




Integrated modeling of waterfowl distribution in western Canada using aerial survey and citizen science (eBird) data

ANTOINE ADDE ^{1,2,†} CLARA CASABONA I AMAT,¹ MARC J. MAZEROLLE ¹ MARCEL DARVEAU,¹
STEVEN G. CUMMING ^{1,2} AND ROBERT B. O'HARA³

¹*Département des sciences du bois et de la forêt, Université Laval, Québec, Québec, Canada*

²*Boreal Avian Modelling Project, University of Alberta, Edmonton, Alberta, Canada*

³*Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway*

Citation: Adde, A., C. Casabona i Amat, M. J. Mazerolle, M. Darveau, S. G. Cumming, and R. B. O'Hara. 2021. Integrated modeling of waterfowl distribution in western Canada using aerial survey and citizen science (eBird) data. *Ecosphere* 12(10):e03790. 10.1002/ecs2.3790

Abstract. Although the exceptional spatiotemporal extent of the Waterfowl Breeding Population and Habitat Survey (WBPHS) has substantially contributed to our understanding of the ecology of North American waterfowl, vast geographical areas remain excluded from the survey. The unprecedented number of observations generated by the recent boom in citizen science initiatives could help resolve these spatial gaps and increase the density of records in regions already covered. The study objective was to assess the value of the integrated species distribution modeling (ISDM) approach for integrating WBPHS and eBird data to model waterfowl distribution across the Canadian western boreal forest, where WBPHS data are sparse. Following the ISDM approach, we used a state-space point process formulation that combined a model for the “true” species distribution and two observation models for how WBPHS and eBird data were generated. Our results highlighted the importance of observational processes related to sampling effort and site accessibility for modeling eBird data. In addition, our models allowed identifying waterfowl–habitat associations related to geoclimatic, forest, and hydrological factors that explained the distribution of target species. To assess the individual contribution of WBPHS and eBird data, we re-fitted the models using only one of the two data sets and compared the results obtained against those from the integrated approach. Waterfowl–habitat associations and predictions derived from the models using both data sets and those fitted with WBPHS data only were close and consistent with the observed species distribution. However, it was more difficult to extract an ecological signal from models fitted with eBird data only. Interestingly, predictions from models combining both data sets were closer to the WBPHS records than the predictions from models fitted with WBPHS data only. By facilitating the combination of all available data sources, we demonstrated the potential of the ISDM approach for modeling and mapping species distributions. We encourage future North American waterfowl modeling attempts to use this method, especially for resolving gaps in the WBPHS coverage. As multiple data sets can be added to the original framework, integration efforts must not be restricted to the additional contribution of eBird data alone and could consider, for example, provincial atlases and regional helicopter surveys.

Key words: boreal forest; citizen science; data integration; eBird; integrated species distribution modeling; point process; state-space model; Waterfowl Breeding Population and Habitat Survey.

Received 16 December 2020; revised 14 May 2021; accepted 27 May 2021. Corresponding Editor: Juan-Carlos Rocha.

Copyright: © 2021 The Authors. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

† **E-mail:** antoine.adde.1@ulaval.ca

INTRODUCTION

Thanks to the economic importance of waterfowl hunting (Johnsgard 2010, Carver 2015, Mattsson et al. 2018) and their status under the Migratory Bird Convention Acts of Canada and the U.S. (ECCC 2020, USFWS 2020), North American waterfowl benefit from one of the most extensive wildlife surveys worldwide: the Waterfowl Breeding Population and Habitat Survey (WBPHS) (USFWS 2019). This annual aerial survey was established in the early 1950s to inform hunting regulations. It now covers ~3.6 million km² of breeding habitat (Smith 1995, Nichols and Williams 2006). The exceptional spatiotemporal coverage of the WBPHS has contributed to maintaining high population levels and improved our understanding of the ecology of North American waterfowl (Sorenson et al. 1998, Barker et al. 2014, Doherty et al. 2015). However, vast areas remain uncovered by the survey (Fig. 1). Consequently, extrapolating results inferred from WBPHS data to these areas is problematic. For example, the widest prediction intervals obtained from WBPHS-borne waterfowl abundance models for Canada are found in regions poorly covered by the survey, including British Columbia and Yukon (Barker et al. 2014, Adde et al. 2020a). This uncertainty is a barrier to the conservation of waterfowl and their habitat in these regions (Ducks Unlimited 2020).

Over the last 15 yr, the boom in citizen science initiatives has generated a vast quantity of ecological records (Dickinson et al. 2010, Amano et al. 2016, Pocock et al. 2017). eBird alone, the world's largest biodiversity-related citizen science initiative (Sullivan et al. 2014), gathers more than 100 million bird sightings each year (<https://ebird.org/home>). Seizing this opportunity to increase the spatiotemporal extent and density of observational data, the scientific community has started to apply citizen science to conservation challenges (Devictor et al. 2010, Chandler et al. 2017, McKinley et al. 2017). In particular, species occurrence records from citizen science have proved useful for inferring distributional patterns (Humphreys et al. 2019), population trends (Fink et al. 2020), and migratory behavior (Fournier et al. 2017) of

avian species. This suggests that citizen science records could help to resolve spatial gaps in the WBPHS coverage. However, because citizen science data are generally obtained under unstructured protocols with uneven spatiotemporal sampling effort, it has been challenging to use them in combination with standardized survey data (Dickinson et al. 2010, Isaac et al. 2014, Geldmann et al. 2016). Filtering procedures and hierarchical modeling methods that, respectively, facilitate data homogenization and fusion have proved useful in meeting these challenges (Pacifi et al. 2017, Fletcher et al. 2019, Miller et al. 2019).

Isaac et al. (2020) formalized the concept of integrated species distribution modeling (ISDM), the practice of fitting species distribution models with more than one observation model. Building upon attempts to combine species occurrence records obtained under heterogeneous observation processes (Dorazio 2014, Pagel et al. 2014, Fithian et al. 2015), ISDM aims to take advantage of all available data while accounting for heterogeneities in the sampling protocol, the spatial structure, and the nature of the data (e.g., presence only, detection/non-detection, or abundance). To facilitate the implementation of the ISDM approach, Isaac et al. (2020) proposed a flexible modeling framework based on a state-space formulation where a single model infers "true" species distributions corrected for sampling biases, while different observation models account for the observation processes of each data set. Species distributions are modeled as a log-Gaussian Cox process (LGCP) (Møller et al. 1998), which aims to estimate an intensity surface of the density of points (i.e., of individuals) in an area (Renner et al. 2015, Soriano-Redondo et al. 2019, Sicacha-Parada et al. 2020). Point process models of this kind facilitate data set integration because they alleviate the need to discretize records into a priori spatial units, thereby preserving the spatial accuracy of the underlying observation while remaining invariant to spatial scale (Illian and Burslem 2017, Miller et al. 2019, Isaac et al. 2020). Computationally expensive LGCPs belong to the class of latent Gaussian models and so can be estimated in a Bayesian context using integrated nested Laplace approximations (INLA) (Rue et al. 2009, Illian et al. 2013, Illian and Burslem 2017) and stochastic partial

differential equations (SPDE) (Lindgren et al. 2011), which speed up model inference.

The objective of this method paper was to assess the value of the ISDM approach for integrating WBPHS and eBird data to model waterfowl distribution across the Canadian western boreal forest, where WBPHS data are very sparse. To do this, we adapted the conceptual method proposed by Isaac et al. (2020) to integrate the two data sets, using a common underlying distribution model, and two data set-specific observation models for how WBPHS and eBird data were generated. We analyzed the fitted ISDM model by examining the parameter estimates and mapping model predictions obtained for three test species. The added value of the integrated approach and the individual contributions of the two data sets were assessed relative to two data set-specific models.

METHODS

Study area

Our study area was the Canadian western boreal forest (CWBF) (Fig. 1), a central breeding ground for many waterfowl species. This area ranks third out of the 25 most important and threatened waterfowl habitats in North America (Ducks Unlimited 2020). We delineated the CWBF using the extent of the boreal bird conservation regions (BCRs) (NABCI 2014) 4, 6, 7, and 8 in Manitoba, Saskatchewan, Alberta, British Columbia, Nunavut, Northwestern Territories, and Yukon, an area of ~3 million km².

Waterfowl data

We retrieved waterfowl data from the WBPHS and eBird databases for three ecologically contrasted example species commonly found in the CWBF: *Mareca americana* (American wigeon; AMWI), *Bucephala albeola* (bufflehead; BUFF), and *Branta canadensis* (Canada goose; CAGO). Details on the ecology of these three species can be found in Mack and Morrison (2006). In short, AMWI is a ground-nesting species whose primary breeding area is within the CWBF. It was of regional conservation concern until the early 2010s. BUFF is an emblematic CWBF cavity nester and the only duck from this nesting guild recorded at the species level in the WBPHS

database. CAGO is a generalist ground-nesting species, with a wide breeding range in North America that extends from the remote arctic tundra to urban areas.

