cumming, & Darveau, 2014; Johnston et al., 2015; Runge, Martin, Possingham, Willis, & Fuller, 2014). Species distribution models have also been incorporated into a host of epidemiological studies, including those aimed at appraising the role of migratory waterfowl in global transmission of highly pathogenic avian influenza among wild bird populations and to domestic poultry (Belkhiria, Hijmans, Boyce, Crossley, & Martínez-López, 2018; Martin et al., 2011; Prosser et al., 2018, 2013, 2009; Russell, 2016; Takekawa et al., 2010). Despite the demonstrated ability of species distribution models to correlate species presence or wildlife disease to underlying environmental factors (Elith & Leathwick, 2009; Guisan & Thuiller, 2000), several challenges remain in accurately forecasting species occurrence and abundance within a spatio-temporal context (Franklin, 2018; Martnez-Minaya, Cameletti, Conesa, & Pennino, 2018; Paradinas, Conesa, López-Quílez, & Bellido, 2017). Species ranges are not static (Gaston, 2003) and avian species in particular exhibit substantial spatial and temporal dynamism (Fink, Damoulas, & Dave, 2013; Runge et al., 2014). As a consequence, modelling techniques that assume static environmental preferences across space/time may not be appropriate for species that exhibit migratory behaviour. Likewise, approaches intended to derive abundance estimates from presence-only observations may assume a linear correlation of presence to abundance when in fact the relationship varies across multiple spatial and temporal scales (Johnston et al., 2015; Oliver et al., 2012; Yin & He, 2014).

To address some of the outstanding concerns with modelling highly mobile species, we employed spatially and temporally explicit techniques to estimate occurrence and relative abundance for ten migratory waterfowl species. Spatially explicit models are differentiated from those that are implicit in that the explicit variety considers the relative positions of individuals (without aggregating occurrences to grid or raster cells) and addresses spatial autocorrelation and spatial dependency as intrinsic parts of the model rather than through iterative fitting or the post hoc adjustment of model outputs (Hui, McGeoch, & Warren, 2006; Perry et al., 2002; Veldtman & McGeoch, 2004; Wiens, 2000). The spatial distribution of a species encodes information about the underlying ecological processes that gave rise to it (Hurtt & Pacala, 1995; Velázquez, Martínez, Getzin, Moloney, & Wiegard, 2016); therefore, it is essential that spatially explicit methods be applied to reveal species ecology through analysis of geography. Indeed, neglecting spatial autocorrelation may result in parameter estimates with smaller uncertainty than is warranted and may produce predictions with considerable error (Dormann, 2007; Elsner et al., 2016; Hoeting, 2009; Renner et al., 2015). In a comparable manner to spatial autocorrelation, temporal autocorrelation can also give rise to model error (Cressie & Wikle, 2011; Paradinas et al., 2017), which may be present whenever sampling has occurred repeatedly, at regular intervals, or over extended periods of time. For all of these reasons, spatially and temporally explicit methods offer distinct advantages over non-explicit alternatives, such as maximum entropy and machine-learning techniques (e.g., random forests, boosted regression trees), particularly when the focus of a study includes species or diseases that vary across time/space (Martinez-Minaya et al., 2018; Paradinas et al., 2017; Runge et al., 2014).

Our current proposal is to combine citizen science data with advanced spatio-temporal modelling techniques to jointly estimate occurrence probability and relative abundance for ten dabbling duck species in the Genus Anas. In addition to their contribution to ecosystem function (Bauer & Hoye, 2014; Figuerola, Charalambidou, Santamaria, & Green, 2010), economic importance (USFWS, 2015) and cultural benefit to humans (Schummer, Coluucy, Mitchell, & Olsen, 2017; Soulliere et al., 2007), several dabbling ducks species are of interest due to possible roles in avian influenza transmission (Van Den Brand et al., 2018; Papp et al., 2017; Ramey, DeLiberto, Berhane, Swanye, & Stallknecht, 2018; Ramey et al., 2014, 2016). Choosing the continental United States as our study area, our primary goals include (a) leveraging freely available citizen science data to estimate seasonal species occurrence and abundance, (b) quantifying inter-annual variation in duck environmental associations and (c) demonstrating a powerful and flexible technique to account for errors arising from both spatial and temporal autocorrelation, as well as the biases inherent to data collected through citizen science efforts lacking standardized sampling protocols.

2 | METHODS

2.1 | Species data

Data for ten species of dabbling duck (Table 1) were secured from the Cornell Lab of Ornithology and National Audubon Society eBird project (Sullivan et al., 2009). In total, more than 1 million records were initially obtained from the eBird database for the three-year

| TABLE 1 | Focal dabbling duck species |
|--------|-----------------------------|
| **Abbreviation** | **Common name** | **Scientific name** |
| ABDU | American black duck | Anas rubripes |
| AMWI | American wigeon | Anas americana |
| BWTE | Blue-winged teal | Anas discors |
| CITE | Cinnamon teal | Anas cyanoptera |
| GADW | Gadwall | Anas strepera |
| AGWT | Green-winged teal | Anas carolinensis |
| MALL | Mallard | Anas platyrhynchos |
| MODU | Mottled duck | Anas fulvigula |
| NOPI | Northern pintail | Anas acuta |
| NSHO | Northern shoveller | Anas clypeata |
period 2015–2017. Following pre-processing (see Section 2.3), 999,377 occurrence records were incorporated for spatio-temporal modelling. The number of records per species ranged between 10,039 and 190,309.