WBPHS data.—Each May, WBPHS observers count adults of waterfowl species seen within 200 m of transects flown by fixed-wing aircraft. Transects are subdivided into segments of ~30 km (Smith 1995) to which observations are registered. We retrieved segment-level ($n = 814$) annual counts of waterfowl species conducted in the CWBF portion of the WBPHS for a 27-yr period (1990–2017). The spatial coverage of WBPHS segments within the CWBF is highly heterogeneous (Fig. 1). The highest WBPHS segment density was found in the southern part of BCR 6 (Boreal Taiga Plains) and BCR 8 (Boreal Softwood Shield). In contrast, BCRs 4 (Northwestern Interior Forest) and 7 (Taiga Shield and Hudson Plains) are almost excluded from the WBPHS. Survey coverage within the boreal has expanded considerably over time (Barker 2015). We chose the 1990–2017 period because it (1) provided a reasonably long time series for many segments to limit the effects of interannual variation in waterfowl populations; and (2) it included the recent period of maximum spatial coverage of the CWBF to allow the best possible characterization of current conditions therein. This period was also consistent with the availability of eBird data (see *Methods: Waterfowl data: eBird data*). WBPHS data for 2018 and 2019 were not available to us as of February 2020, when the present analysis was conducted. Data from 2007 were excluded because of a deviation in that year from the usual survey design (Barker et al. 2014).

eBird data.—eBird is an assemblage of species observations reported by members of the public with no central coordination of sample characteristics (Sullivan et al. 2014). Volunteer observers submit eBird records to <https://ebird.org> in the form of checklists listing counts of encountered species. Checklist metadata include information on the general context of the observations (e.g., geographical coordinates, date, and time) and the sampling effort (e.g., checklist duration, distance traveled, and the number of observers). On 20 February 2020, we retrieved eBird records collected across our

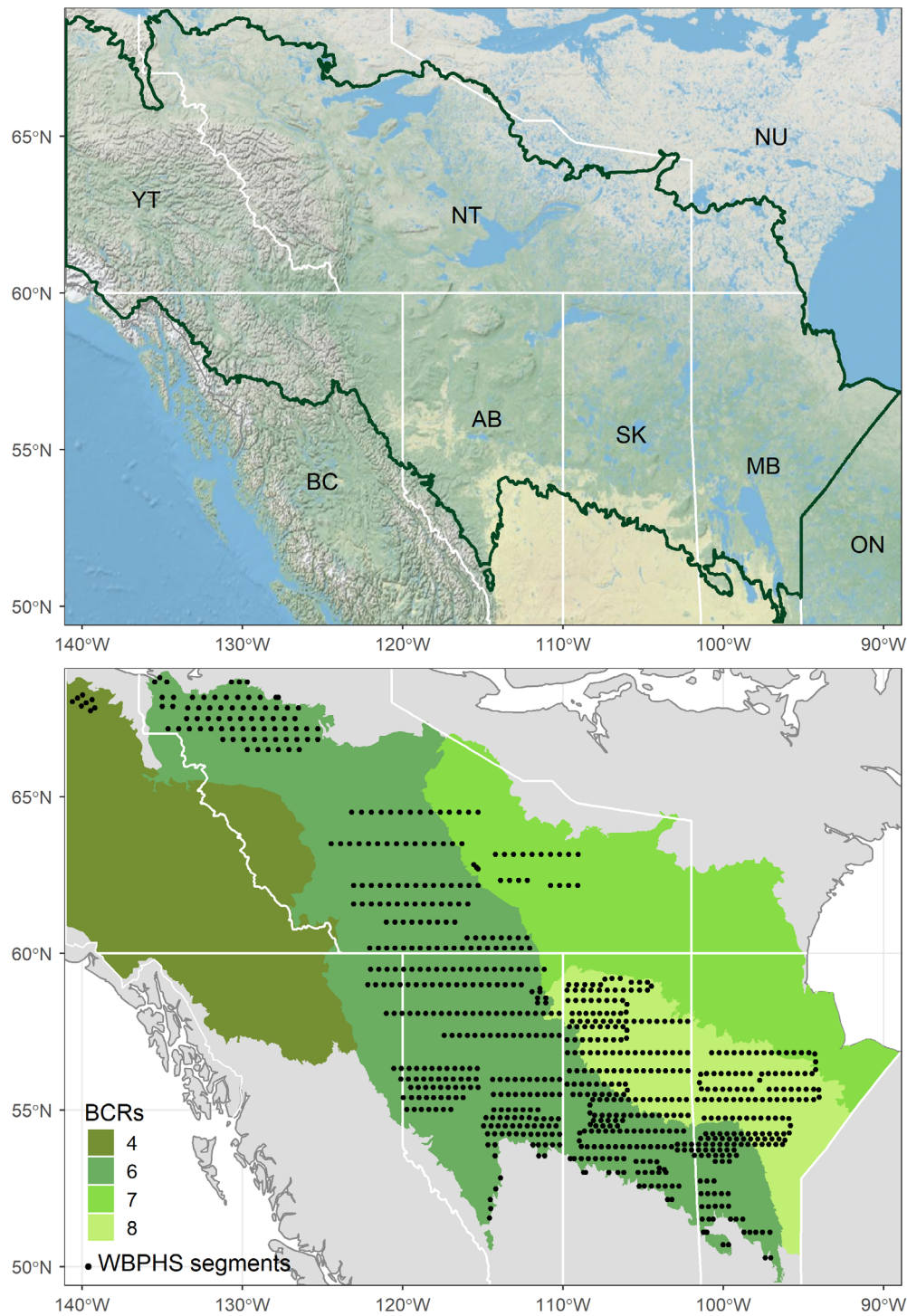


Fig. 1. Study area in western Canada with its coverage by Waterfowl Breeding Population and Habitat Survey (WBPHS) segments. *Top*. Delineation of the Canadian western boreal forest (CWBF; green line), our study area. White lines: provinces (AB: Alberta; BC: British Columbia; MB: Manitoba; NT: Northwest Territories; NU: Nunavut; ON: Ontario; SK: Saskatchewan; YT: Yukon Territory). *Bottom*. Location of the 814 WBPHS segments

(Fig. 1. *Continued*)

used in our study. Segment, shown as black dots are ~30 km units arrayed along linear transects BCRs: Bird Conservation Regions used to delineate the CWBF (outlines from NABCI 2014; 4: Northwestern Interior Forest; 6: Boreal Taiga Plains; 7: Taiga Shield and Hudson Plains; 8: Boreal Softwood Shield).

study area during the breeding season (April–July) for the period 1990–2019. Data were filtered according to the best practices for using eBird data (Strimas-Mackey et al. 2020). Specifically, we included (1) only completed checklists (i.e., all the birds observed are reported); (2) surveys conducted between 5:00 a.m. and 9:00 p.m.; (3) covering <5 km; (4) collected over a period of no longer than 5 h; and (5) with no more than 10 observers.

Covariates

Observational covariates.—We considered five candidate observational covariates to explicitly account for sampling biases in our models (Table 1). Observational covariates for eBird data were related to checklist effort (2/5) and site accessibility (3/5). The two checklist-effort covariates, “checklist duration” and “distance traveled,” were directly retrieved from the eBird checklists. The three site-accessibility covariates, “distance to road,” “road density,” and “travel time to city,” were extracted at each checklist location and

standardized (centered and scaled to a mean value of zero and unit variance) (Appendix S1: Fig. S1). There were no strong pairwise correlations among the five covariates (Pearson product–moment correlation coefficient: $|r| < 0.70$).

Ecological covariates.—We included a literature-based (Adde et al. 2020a, b, Adde et al. 2021) suite of 14 candidate ecological covariates aimed at explaining the large-scale distribution of waterfowl across our forest-dominated study area (Table 1 and Appendix S1: Fig. S2). The covariates were classified as “Geoclimatic” (4/14), “Hydrologic” (2/14), and “Forest composition” (8/14). For consistency with the scale at which WBPBS records are provided (~30 km × 400 m segments) and to avoid issues related to multiresolution, we resampled all covariates to the 300-arcsecond grid on which the geoclimatic covariates were provided (cell areas ranged from 31 to 56 km²) to resample all 14 ecological covariates. All ecological covariates were standardized to zero mean and unit variance. There were no strong pairwise correlations among covariates ($|r| < 0.70$).

Table 1. Definitions, spatial resolution, and temporal coverage of the source data sets used to compute the candidate covariates.