2.2 Study domain and environmental data

The study domain encompassed the entire continental United States and exhibited an area exceeding 7.7 million km$^2$. Environmental data (see Appendix S1: Supplement A) included two spatial datasets used to characterize land cover and dominant agricultural land use across the study domain: the 2011 National Land Cover Database (Homer et al., 2011) and the 2017 Cropland Data Layer (CropScape, 2017). A “distance to water” covariate providing the Euclidean distance (km) to the nearest surface water class (inclusive of wetlands) within the National Land Cover Database (NLCD) was also calculated using the spatstat R package (Baddeley & Turner, 2005). Both the NLCD and Cropland datasets were categorical rasters at 30-m resolution. A topographical roughness index, a topographical wetness index and a continentality raster (30 arc-seconds continuous rasters) from the ENVIREM dataset (Title & Bemmels, 2017) were adopted to model topographical heterogeneity across the study domain. Indices for topographical roughness and wetness were intended to help differentiate wetlands and surface waters from other landscape components while continentality (average coldest month temperature subtracted from average warmest month temperature in °C) aided in distinguishing between terrestrial and marine climate regimes (Sayre, Comer, Warner, & Cress, 2009). Spatial and temporal climate variability were captured through inclusion of minimum monthly temperature and monthly precipitation rasters from the PRISM Climate Group at a 4-km resolution (PRISM, 2018). Additionally, a continuous raster (30 m) giving human population density (CIESIN, 2016) was used as one of two covariates to help control for possible observer bias associated with the reporting of duck occurrence locations (see Section 2.3).

2.3 Preliminary analysis

Considering our point-referenced observation data and interest in quantifying both occurrence and abundance across a geographically expansive study domain, we adopted the stochastic partial differential equation (SPDE) approach (Lindgren, Rue, & Lindström, 2011), such that our joint modelling framework incorporated a two-tiered conditionally dependent structure with shared components. For clarification, we use the term “joint model” to refer to a single model consisting of multiple parts, levels or submodels that might typically be implemented as separate or individual analyses. The first level of our two-part joint model can be viewed as a submodel for duck occurrence probability, whereas its second level is a submodel for relative abundance. To avoid repeated and vague use of the word “model,” the terms “tier” or “stage” are used to reference either the joint model’s occurrence probability or relative abundance level specifically. Although our model exhibits a two-part structure that is comparable to that associated with occupancy models, which jointly estimate detection probability (observation process) with species occurrence probability (Johnston, Fink, Hochachka, & Kelling, 2018; MacKenzie et al., 2006), we emphasize that our model does not directly estimate detection. Rather than combining detection and occurrence probability, we propose a joint modelling approach that joins species occurrence probability with estimated relative abundance, such that species occurrence modelled in the first stage is passed as a covariate to the model’s second stage for improved abundance estimation. Stated differently, the model’s second stage (relative abundance) is conditional on the first stage (occurrence probability). The SPDE method was selected because of its ability to ease processing time associated with the spatial random effects used to quantify spatial autocorrelation and other spatial error in the joint model. Rather than using spatial grids or polygons, the SPDE method enables spatial random effects to be estimated across a triangulated mesh, which greatly reduces the time needed to run the models (Lindgren et al., 2011).

Prior to constructing species distribution models, pre-processing and exploratory analyses were performed to clean and better organize duck occurrence and environmental data for model fitting. Steps were also taken to avoid covariate multicollinearity and similar statistical issues commonly encountered during species distribution modelling.

A major advantage of the eBird database is that criteria describing the field survey method used to collect data (i.e., point, transect and aerial survey) are included with each record, as are the number of participating surveyors, the duration of surveys and other essential information (Fink et al., 2010; Sullivan et al., 2014). For the current study, records obtained from the eBird database were restricted to the most recent three-year period (2015–2017), included only complete checklists (i.e., all species present were documented and reported) and were filtered to remove records not yet reviewed by a participating member of the eBird organization (see Sullivan et al., 2009 for review standards). To reduce the range of values used in estimating abundance, all reported counts were truncated at 100 birds. Truncating counts were accomplished by recoding bird counts greater than 100 as 100. To avoid arbitrariness in choosing record filtering criteria and in selecting truncating thresholds, we followed comparable strategies to those reported in the literature (Cardador, Carrete, Gallardo, & Tella, 2016; Fink et al., 2010; Johnston et al., 2015; McCormack, Zellner, & Knowles, 2010; Zipkin et al., 2017). Two actions were also taken to reduce the observer bias that is often inherent to citizen science data (Dormann, 2007; Johnston et al., 2018; Kadmon, Farber, & Danin, 2003; Renner et al., 2015). Firstly, an observation effort covariate was constructed by multiplying each record’s reported number of observers by the total observation period duration (minutes) and then scaling the product to the range from zero to one. The resulting covariate reflected sampling effort (observer hours) with values close to zero indicating relatively minimal observation effort and a value of one signifying the maximum. Secondly, human population density was included as an independent variable with
all models to account for the possibility that duck observations were proportional to the number of persons residing in the vicinity (Kadmon, Farber, & Danin, 2004). Beyond providing a means of accounting for observer bias during model fitting, the effort and human population variables facilitated prediction under assumptions of maximal observer effort and a proportionate collection of observations across the landscape regardless of human accessibility. That is, these variables allowed us to model both the observations and the observation process (Besbes & Morgan, 2014; Gould & Nichols, 1998; Humphreys, Elsner, Jagger, & Pau, 2017; Zipkin et al., 2017).