Category	Theme	Data set	Period or year	Spatial resolution	Covariates
Observational	Checklist effort	eBird checklists (Sullivan et al. 2014)	1990–2020	Spatial point	Distance traveled; checklist duration
Observational	Site accessibility	National Road Network (Statistics Canada 2012)	2010s	1:50,000	Distance to road; road density
Observational	Site accessibility	Malaria Atlas Project (Weiss et al. 2018)	2015	1 km	Travel time to cities (> 50,000 inhabitants) via surface transport
Ecological	Geoclimatic	WorldClim (Fick and Hijmans 2017)	1970–2000	300 arcsec	Mean temperature; mean precipitation; mean climate moisture index
Ecological	Geoclimatic	ENVIREM (Title and Bemmels 2018)	2000s	300 arcsec	Mean topographic wetness index
Ecological	Hydrologic	Land Cover of Canada (Latifovic et al. 2017)	2010	30 m	% area of wetland; % area of open water
Ecological	Forest	kNN-Canada’s forest attributes (Beaudoin et al. 2017)	2011	250 m	% of aboveground tree biomass (<i>Betula papyrifera</i> ; <i>Larix laricina</i> ; <i>Picea glauca</i> ; <i>Picea mariana</i> ; <i>Pinus banksiana</i> ; <i>Pinus contorta</i> ; <i>Populus balsamifera</i> ; <i>Populus tremuloides</i>)

Integrated modeling framework

We analyzed the data for each of the three waterfowl species separately. To predict the spatial distribution of a species across the CWBF, we followed the ISDM approach (Isaac et al. 2020) by formulating a state-space point process model, which was then fitted in a Bayesian framework using INLA (Rue et al. 2009). INLA is a fast and flexible alternative to traditional Markov chain Monte Carlo methods for approximate Bayesian inference in latent Gaussian models (Rue et al. 2009), a wide class of regression models that includes LGCPs (Møller et al. 1998). By facilitating model formulation and fitting in complex and high-dimensional settings, INLA has contributed to the recent popularity of point process models in ecological sciences (Illian et al. 2013, Soriano-Redondo et al. 2019, Opitz et al. 2020). Data preparation and model fitting were conducted using the R-package “PointedSDMs” (<https://github.com/oharar/PointedSDMs>; not yet on CRAN), which is built on the widely used “R-INLA” package (Lindgren and Rue 2015). The complete R-code to implement our models is provided in Data S1.

Fig. 2 illustrates the hierarchical modeling structure used in this study. Our state-space formulation can be thought of as the combination of a process model for the “true” species distribution with two data set-specific observation models for how WBPBS and eBird data were generated. The “true” species distribution is

treated as an unobserved state $\lambda(s)$ function of ecological covariates X and parameters ϕ such as $p(\lambda(s), X, \phi)$. The observation models link the recorded species distribution to the underlying state such that the likelihood for a data set Y_i is $\Pr(Y_i|\lambda(s), \theta_i)$, with θ_i denoting the parameters of the observation models. Combining these elements, the full likelihood for the model becomes:

$$L(Y_i|X, \phi, \theta_i) \propto p(\lambda(s), X, \phi) \prod_{i=1}^M \Pr(Y_i|\lambda(s), \theta_i) \quad (1)$$

Process model.—The true species distribution was modeled as a log-Gaussian Cox process (Møller et al. 1998) with intensity $\lambda(s) = e^{\eta(s)}$ defining the expected number of points (i.e., individuals) at location s . The intensity was a function of ecological covariates $X(s)$ and $u(s)$ a spatially continuous Gaussian random field that aimed to account for unmeasured covariates and potential spatial autocorrelation:

$$\eta(s) = \sum_{i=1}^P \beta_i X_i(s) + u(s) \quad (2)$$

For computational efficiency, we used the SPDE approach to model the spatial field in the form of a Gaussian Markov random field with zero mean and Matérn covariance function (Lindgren et al. 2011). Briefly, the INLA-SPDE approach evaluates the continuous random field as a discretely indexed random process based on a spatial mesh defined by triangulation of the

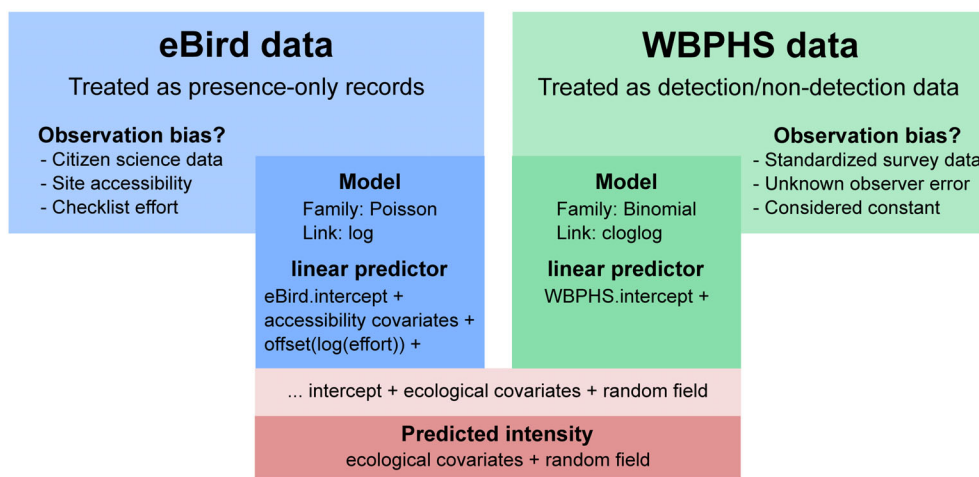


Fig. 2. Schema of the hierarchical modeling structure used in our study.

study area (Appendix S1: Fig. S3). The mesh design was exploratory, and here, we sought a trade-off between the computational costs and precision of the spatial field. For additional details on the INLA-SPDE approach, the reader is referred to the foundational paper and other practical implementations (e.g., Cameletti et al. 2013, Blangiardo and Cameletti 2015, Krainski et al. 2019).

Given this structure (Eq. 2), the expected number of points within an area B follows a Poisson distribution with mean:

$$\mu(B) = \int_B \lambda(s) ds \quad (3)$$

This integral is analytically intractable but can be approximated numerically using integration points obtained after discretizing the study area into triangles (Simpson et al. 2016) such as:

$$\mu(B) \approx \sum_{s=1}^{nB} A(s)e^{\eta(B(s))} \quad (4)$$

where nB is the number of integration points in B , each located at $B(s)$, and $A(s)$ is the area of the triangle around s . Under this framework, the value of the intensity is only calculated at the integration points. For any other location, the intensity is interpolated between the three points that form the corners of the surrounding triangle. The mesh designed for computing the Gaussian Markov random field was used to define the integration points (Appendix S1: Fig. S3).

WBPHS observation model.—We treated the WBPHS counts as simplified segment-level detection/non-detection data (i.e., detection when waterfowl count $N > 0$; non-detection otherwise) replicated over the number of years surveyed. This simplification was applied because we had no information on the distribution of individual waterfowl across the ~30 km segments, which is necessary for point process intensity estimation. WBPHS segment centers were used as model points in the analyses. To integrate the binary detection/non-detection information with the process model, we modeled the probability of having a waterfowl count > 0 using a binomial model with a complementary log-log link function (cloglog; $f(x) = \log(-\log(1 - x))$). As demonstrated in previous applications (Kéry and Royle 2016, Bowler et al. 2019),

this is equivalent to expressing the detection/non-detection probability as a function of the intensity $\lambda(s)$ of an underlying Poisson process, such as:

$$\begin{aligned} \Pr(N(s) > 0) &= 1 - \Pr(N(s) = 0) = 1 - e^{-\lambda(s)} \\ &= 1 - e^{-e^{\eta(s)}} \end{aligned} \quad (5)$$

which corresponds to the inverse of the cloglog link function

$$\log(-\log(1 - \Pr(N(s) > 0))) = \eta(s) \quad (6)$$

where $\eta(s)$ includes the shared process model components from Eq. 2 (the set of ecological covariates and the random field) and an intercept. This extra intercept simply assumed a constant observation bias for the WBPHS data.

eBird observation model.—Because they are prone to high sampling biases and absences cannot be confidently assessed, we treated eBird records as presence-only data (i.e., presence when the species was on the checklist) emerging from a thinned Poisson process of the underlying LGCP (Eq. 2). The observed pattern of points (eBird records) being a thinned-out version of the complete distribution of individuals (Isaac et al. 2020). The intensity of the thinned LGCP is $\lambda(s)b(s)$, with $b(s)$ the thinning probability, which we assumed to be dependent on the observational covariates $Z_j(s)$ for sampling effort and site accessibility. As demonstrated in Renner et al. (2015), $\lambda(s)$ and $b(s)$ can be modeled with a log link function using a standard generalized linear model formulation:

$$\log(\lambda(s)b(s)) = \sum_{i=1}^P \beta_i X_i(s) + u(s) + \sum_{j=1}^Q \delta_j Z_j(s) \quad (7)$$

Accessibility covariates such as “distance to road” and “travel time to cities” were included as components of the vector of regressors $Z(s)$. In contrast, measures of sampling effort (i.e., “checklist duration” and “distance traveled”) were included as log-transformed offsets to adjust for differential exposures. In preliminary analyses, effort covariates were also tested as standard regressors, but we obtained better model performances (see Covariate selection procedure) when using these variables as offsets. Similarly to the WBPHS observation model, an

intercept was also estimated for the eBird observation model.

Prior specification

The prior distributions for fixed effect coefficients were set to the R-INLA “non-informative” defaults, allowing the data to predominate in calculating the posterior distributions. Accordingly, Gaussian priors with 0 mean and precision of 0.001 were used for the fixed effects. Preliminary tuning analyses showed that the choice of priors had no effect on fixed parameter estimates.

Despite multiple tests of penalized complexity priors (Simpson et al. 2017), preliminary analyses revealed systematic overfitting of the random field (i.e., perfectly reproducing spatial patterns found in observed data). To avoid this problem, fixed prior values were used for the parameters “prior.range,” the distance at which spatial correlation declines to ≈ 0.1 (Krainski et al. 2019), and “prior.sigma,” the standard deviation of the field. On an exploratory basis, we searched for a combination of “prior.range” and “prior.sigma” values that would result in a random field that explained up to 25% of the variations in model predictions. This objective was achieved by fixing “prior.range” to 0.25° and “prior.sigma” to 0.1.

Covariate selection procedure

The covariate selection procedure was conducted first for observational and then for ecological covariates. The small number of candidate observational covariates ($n = 5$) allowed us to analyze the models fitted using all 32 possible combinations of these covariates (Data S2). Each candidate model included the intercepts and the random field, but no ecological covariates. The 32 models were ranked by the Watanabe-Akaike information criteria (WAIC; Watanabe 2010), a Bayesian approach for estimating the out-of-sample expectation (Gelman et al. 2014). The models were ranked by WAIC in increasing order. We retained the top-ranking model.

Candidate ecological covariates ($n = 14$) were added to the best observation model. As an exhaustive evaluation of the 16,384 possible variable subsets was infeasible, we applied a backward elimination procedure. We started by fitting a full model that included all 14 ecological covariates. We examined the 95% credible

intervals (95% CrI) of the estimated coefficients and identified those that included zero. A final model was then fitted, excluding all these “non-significant” ecological covariates.

Model checking and spatial predictions

As a preliminary visualization step to help interpret model results, we mapped WBPHS and eBird data used in the model for each of three waterfowl species. We mapped WBPHS data as a segment-level ratio of the number of years when the species was detected at least once (NPresence) and the number of years the segment was surveyed (NTrials). For eBird data, we mapped the location of each recorded presence point over a plot of the smoothed density obtained using a kernel density estimator (Silverman 1986).

We interpreted the model inferential properties using the mean, standard deviation (SD), and 95% CrI of the posterior distributions of the model parameters. A regular grid of 0.20-degree resolution was used to map model predictions for each species, obtained as a function of both the ecological covariates and the estimated random field. We mapped the predicted intensity and estimated SD for each grid cell ($n = 12,933$). To evaluate the common variance between the mapped intensities of two species, we calculated Pearson’s R^2 for each species pair (AMWI vs. BUFF, AMWI vs. CAGO, and BUFF vs. CAGO). To disentangle the effects of ecological covariates from the random fields, we re-mapped the predicted intensity obtained as a function of either (1) the ecological covariates only or (2) the random field only. We used Pearson’s R^2 to evaluate the proportion of variance in full-model predictions explained by covariate-only and random field-only predictions.

Integrated vs. single data set models

We assessed the behavior of models fitted with WBPHS or eBird data only and investigated the individual contributions of the two data sets to the integrated models. To do so, we iteratively re-fitted our models using only one of the two data sets and compared the results obtained with those of the integrated approach. Using Pearson’s R^2 , we estimated the proportion of variance in integrated model predictions explained by WBPHS- and eBird-only predictions. In addition, we compared the fit of the models to the

recorded values by computing the R^2 between (1) the recorded segment-level NPpresence/Ntrials ratios (WBPHS data) and (2) the intensity predicted at the center of each segment. This exercise was done individually for each of the three species and prediction sets (integrated data sets, WBPHS only, and eBird only).

Finally, we summarized the BCR-level distribution of the total predicted intensity and standard deviation for each of the three prediction sets by calculating the percentages of the total intensity and standard deviation that belonged to a given BCR weighted by the number of grid cells in that BCR. These statistics were calculated for the three species individually. Because our study area included four BCRs, the weighted percentages for each BCR should have been approximately 25% if the predicted intensity and standard deviation were homogeneously distributed.

RESULTS

Recorded waterfowl data

WBPHS data.—Waterfowl data extracted from the WBPHS are displayed in Fig. 3 (left panel). At the study area level, the mean NPpresence/Ntrials ratios (NP/NT) were 0.49, 0.32, and 0.32 for BUFF, AMWI, and CAGO, respectively. The three species were commonly encountered in the southern part of our study area (<55° N) with NP/NT of 0.56, 0.49, and 0.35 for BUFF, CAGO, and AMWI, respectively. At higher latitudes (>65° N), the probability of encountering BUFF and CAGO decreased (NP/NT < 0.20), but not for AMWI (NP/NT = 0.51). The lowest NP/NT values for AMWI were found in the eastern portion of our study area (<110° W), particularly across BCR 7 (NP/NT = 0.16; see Fig. 1 for BCRs map).

eBird data.—After applying the filtering procedure (see *Methods: Waterfowl data: eBird data*), the number of eBird records for AMWI, BUFF, and CAGO was 5901, 6334, and 11,696, respectively. Compared to WBPHS data, it was more difficult to identify species-specific distributional patterns from eBird data (Fig. 3; right panel). eBird records were highly clustered around a few hotspots and were characterized by several linear patterns corresponding to the road network. As revealed by the kernel density surface of each

species, the highest density of points was found in the westernmost BCR 4, in an area located around 61° N and 134° W (Whitehorse and its surroundings; Yukon's capital city) (Fig. 3). The small bounding box between 60–62° N and 132–138° W gathered 40, 27, and 12% of the total AMWI, BUFF, and CAGO records, respectively. Several smaller hotspots were found in the southern part of BCR 6 (<57° N), where CAGO had the highest percentage of total records (68%), and AMWI had the lowest (38%). We also noticed two CAGO hotspots in the eastern portion of our study area (51° N/97° W and 52° N/92° W), which were unidentifiable for the other two species.

From a species distribution modeling perspective, Fig. 3 shows that the presence of eBird data in several areas not covered by the WBPHS could provide important additional information to capture the ecological signal underlying the “true” waterfowl distribution more effectively. This result applies in particular to BCRs 4 and 8, which are almost excluded from the WBPHS coverage. However, the complex distribution of eBird records indicates that their integration would be challenging.

Integrated waterfowl distribution models

Covariate selection and estimates.—Our integrated models were successfully fit for the three species (Fig. 4; see Appendix S1: Tables S1–S3 for full-model summaries). The details of the two steps of the covariate selection procedure (1: observational covariates, 2: ecological covariates) are shown in Data S2.

For all species, three accessibility covariates (“distance to road,” “road density,” and “travel time to city”) and one checklist-effort covariate (“checklist duration”) were retained after Step 1. The intensity of the point process decreased with an increase in travel time to cities. “Road density” had a negative effect on the intensity of the point process for AMWI, but was not significant for BUFF and CAGO. The intensity increased with an increase in the distance to the nearest road. The checklist duration was retained as an offset (see eBird observation model).