Several Gaussian random fields (i.e., spatial random effects) were constructed to quantify the spatial relationships among duck observations in our dataset. The random fields served as error terms to measure spatial autocorrelation and other uncertainty associated with the observed point patterns within and between each tier of the model. To create the Gaussian random fields, a three-dimensional, triangulated mesh scaled to represent the curvature of the Earth’s surface was constructed using guidelines provided by Lindgren et al. (2011). The mesh intersections (vertices) also served as quadrature locations to support numerical integration for some model components. We avoid using the term “pseudo-absence” as quadrature locations were intended to characterize the background environment (Renner et al., 2015) and should not be interpreted as locations devoid of ducks. Tools available in the r-INLA package for R (Lindgren & Rue, 2015) facilitated mesh construction by providing commands to control the distance between mesh vertices, the maximum length of triangle edges, the minimum permissible angle at edge intersections and the length of the mesh outer extension. The final mesh included 4,953 location-specific vertices arranged in a regular, but not uniform, pattern across the study domain.

To organize data, a spatial point dataset was constructed by combining duck occurrence locations with a separate set created from the mesh vertices. Using the raster R package (Hijmans, 2016), location-specific values for the variables listed in Section 2.2 were extracted to the combined spatial points dataset and then scaled and centred for ease of computation and later interpretation. In the case of time-stamped temperature and precipitation data, extractions were undertaken to match monthly climate estimates to the corresponding year and month that individual ducks were observed. The resulting dataset included an attribute table with rows corresponding to the locations of duck occurrences or quadrature points and columns recording the coincident covariate for each geographic location. To enable season-specific modelling and comparison, reported duck observation dates were used to disaggregate the full spatial point dataset into spring (16 February –31 March), breeding (01 April –31 August), fall (01 September –30 November) and winter (01 December –15 February) subsets. In total, forty spatial point datasets were created and modelled; four season-specific models for each of ten focal duck species.

To gauge the potential for multicollinearity among covariates, collinearity diagnostics for independent variables as described by Belsley, Kuh, and Welsch (1980) were estimated for each duck species and season using the perturb R package (Hendricks & Pelzer, 2004). This method adds random noise to independent variables and then calculates variance decomposition proportions and a condition index to assess collinearity. Evaluation of model variables resulted in condition index scores for each species and season combination ranging between 11.2 and 19.8. These scores fell well below the Belsley et al. (1980) proposed threshold of 30 giving us confidence that collinearity would not bias model results.

As a final step before model fitting, a minimum of 25% of occurrence points were randomly selected from each duck species and season for model testing and evaluation. In the case of the abundant mallard (MALL), additional points were randomly selected for evaluation to reduce computer processing time and improve performance (see, Section 2.5). The final column of Table 2 provides the number of records used to fit and evaluate models for each focal species.

### Table 2 Occurrence probability evaluation summary

| Species | Sensitivity (CI) | Specificity (CI) | AUC (CI) | TSS (CI) | Train/Test |
|---------|-----------------|-----------------|----------|---------|------------|
| ABDU    | 0.93 (0.89–0.98)| 0.96 (0.95–0.97)| 0.99 (0.99–0.99) | 0.89 (0.86–0.92) | 29,890/74,727 |
| AMWI    | 0.89 (0.85–0.93)| 0.91 (0.88–0.94)| 0.96 (0.95–0.97) | 0.80 (0.76–0.84) | 55,757/13,939 |
| BWTE    | 0.85 (0.75–0.94)| 0.92 (0.81–1.00)| 0.96 (0.93–0.99) | 0.76 (0.67–0.86) | 45,590/11,398 |
| CINT    | 0.85 (0.78–0.93)| 0.98 (0.95–1.00)| 0.99 (0.98–0.99) | 0.83 (0.77–0.89) | 15,491/3,873 |
| GADW    | 0.91 (0.86–0.95)| 0.88 (0.84–0.93)| 0.96 (0.95–0.97) | 0.79 (0.76–0.81) | 69,874/17,469 |
| AGWT    | 0.87 (0.87–0.91)| 0.89 (0.86–0.93)| 0.96 (0.95–0.97) | 0.78 (0.75–0.81) | 55,558/13,890 |
| MALL    | 0.96 (0.93–0.99)| 0.81 (0.66–0.92)| 0.95 (0.94–0.97) | 0.75 (0.65–0.85) | 190,309/207,315 |
| MODU    | 0.92 (0.83–1.00)| 0.98 (0.98–0.99)| 1.00 (0.99–1.00) | 0.91 (0.81–1.00) | 10,039/2,510 |
| NOPI    | 0.84 (0.79–0.90)| 0.93 (0.91–0.96)| 0.96 (0.96–0.97) | 0.78 (0.74–0.82) | 35,650/8,913 |
| NSHO    | 0.90 (0.87–0.93)| 0.90 (0.86–0.94)| 0.96 (0.95–0.97) | 0.79 (0.76–0.83) | 61,646/15,412 |

Note: Sensitivity, specificity and the area under the receiver operating characteristic curve (AUC) are scaled from 0 to 1, and the True Skill Statistic (TSS) score ranges from –1 to 1. Higher scores signify improved performance for all metrics. Train/Test column provides the count of records used to fit and evaluate models, respectively. Parenthetical values provide the 95% confidence interval. Seasonal models are pooled by species.
2.4 | Statistical model

The joint model’s first tier estimated duck occurrence probability and was constructed as a spatially and temporally explicit point model with species presence/absence (0, 1) as response $y_i$. Incorporating a Binomial likelihood, the first tier can be expressed as,

$$y_{i,(t)} \sim \text{Binomial}(\alpha_{i,t})$$

$$\log it(\alpha_{i,t}) = \beta_0 + \beta_1 \text{Ppt}_{i,(t)} + \beta_2 \text{Con} t_{i,(t)}$$

$$+ \beta_3 \text{Trew}_{i,(t)} + \beta_4 \text{Tri}_{i,(t)} + \beta_5 \text{Pop}_{i,(t)}$$

$$+ \beta_6 \cdot \text{fCarp}_{i,(t)} \cdot \text{Crop}_{i,(t)} + \beta_7 \cdot \text{fLC}_{i,(t)} \cdot \text{Cs}_{i,(t)}$$

$$+ \beta_8 \cdot \text{fWd}_{i,(t)} \cdot \text{Wds}_{i,(t)} + \beta_9 \cdot \text{fTM}_{i,(t)} \cdot \text{mT}_{i,(t),0} + \text{W}_{2}(s,t),$$

where $\beta_0$ is the intercept, $\beta_1$ is the coefficient of annual precipitation (mm/year), $\beta_2$ is the coefficient of continentality (°C), $\beta_3$ is the coefficient of topographical wetness index, $\beta_4$ is the coefficient of topographical roughness index, and $\beta_5$ is the coefficient of human population density (persons/km²). The function $\beta_6 \cdot \text{fCarp}_{i,(t)} \cdot \text{Crop}_{i,(t)}$ describes an independent and identically distributed (iid) random effect to account for agricultural crops from the Cropland database. The function $\beta_7 \cdot \text{fLC}_{i,(t)} \cdot \text{Cs}_{i,(t)}$ describes a similar iid to evaluate land cover. The functions $\beta_8 \cdot \text{fWd}_{i,(t)} \cdot \text{Wds}_{i,(t)}$ and $\beta_9 \cdot \text{fTM}_{i,(t)} \cdot \text{mT}_{i,(t),0}$ respectively, model distance to water (Wds, km) and minimum temperature (mT, °C) as order-one random walks. Comparable to regression splines, random walks allow for smooth, non-linear fitting of these covariates, which were suspected of having a dynamic or highly variable response due to duck water dependence and seasonal migratory behaviour, respectively. Lastly, $\text{W}_{2}(s,t)$ provides a spatially structured random effect for estimation of spatial error. The model considers correlated spatial processes at locations $s$ ($s = 1, 2, 3, \ldots, n$) under time steps $t$ ($t = 1, 2, 3, \ldots, n$) for the years 2015–2017 with precipitation (Ppt) and minimum temperature (mT) varying in both time/space and other environmental covariates exhibiting only spatial variation.

The model’s second-tier incorporated observed bird counts as the response variable and had the same covariates as the first-tier linear predictor (Equation 1), except that the second-tier also included the observer effort variable described during preliminary analysis (Section 2.3) and a shared spatial field created by copying the spatial effect $\text{W}_{1}(s,t)$ from the first tier (Equation 1). Rather than coding unobserved locations associated with mesh vertices as zero as done in tier-one, the vertex locations were set to "NA" (missing data) in the second-tier resulting in a zero-truncated response vector. Adopting a Poisson likelihood with a mean $\mu$ and log link function, the second-tier process was,

$$\log (\mu_{i,t}) = \beta_0 + \beta_1 \text{Ppt}_{i,(t)} + \beta_2 \text{Con} t_{i,(t)}$$

$$+ \beta_3 \text{Trew}_{i,(t)} + \beta_4 \text{Tri}_{i,(t)} + \beta_5 \text{Pop}_{i,(t)}$$

$$+ \beta_6 \cdot \text{fCarp}_{i,(t)} \cdot \text{Crop}_{i,(t)} + \beta_7 \cdot \text{fLC}_{i,(t)} \cdot \text{Cs}_{i,(t)}$$

$$+ \beta_8 \cdot \text{fWd}_{i,(t)} \cdot \text{Wds}_{i,(t)} + \beta_9 \cdot \text{fTM}_{i,(t)} \cdot \text{mT}_{i,(t),0} + \text{W}_{2}(s,t) + \text{W}_{3}(s,t),$$

where $\beta_0$ is the intercept and $\beta_1$ through $\beta_9$ provide coefficients for the same covariates described under the first tier. Observer effort, which varies by location and year, is quantified by $\text{Eff}_{i,(t)}$. $\text{W}_{1}(s,t)$ gives the spatial effect for second-tier observations, and $\text{W}_{2}(s,t)$ is a copy of $\text{W}_{1}(s,t)$ from the first tier that is shared with the second-tier. The shared field controls for correlation between our two model tiers (Jones-Todd, Swallow, Illian, & Toms, 2017) while allowing predicted occurrence probability to inform duck abundance. This general model specification described above has also been proposed for dealing with zero inflated data, which is a common issue with species distribution and epidemiological studies (Arab, 2015; Paradinas et al., 2017; Sadykova et al., 2017; Serra, Saez, Juan, Varga, & Mateu, 2013).