The final models (after Step 2) included nine ecological covariates for AMWI and BUFF and ten for CAGO (Fig. 4). Of the initial set of 14 candidate ecological covariates, only “% of open water”

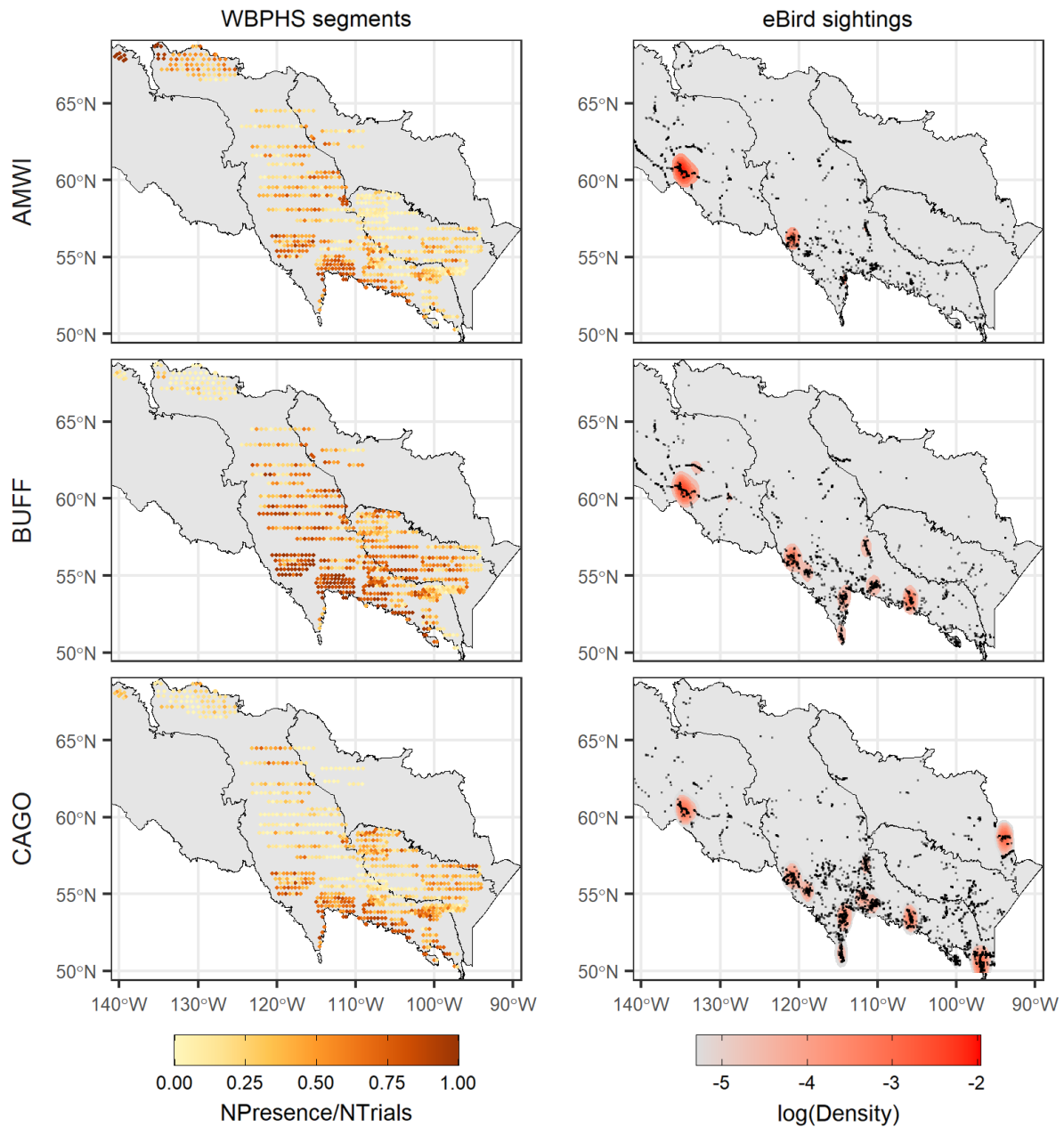


Fig. 3. Spatial distribution of the data extracted from the Waterfowl Breeding Population and Habitat Survey (WBPHS) (left) and eBird (right) data sets. $N_{\text{Presence}}/N_{\text{Trials}}$: segment-level ratio between the numbers of years the species was detected at least once (N_{Presence} ; min = 0 and max = 27) and the numbers of years the segment was surveyed (N_{Trials} ; min = 14 and max = 27). $\log(\text{Density})$: logged kernel density surface of eBird records (black dots).

was not selected in the final models. The highest number of shared covariates was between BUFF and CAGO ($n = 8$) and the lowest between AMWI and BUFF ($n = 5$). Four covariates were

shared by the three species: “% of wetland,” “% AGTB of *Picea glauca*,” “% AGTB of *Picea mariana*,” and “% AGTB of *Populus tremuloides*” (AGTB: aboveground tree biomass).

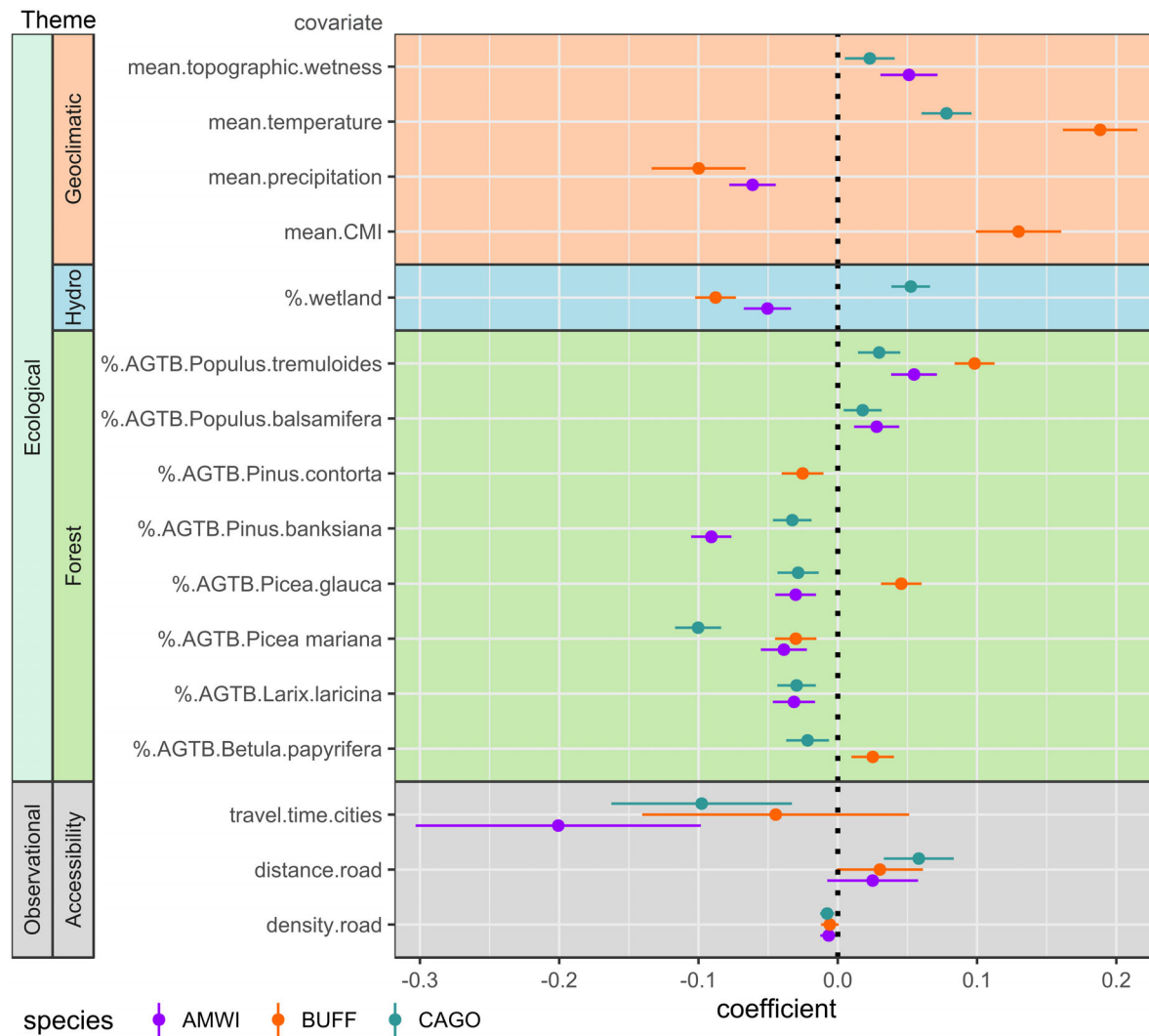


Fig. 4. Posterior estimates of the regression coefficients (mean and 95% credible intervals) of the point process intensity models for American wigeon (AMWI), bufflehead (BUFF), and Canada goose (CAGO). See Table 1 for details on covariates.

Spatial predictions.—For each species, the maps of the predicted intensity and SD are shown in Fig. 5. The broad spatial patterns in predicted intensity were visually close between species, with the highest degree of similarity found between BUFF and CAGO ($R^2 = 0.32$) and the lowest between AMWI and BUFF ($R^2 = 0.08$). The most obvious common pattern was found in the southern part ($<60^\circ$ N) of BCR 6, where the three models predicted the highest densities. Consistent with the presence of the largest eBird cluster (Fig. 3), a high-intensity spot was

predicted for the three species around 61° N and 134° W.

For the three species, the SD was relatively low (<0.10) and homogeneous across our study area (Fig. 5). Locally higher SD occurred in areas where the model covariates had the highest values. For example, the SD hotspot for BUFF predictions around 60° N and 128° W (Fig. 5, middle row) matches the high *Pinus contorta* biomass values in this area (Appendix S1: Fig. S2).