Because modelling was conducted under a Bayesian framework, prior distributions were required for model parameters. Penalized complexity (PC) priors (Fuglstad, Simpson, Lindgren, & Rue, 2018; Simpson, Illian, Lindgren, Sørbye, & Rue, 2015) were selected for the three spatial random effects $\text{W}_{1}$ based on the geometry of the mesh used to define our study domain. We provided estimates for the lower tail quantile and probability of the spatial range ($\nu_0 = 1$ and $\rho = 0.5$) as well as the upper tail quantile and probability for the standard deviation ($\sigma_0 = 1$ and $\alpha = 0.01$). PC priors were also utilized for the non-linear effects modelled as functions of distance to water ($f_{\text{Wd}}, Wds$) and minimum temperature ($f_{\text{mT}}, \text{mT}$). Distance to water priors were set as $\mu = 0.5$ and $\alpha = 0.0001$ and minimum temperature with $\mu = 1$ and $\alpha = 0.0001$. Our priors and multivariate likelihoods were then combined to obtain the posterior distributions using the INLA method (Rue, Martino, & Chopin, 2009). INLA is a computationally efficient alternative to Markov chain Monte Carlo (MCMC) that can be applied to fit latent Gaussian models. Because INLA uses deterministic functions to approximate posterior distributions, it avoids the convergence and mixing problems associated with MCMC and can be used to model complex datasets that exhibit spatial and temporal structure (Rue et al., 2009).

2.5 | Model fitting and evaluation

Model fitting and evaluation were performed as an iterative process in which models were constructed for each focal species and season and then validated against the testing dataset randomly partitioned during pre-processing. As occurrence probability comparison metrics, we calculated model sensitivity (proportion of correctly predicted presences), specificity (proportion of correctly predicted absences) and the area under the receiver operating characteristic curve (AUC) using the PresenceAbsence package (Freeman & Moisen, 2008) for R. Because the AUC may not be appropriate for comparison between species distribution models (Lobo, Jimenez-Valverde, & Real, 2008; Shabani, Kumar, & Ahmadi, 2016), we also determined the True Skill Statistic (TSS) by subtracting a value of one from the sum of the sensitivity and specificity estimates (Allouche, Tsoar, & Kadmon, 2006). Although the total number of presences available in our dataset differed by species, season and year, presence counts were largely balanced by background points drawn from the triangulated mesh. That stated, we recognized that a 0.5 threshold
may not be optimal (Guisan & Theurillat, 2000) for all forty species-season combinations that were modelled and likewise acknowledge that a variety of other threshold selection criteria have been proposed (Bean, Stafford, & Brashares, 2012; Cramer, 2003; Liu, Berry, Dawson, & Pearson, 2005; Raxworthy, Ingram, Rabibisoa, & Pearson, 2007).

To assess the accuracy of duck abundance estimates, we calculated the Spearman rank correlation (referred to as correlation or “Cor” here forward) between model predicted abundance and the number of observed birds reported with eBird records as well as the correlation between predicted abundance and duck presence. In the case of mallard (MALL) models, the initial correlation between predicted and observed abundance fell below 0.50 causing us to re-run the models after reducing the MALL dataset. As described in Section 2.3, this resulted in a greater proportion of points being retained for model testing.

### RESULTS

Table 2 provides a model evaluation summary for predicted occurrence probability. The mean average TSS across all duck species and seasons exceeded 0.80 with MALL exhibiting the lowest TSS (0.75 [0.65, 0.85] 95% confidence interval) and MODU showing the highest TSS (0.91 [0.81, 1.00] 95% confidence interval). Accuracy assessments for model estimated abundance are given in Table 3. The mean average correlation for all species and seasons was over 0.65 when compared to eBird reported bird counts (Table 3[Cor-A]) and approximately 0.68 relative to duck presence/absence (Table 3[Cor-B]). As with occurrence probability, MALL exhibited the lowest accuracy statistic with an average correlation of 0.51 across pooled seasons.

### Abundance evaluation summary

| Species | Cor-A (Cl) | Cor-B (Cl) |
|---------|------------|------------|
| ABDU    | 0.74 (0.66–0.81) | 0.74 (0.66–0.82) |
| AMWI    | 0.77 (0.74–0.79) | 0.78 (0.75–0.81) |
| BWTE    | 0.67 (0.59–0.75) | 0.68 (0.58–0.78) |
| CINT    | 0.58 (0.38–0.78) | 0.58 (0.37–0.79) |
| GADW    | 0.75 (0.75–0.75) | 0.78 (0.76–0.80) |
| AGWT    | 0.75 (0.74–0.77) | 0.77 (0.76–0.79) |
| MALL    | 0.51 (0.36–0.69) | 0.55 (0.41–0.61) |
| MODU    | 0.53 (0.44–0.63) | 0.54 (0.44–0.63) |
| NOPI    | 0.72 (0.67–0.77) | 0.72 (0.67–0.77) |
| NSHO    | 0.76 (0.74–0.78) | 0.78 (0.75–0.81) |

Note: Cor-A lists the Spearman’s rank correlation (Cor) estimate for predicted species density (birds/km²) to eBird reported bird counts. Cor-B gives the correlation for predicted species density to reported bird occurrences (presence/absence). Parenthetical values provide the 95% confidence interval. Seasonal models are pooled by species.

By comparison, predicted abundance correlation to eBird counts was highest for AMWI (0.77 [0.74, 0.79] 95% confidence interval). Overall, accuracy statistics for occurrence probability appeared more consistent across species than those for abundance. More simply, greater between species variability is apparent in the Cor-A measure than in the TSS, sensitivity, specificity or AUC metrics. This disparity was also evident in the degree of inter-annual temporal correlation exhibited by the occurrence and abundance tiers of the joint model. The degree of temporal correlation for the occurrence probability tier was greater than 0.99 for all species and seasons, whereas inter-annual correlation for the abundance level ranged between 0.55 and 0.98 with a mean average of 0.79. Temporal correlation...
estimates for occurrence probability and abundance are provided for all models in data Appendix S1: Supplement A.

Plotting predicted NSHO occurrence against predicted log-scaled abundance (Figure 1) illustrates that although the two estimates are highly correlated (Table 3[Cor-B]), the relationship is non-linear with a rapid increase in abundance at locations having a greater than a 90% chance of occurrence. Note that the upper tail of each plot shows exponential increase.

Maps depicting predicted occurrence probability and estimated relative abundance for all focal duck species are included with data Appendix S1: Supplement B data. For demonstrative purposes, Figure 2 displays model predicted occurrence probability for the northern shoveller (NSHO) and Figure 3 maps NSHO seasonal relative abundance. Considerable seasonal variability is apparent for both occurrence and abundance. To illustrate seasonal variation at a finer scale, Figure 4 compares NSHO occurrence probability for the breeding and winter seasons across the Prairie Pothole Region. The Prairie Pothole Region is recognized as being of critical importance to NSHO and a diverse range of other waterfowl (Fairbairn & Dinsmore, 2001; Klett, Shaffer, & Douglas, 1988; Reynolds, Shaffer, Renner, Newton, & Batt, 2001).

Although duck association with land cover varied little throughout the year, seasonal variability was evident for other model covariates (see Appendix S1: Supplement A). The sampling effort and human population density covariates used to account for data collection bias demonstrated robust performance across species as judged by the magnitudes of posterior densities and credible intervals. Using NSHO as an example, sampling effort and human population density were particularly strong for the spring season, during which as many as 33 additional birds/km² can be identified (exponentiating the mean posterior densities) under maximum sampling effort. Approximately one additional bird was reported for each doubling of local human population density beyond the modelled domain-wide average of 40.2 persons/km².

4 | DISCUSSION

We presented a spatially and temporally explicit species distribution model for joint estimation of occurrence and abundance across the continental United States. With a focus towards migratory dabbling ducks, our primary objectives included undertaking a statistically rigorous treatment of potentially biased citizen science data, quantifying possible time-varying changes to duck environmental associations and demonstrating a robust and flexible method to account for spatial and temporal autocorrelation. Because point process models offer ecologically relevant information pertaining to both the observed pattern and underlying process (Fithian, Elith, Hastie, & Keith, 2015; Illian et al., 2013; Simpson et al., 2015; Velázquez et al., 2016), we utilized point-specific observation data from the eBird database (Sullivan et al., 2009) and opted to implement a geostatistical approach under a Bayesian perspective (Lindgren & Rue, 2015; Lindgren et al., 2011; Rue et al., 2009). Observation data from the eBird database were found subject to observer bias (Link & Sauer, 1998), and accordingly, the effort and human population density covariates developed to control for sampling bias were shown to be important for all species and the majority of seasons (see Appendix S1: Supplement A). Though biased, we concur with Fink et al. (2010) that a major benefit of using the eBird database is that it provides observer-specific criteria to aid in quantifying survey effort for each record. It is also worth noting that our survey effort covariate had a stronger model influence than did the human population density variable and, like the environmental covariates, the strength and importance of both varied by season.

All modelled climate and topographical covariates showed seasonal variability. The lowest degree of seasonal change was seen in the land cover and agricultural variables derived from the NLCD and CropScape with the greatest being apparent in the temperature and water proximity covariates. The CropScape covariate did suggest the possibility that dabbling duck selection of rice, alfalfa, oats and other cultivars may change throughout the year (see Appendix S1:

![FIGURE 2](image-url)  
NSHO predicted occurrence. Figures depict predicted occurrence probability for the (a) spring, (b) breeding, (c) fall and (d) winter seasons. Warm colours (reds) highlight regions with greater than a 50% chance of occurrence, and cool colours (blues) indicate occurrence probabilities below 50%.
Supplement A); however, we caveat this statement by stressing that associations with different crops may also be artefacts of migration paths being coincident with agricultural regions, planting schedules or water availability connected to seasonal flooding. Analysis beyond correlative species distribution modelling would be necessary to fully evaluate any hypothesized crop preferences. Regardless, analyses presented in the current study are sufficient to demonstrate the need to account for spatial and temporal variability when undertaking species distribution and abundance modelling more broadly.