The R^2 values between (1) the predictions obtained as a function of both the ecological

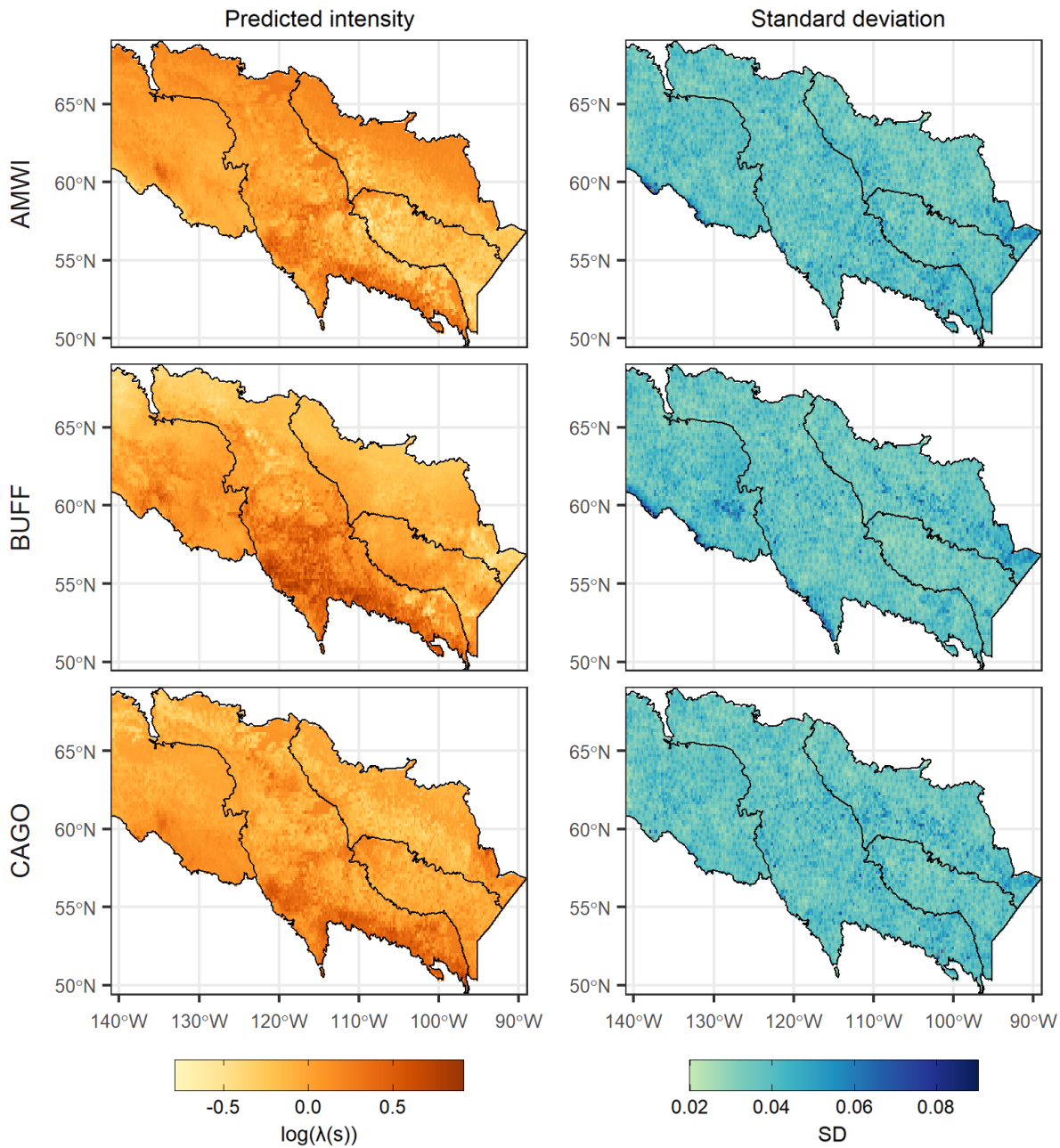


Fig. 5. Maps of the predicted intensity ($\log(\lambda(s))$) and standard deviation (SD) from the integrated models estimated for American wigeon (AMWI), bufflehead (BUFF), and Canada goose (CAGO).

covariates and the random field, and (2) the predictions obtained as a function of the ecological covariates only were 0.99, 0.98, and 0.97 for BUFF, AMWI, and CAGO, respectively. Conversely, the R^2 between (1) and (2) the predictions obtained as a function of the random field only

dropped to 0.23, 0.17, and 0.08 for CAGO, BUFF, and AMWI, respectively. These results and the associated prediction maps (Appendix S1: Fig. S4) revealed the main contribution of the ecological covariates in the integrated model predictions. The random field remained useful for

capturing local hotspots of high intensity (visually related to eBird clusters; Fig. 3) that the covariates alone could not explain (Appendix S1: Fig. S4).

Integrated vs. single data set models

Comparison of covariate estimates.—The posterior estimates of the coefficients obtained after refitting our models with a single data set are shown for a representative subset of ecological covariates in Fig. 6 (see Appendix S1: Tables S4–S6 for all covariates). The absolute coefficient values were the highest for WBPBS-only models in 25 out of the 28 possible cases. In contrast, they were the lowest for the eBird-only models (20/28). The coefficient values for the integrated models were generally in between (19/28). With regard to the 95% CrI, many covariates (15/28) lost their effect in the eBird-only models, revealing the complexity of capturing an ecological signal from eBird data. In a few cases, the eBird model coefficients remained significant but had an opposite effect to those revealed in the integrated and WBPBS-only models. For example,

this occurred for the percentage of wetlands with BUFF or *Picea glauca* with CAGO and AMWI, all of which became positive in the eBird-only models (Fig. 6).

Comparison of spatial predictions.—The area-weighted cumulative percentages of the total predicted intensity (PTPI) and uncertainty (PTPU; i.e., SD) accounted for by each BCR for the integrated and single data set predictions are shown in Fig. 7. The associated prediction maps are provided in Appendix S1: Figs. S5 and S6.

Across the four BCRs, WBPBS-only PTPI was the most heterogeneous ($SD = 7.51$) (Fig. 7, bottom panel). The highest WBPBS-only PTPI occurred in BCR 6, ranging from 37% for BUFF to 32% for CAGO. The highest positive differences between WBPBS-only and integrated PTPI were also found in BCR 6 (≥ 5 percentage points). Conversely, the highest negative differences between WBPBS-only and integrated PTPI were found in BCR 4 for both AMWI (9 percentage points) and CAGO (6 percentage points). eBird-only PTPI was highly homogeneous across the four BCRs ($SD = 0.88$) (Fig. 7, bottom panel).

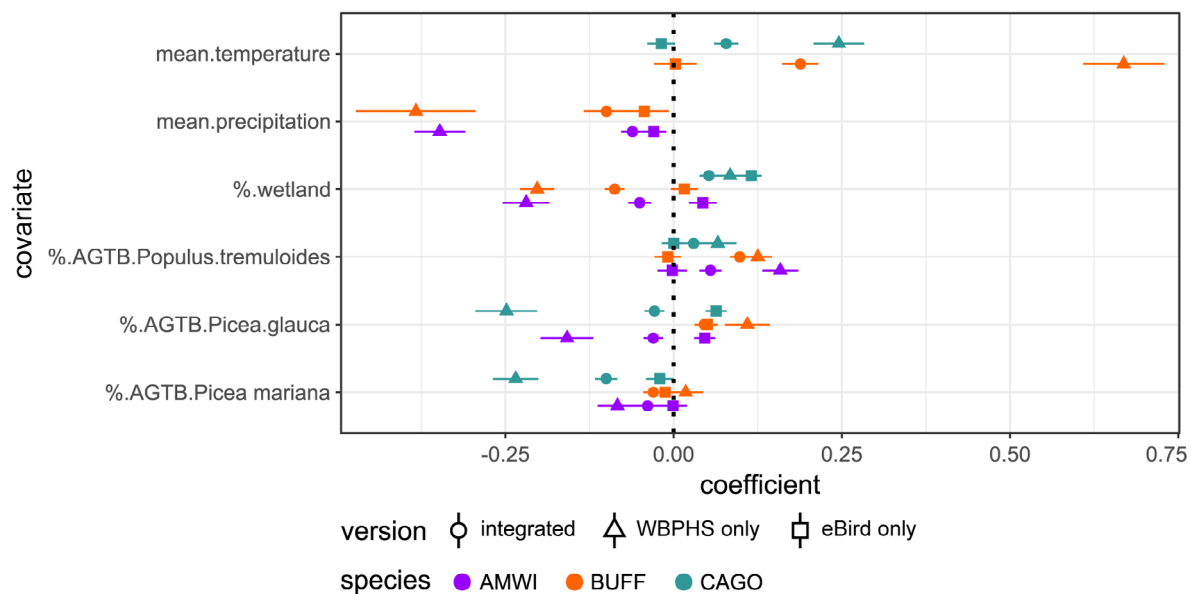


Fig. 6. Comparative posterior estimates of the regression coefficients (mean and 95% credible intervals) obtained for the models fitted with the Waterfowl Breeding Population and Habitat Survey data only (WBPBS only), the eBird data only (eBird only), and the two data sets (integrated). See Table 1 for details on covariates. See Appendix S1: Tables S4–S6 for all covariates. AMWI: American wigeon; BUFF: bufflehead; CAGO: Canada goose.