Model results indicate that dabbling duck distributions exhibit high levels of spatial and temporal correlation. When evaluated for occurrence probability over the three-year period 2015 through 2017, each duck species presented in this study displayed spatial distributions with near-synchronous, season-specific overlap between years \( r = 0.99 \). Comparably, the degree of correlation between seasonal duck abundances exceeded \( r = 0.55 \) for the same time period. In addition to highlighting the need for models to account for temporal correlation when forecasting migratory species across

**FIGURE 3** NSHO predicted abundance. Figures show predicted abundance for the (a) spring, (b) breeding, (c) fall and (d) winter seasons. The figure legend is scaled from 0 to 100\(^{+}\) birds/\( \text{km}^2 \) with lighter yellow colours signifying lower bird densities and darker reds emphasizing locations with relatively higher abundance.

**FIGURE 4** Comparing NSHO occurrence over the Prairie Pothole Region (PPR). Figures depict predicted occurrence probability for the (a) breeding and (b) winter seasons. Warm colours (reds) highlight regions with greater than a 50% chance of occurrence, and cool colours (blues) indicate occurrence probabilities below 50%. Smooth black line approximates PPR boundaries in the United States. Note that this figure is scaled differently than Figure 2 and that predicted occurrence probabilities outside of the PPR are set to zero (tan colour) for improved visual clarity.
environmental and geographic space, the strength of these correlations may have important implications for waterfowl conservation and wildlife disease risk assessment. For example, Figure 1 serves to emphasize that although duck presence is highly correlated to duck abundance, this relationship is non-linear and ducks may be substantially more abundant at some locations and times than at others. That is, a correlation of $r = 0.55$ still allows for sizable inter-annual variability and the locations of densest duck aggregation may change year to year. As an applied matter, it is an important consideration that not all suitable habitat is created equal; some areas of suitable habitat may support substantially more ducks than do others. In an epidemiological context, if wildlife disease transmission rates correspond to population density, some locations may be at considerably greater risk than others and these higher-risk locations may move in the landscape from year to year.

Finally, we believe that our modelling framework helps address some of the outstanding concerns with modelling highly mobile species and underscores that spatially and temporally explicit methods grant distinct advantages over non-explicit alternatives. Our results for ten dabbling duck species appear reasonably robust considering use of publicly sourced data and the areal extent of our study area. Species distribution modelling is not a substitute for field survey; however, when formal survey data are not available and cannot be practically acquired, the presented framework offers a realistic alternative for modelling highly mobile species that vary throughout time and across space.

**ACKNOWLEDGEMENTS**

We thank J. Andrew Royle for assistance in reviewing the manuscript, and Yassine Kodad and Kevin Laurent for early stages of analysis. This research was supported by the U.S. Geological Survey Environmental Health Mission Area’s Contaminant Biology Program and the U.S. Geological Survey Ecosystems Mission Area. In addition, USDA’s Animal and Plant Health Inspection Service supported this work under Cooperative Agreement 6000001762. Computing for this project was performed on the High Performance Computing Cluster at the Florida State University Research Computing Center and at the Michigan State University Institute for Cyber-Enabled Research. The use of trade, firm or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

**DATA ACCESSIBILITY**

All analysis and modelling were performed using the open-source R language for statistical computing (R Core Team, 2017) with freely available data. Occurrence data for all modelled species are available at no cost through the eBird website (https://ebird.org/home).

**ORCID**

John M. Humphreys https://orcid.org/0000-0002-8464-0886

**REFERENCES**

Allouche, O., Tsoar, A., & Kadmon, R. (2006). Assessing the accuracy of species distribution models: Prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology*, 43(6), 1223–1232. https://doi.org/10.1111/j.1365-2664.2006.01214.x

Arab, A. (2015). Spatial and spatio-temporal models for modeling epidemiological data with excess zeros. *International Journal of Environmental Research and Public Health*, 12(9), 10536–10548. https://doi.org/10.3390/ijerph120910536

Baddeley, A., & Turner, R. (2005). spatstat: An R Package for Analyzing Spatial Point Patterns. *Journal of Statistical Software*, 12(6), 1–42.