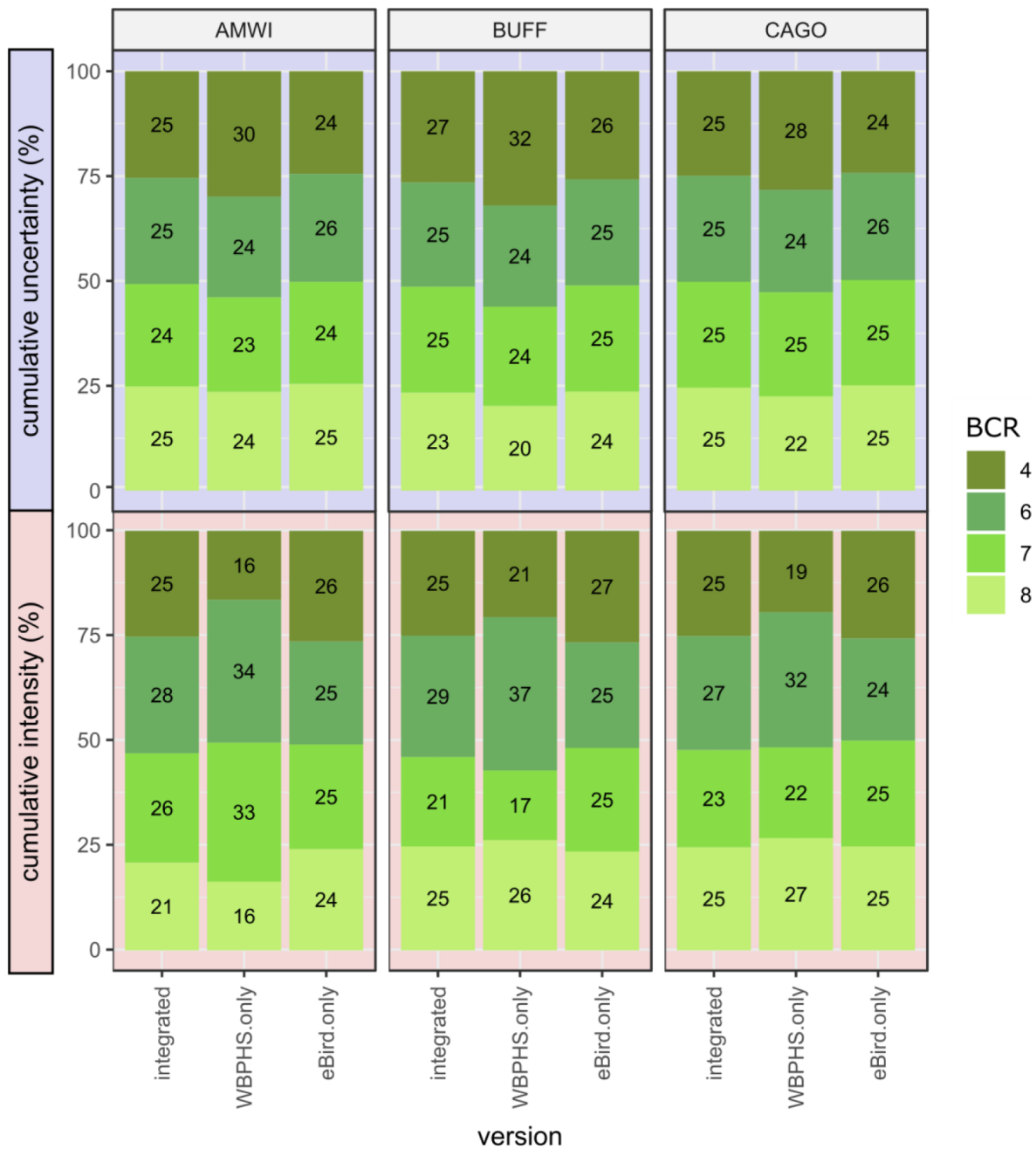


Fig. 7. Cumulative area-weighted percentages of total predicted intensity (bottom) and uncertainty (i.e., standard deviation) (top) accounted for by each bird conservation region (BCR) for the models fitted with the two data sets (integrated), the Waterfowl Breeding Population and Habitat Survey data only (WBPHS only), and the eBird data only (eBird only). AMWI: American wigeon; BUFF: bufflehead; CAGO: Canada goose.

The highest PTPU for the three species came from WBPHS-only predictions and was found in BCR 4 (Fig. 7, top panel). Compared to WBPHS-only PTPU (SD = 3.38), integrated PTPU was

more homogeneously distributed across the four BCRs (SD = 0.90) and was lower in the areas poorly covered by the WBPHS (negative difference up to 5 percentage points in BCR 4). PTPU

derived from the eBird-only predictions was spatially homogeneous ($SD = 0.69$).

Visually, the WBPHS-only predictions were close to the integrated predictions (Appendix S1: Fig. S5). This was not the case for the eBird-only predictions, which were relatively flat across our study area, although a spot of higher intensity in BCR 4 was identifiable for the three species (Appendix S1: Fig. S5). At the grid-cell level, the R^2 between the WBPHS-only and integrated predictions were 0.90, 0.77, and 0.68 for BUFF, CAGO, and AMWI, respectively. For the eBird-only predictions, these values decreased to 0.02, 0.01, and 0.01 for AMWI, BUFF, and CAGO, respectively.

Comparison of model fits.—For the three species, the R^2 between recorded segment-level NP/NT and the intensity predicted at the segment centers revealed that integrated predictions had the best fits to recorded data ($R^2 = 0.39 \pm 0.06$), although these values were close to those obtained for WBPHS-only predictions ($R^2 = 0.37 \pm 0.05$). eBird-only predictions poorly fitted recorded data ($R^2 = 0.01 \pm 0.01$).

DISCUSSION

In this method paper, we successfully adapted the ISDM approach, recently conceptualized by Isaac et al. (2020), for modeling and mapping the large-scale distribution of North American boreal waterfowl. Using a state-space point process formulation, we showed that it is possible to efficiently combine aerial survey (WBPHS) and citizen science (eBird) data to benefit from the complementarity of their records and spatial extents. By allowing for multiple observation models, the state-space formulation facilitated data integration while allowing explicit accounting for how the data sets were generated. This was useful to distinguish the ecological signal explaining the “true” species distribution from purely observational processes. This hierarchical formulation presents the possibility of using distinct observation models for each data set, allowing them to be treated separately and making the most of their specificities. In our case, we took advantage of the standardized and replicable design of the aerial survey to model WBPHS data as a binomial detection/non-detection process, while opportunistic eBird data were treated as

presence-only records. With minimal effort, other data types can be considered (e.g., point counts or expert range maps), along with multiple complementary data sets with associated observation models.

Explicitly modeling the observational processes was necessary for an optimal use of eBird data. Indeed, citizen science records are most often affected by observational biases related to sampling effort and site accessibility (Warton et al. 2013, Bonnet-Lebrun et al. 2020, Sicacha-Parada et al. 2020). Although preliminary filtering procedures are useful for removing outlier records and improving the general quality of the inference (Boria et al. 2014, Robinson et al. 2018, Steen et al. 2019), they probably remain insufficient to account for the observational noise in these data. This issue is perhaps most evident for remote study areas, such as ours, where available citizen science records are massively clustered around a few hotspots corresponding to the main human settlements. In this respect, our study area is certainly one of the places where this approach is the most challenging to apply, but where a real need also exists. Our results confirmed the importance of the data generation process: For all three example species considered in our study, the covariate selection procedure identified four out of the five candidate observational covariates (“checklist duration,” “distance to road,” “road density,” and “travel time to city”) as being important to model waterfowl distribution across our study area. The purpose of our study was mostly demonstrational and, for this reason, we used covariates that were already available. Thus, it is very likely that refined observational covariates could have enhanced the quality of the models. Particularly, using a threshold <50,000 inhabitants for the covariate “travel time to cities” (Weiss et al. 2018) could have been of particular interest for our study area, as we showed that smaller cities, such as Whitehorse (~25,000 inhabitants), were major hotspots for eBird records.