Barker, N. K. S., Cumming, S. G., & Darveau, M. (2014). Models to predict the distribution and abundance of breeding ducks in Canada. *Des Modèles Pour Prédire La Répartition Et L’Abondance Des Canards Nichant Au Canada*, 9(2), 41–82. https://doi.org/10.5751/ACE-00699-090207

Bauer, S., & Hoye, B. J. (2014). Migratory animals couple biodiversity and ecosystem functioning worldwide. *Science*, 344(6179). https://doi.org/10.1126/science.1242552

Bean, W. T., Stafford, R., & Brashares, J. S. (2012). The effects of small sample size and sample bias on threshold selection and accuracy assessment of species distribution models. *Ecography*, 35(3), 250–258. https://doi.org/10.1111/j.1600-0587.2011.06545.x

Belkhiria, J., Hijmans, R. J., Boyce, W., Crossley, B. M., & Martinez-Lopez, B. (2018). Identification of high risk areas for avian influenza outbreaks in California using disease distribution models. *PloS ONE*, 13(1), e0190824. https://doi.org/10.1371/journal.pone.0190824

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Hoboken, New Jersey: John Wiley & Sons, Inc. https://doi.org/10.1002/0471725153.

Besbeas, P., & Morgan, B. J. T. (2014). Goodness-of-fit of integrated population models using calibrated simulation. *Methods in Ecology and Evolution*, 5(12), 1373–1382. https://doi.org/10.1111/2041-210X.12279

Cardador, L., Carrete, M., Gallardo, B., & Tella, J. L. (2016). Combining trade data and niche modelling improves predictions of the origin and distribution of non-native European populations of a globally invasive species. *Journal of Biogeography*, 43(5), 967–978. https://doi.org/10.1111/jb.12694

CIESIN · Center for International Earth Science Information Network, Columbia University (2016). *Gridded Population of the World, Version 4 (GPw4): Population Density*. Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC). Available at http://sedac.ciesin.columbia.edu/data/set/gpw-v4-population-density. Accessed 03/2018

Cramer, J. S. (2003). Logit models from economics and other fields. Cambridge, UK: Cambridge University Press.

Cressie, N., & Wikle, C. (2011). *Statistics for spatio-temporal data*. Hoboken, NJ: Wiley and Sons.

CropScape (2017). USDA National Agricultural Statistics Service Cropland Data Layer. 2017. Published crop-specific data layer. Washington, DC: USDA-NASS. Retrieved from https://nassgeodata.gmu.edu/CropScape/

Dormann, C. F. (2007). Effects of incorporating spatial autocorrelation into the analysis of species distribution data. *Global Ecology and Biogeography*, 16(2), 129–138. https://doi.org/10.1111/j.1466-8238.2006.00279.x

Elith, J., & Leathwick, J. (2009). Species distribution models: Ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics*, 40(1), 677–697. https://doi.org/10.1146/annurev.ecolsys.110308.120159

Elsner, J. B., Fricker, T., Widen, H. M., Castillo, C. M., Humphreys, J., Jung, J., ... Dixon, P. G. (2016). The relationship between elevation roughness and tornado activity: A spatial statistical model fit to data from the central great plains. *Journal of Applied Meteorology and Climatology*, 55(4), 849–859. https://doi.org/10.1175/JAMC-D-15-0225.1
and Wildlife-Associated Recreation. Tech. rep., U.S. Fish and Wildlife Service, Washington, D.C., USA.

Van Den Brand, J. M., Verhagen, J. H., Veldhuis Kroeze, E. J., Van De Bildt, M. W., Bodewes, R., Herfst, S., ... Kuiken, T. (2018). Wild ducks excrete highly pathogenic avian influenza virus H5N8 (2014–2015) without clinical or pathological evidence of disease article. Emerging Microbes and Infections, 7(1), 1–10. https://doi.org/10.1038/s41426-018-0070-9

Velázquez, E., Martínez, I., Getzin, S., Moloney, K. A., & Wiegand, T. (2016). An evaluation of the state of spatial point pattern analysis in ecology. Ecography, 39,1042–1055. https://doi.org/10.1111/ecog.01579

Veldman, R., & McGeoch, M. A. (2004). Spatially explicit analyses unveil density dependence. Proceedings of the Royal Society B: Biological Sciences, 271(1556), 2439–2444.

Wiens, J. A. (2000). Ecological heterogeneity: Ontogeny of concepts and approaches. In M. Hutchings, E. Stewart, & J. Jones (Eds.), The ecological consequences of environmental heterogeneity (pp. 9–31). Oxford, UK: Blackwell Science.

Yin, D., & He, F. (2014). A simple method for estimating species abundance from occurrence maps. Methods in Ecology and Evolution, 5(4), 336–343. https://doi.org/10.1111/2041-210X.12159

Zeng, Q., Zhang, Y., Sun, G., Duo, H., Wen, L., & Lei, G. (2015). Using species distribution model to estimate the wintering population size of the endangered scaly-sided merganser in China. PLoS ONE, 10(2), 1–12. https://doi.org/10.1371/journal.pone.0117307

Zipkin, E. F., Rossman, S., Yackulic, C. B., Wiens, J. D., Thorson, J. T., Davis, R. J., & Grant, E. H. (2017). Integrating count and detection nondetection data to model population dynamics. Ecology, 98(6), 1640–1650. https://doi.org/10.1002/ecy.1831

BIOSKETCH

John M. Humphreys is a Research Associate in Spatial Ecology at The Michigan State University. His research interests include spatial–temporal statistics, disease ecology, animal movement analysis and applied wildlife ecology.

Author contributions: D.P. and J.M. conceived the study concept; J.H. led and J.S./D.P. contributed to design of the methodology and data analysis. All authors contributed to writing the manuscript and gave final approval for publication.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.