The “true” species distribution was modeled according to a flexible LGCP aimed at estimating the continuous density of points (i.e., individuals) in an area (Møller et al. 1998). The point process approach is becoming increasingly popular in ecological sciences, in particular thanks to the development of tools and R-packages facilitating

its application (Illian et al. 2013, Illian and Burslem 2017, Bachl et al. 2019). However, it remains far from being the norm in species distribution modeling studies, which are dominated by the grid-cell paradigm. Discretizing the spatial domain into grid cells comes at the expense of fixing the spatial scale and makes it difficult to consider within-grid heterogeneity (Illian and Burslem 2017, Miller et al. 2019, Isaac et al. 2020). Conversely, in theory, point processes are spatially continuous, so there are no more problems of scale. Although the scale issue does not exist, in practice, it is converted to an approximation problem. To model a continuous intensity surface, we divided our study area into triangles whose vertices served as integration points (Simpson et al. 2016). In the exploratory mesh design defining the triangle location and number, we sought a trade-off between the computational costs and the quality of the approximation. Complementary analyses conducted on a spatial subset of our study area (Appendix S1: Fig. S7) and the entire CWBF (Appendix S1: Table S7) revealed that differences in alternative mesh resolutions had little influence on model estimates.

The predicted intensities were obtained as a function of both the ecological covariates and random fields. All but one of the 14 candidate ecological covariates were selected in at least one of the models. The significance of the candidate covariates was not surprising, as they were chosen with regard to the existing literature (Adde et al. 2020a, b, Adde et al. 2021). Because of the easily interpretable model parameters, it was possible to assess the ecological meaning of our results. We confirmed the importance and predictive value of the Canada's forest attribute covariates (Beaudoin et al. 2017) to model the large-scale distribution of Canadian waterfowl (Adde et al. 2020a). For the three waterfowl species we considered, the intensity of the point process increased with an increase in the AGTB of *Populus tremuloides*, but decreased with the two *Picea* species. The negative effect of the percentage of wetlands on AMWI and BUFF, and the non-significance of the percentage of open water for the three species might seem more unexpected. However, at the spatial extent of our analysis (~3 million km²), and at the scale at which the ecological covariates were extracted (300-arcsecond), it is likely that there is a

confounding effect between causal relationships and spatial proxies. In other words, for some covariates such as the percentage of wetlands, the significance of the association could simply reflect a large-scale spatial correspondence between recorded waterfowl and covariate distributions without any causality link. It is also probable that wetland–waterfowl associations would be more easily identifiable at a finer spatial scale.

The random fields were able to capture the many otherwise-unexplained residuals that were independent of the ecological covariates, including spatial autocorrelation. These residuals were gathered around the main clusters of the eBird records. To avoid issues related to the overfitting of the random field revealed during preliminary analyses, we chose to use fixed prior values for the parameters controlling the range and the standard deviation of the field. We recognize that this strategy is somewhat subjective and goes against the complete reproducibility of the approach, as other prior values would probably be more suitable if different data were used. In addition to the extremely clustered spatial distribution of eBird data, issues related to the overfitting of the random field and eBird covariate coefficient estimates with unexpected magnitude could have also been linked to other important differences in the spatial processes underlying the two data sets (i.e., point observations for eBird and transect observations for WBPBS). Because each data set exhibits specific spatial patterns and degrees of autocorrelation, it is difficult to specify spatial parameters that can adequately account for these factors simultaneously. Complementary analyses aimed at tackling these issues showed that neither increasing the mesh resolution (Appendix S1: Table S7) or including an independent and identically distributed effect (IID) based on unique mesh node identifiers (Appendix S1: Table S8) provided an effective solution.

Comparing models fitted with different data sets is not straightforward because their information criteria are not comparable. In our study, we compared the models fitted with either the two data sets (integrated) or a single data set (WBPBS and eBird only) based on their coefficients and spatial predictions. Although this comparative analysis was useful to assess the

potential contribution of each of the two data sets, it remains complex to determine which approach (integrated vs. single data set) provides the “best” results and to quantify the contribution of each. When available, the use of a third independent and reliable data source might help answer these questions, but this seems to run counter to the general philosophy of the ISDM approach, which is to take advantage of the maximum amount of available data. Overall, we demonstrated that the outputs resulting from the integrated and WBPHS-only models were quite similar and consistent with the recorded species occurrences. Compared to the WBPHS-only models, the integrated models seem to have the advantage of smoothing the uncertainty more homogeneously across our study area, particularly in the areas not covered by the WBPHS. It was much more difficult to extract an ecological signal from the eBird data, for which the observational noise seemed to dominate. Accordingly, spatial predictions obtained with eBird-only data did not match the recorded species distribution. From an ecological modeling perspective, for our study area, it appears that citizen science records alone are still far from being robust alternatives to standardized aerial inventory data. Interestingly, spatial predictions derived from the integrated models matched the WBPHS records slightly better than the WBPHS-only models. The greater amount of information obtained by combining the two data sets could have helped refine the associations with the ecological covariates, resulting in a better fit. It also shows that although the eBird data remain extremely noisy, their role in the final model after consideration of the observational process is non-zero and even seems to improve the quality of the predictions.

CONCLUSION

By enabling the combination of all available data sources within a single hierarchical modeling framework, we demonstrated the potential of the ISDM approach for modeling and mapping large-scale species distributions. We encourage future North American waterfowl modeling attempts to use this method to resolve spatial gaps in the WBPHS coverage. Although this will require a considerable effort in data preparation,

integration efforts should not be restricted to the additional contribution of eBird data, of which the contribution at the scale of our study area proved to be limited. As multiple data observation models can be added to the original framework, we encourage testing for the potential contribution of provincial atlases, helicopter surveys, and all smaller past inventories. Notably, the ISDM approach applied in our study can be easily reproduced and transferred to other taxa or study areas by using the R-package “PointedSDMs” (<https://github.com/oharar/PointedSDMs>), which facilitates both data preparation and model formulation.

ACKNOWLEDGMENTS

This work was supported by a Natural Sciences and Engineering Research Council of Canada Strategic Partnership Grant for Projects (NSERC STPGP 494135) and by the Boreal Avian Modelling Project (www.borealbirds.ca). A. Adde benefited from additional scholarships from the Institut Hydro-Québec en environnement, développement et société, Laval University, and the Centre d'étude de la forêt. The U.S. Fish and Wildlife Service and Canadian Wildlife Service collected and supplied the Waterfowl Breeding Population and Habitat Survey data and we acknowledge their support regarding the database. This research was enabled in part by the support provided by Calcul Québec (www.calculquebec.ca) and Compute Canada (www.computeCanada.ca).

LITERATURE CITED

- Adde, A., N. Barker, M. Darveau, and S. Cumming. 2020a. Predicting spatiotemporal abundance of breeding waterfowl across Canada: a Bayesian hierarchical modelling approach. *Diversity and Distributions* 26:1248–1263.
- Adde, A., M. Darveau, N. Barker, L. Imbeau, and S. Cumming. 2021. Environmental covariates to model the distribution and abundance of breeding ducks in northern North America: a review. *Écoscience* 28:33–52.
- Adde, A., D. Stralberg, T. Logan, C. Lepage, S. Cumming, and M. Darveau. 2020b. Projected effects of climate change on the distribution and abundance of breeding waterfowl in Eastern Canada. *Climatic Change* 162:2339–2358.
- Amano, T., J. D. L. Lamming, and W. J. Sutherland. 2016. Spatial gaps in global biodiversity information and the role of citizen science. *BioScience* 66:393–400.

- Bachl, F. E., F. Lindgren, D. L. Borchers, and J. B. Illian. 2019. *inlabru*: an R package for Bayesian spatial modelling from ecological survey data. *Methods in Ecology and Evolution* 10:760–766.
- Barker, N. S. 2015. Modelling waterfowl abundance and distribution to inform conservation planning in Canada. Dissertation. Laval University, Quebec, Canada.
- Barker, N., S. Cumming, and M. Darveau. 2014. Models to predict the distribution and abundance of breeding ducks in Canada. *Avian Conservation and Ecology* 9:7.
- Beaudoin, A., P. Y. Bernier, P. Villemaire, L. Guindon, and X. J. Guo. 2017. Tracking forest attributes across Canada between 2001 and 2011 using a k nearest neighbors mapping approach applied to MODIS imagery. *Canadian Journal of Forest Research* 48:85–93.
- Blangiardo, M., and M. Cameletti. 2015. *Spatial and spatio-temporal Bayesian models with R-INLA*. John Wiley & Sons, Chichester, UK.
- Bonnet-Lebrun, A.-S., A. A. Karamanlidis, M. de Gabriel Hernando, I. Renner, and O. Gimenez. 2020. Identifying priority conservation areas for a recovering brown bear population in Greece using citizen science data. *Animal Conservation* 23:83–93.
- Boria, R. A., L. E. Olson, S. M. Goodman, and R. P. Anderson. 2014. Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. *Ecological Modelling* 275:73–77.
- Bowler, D. E., E. B. Nilsen, R. Bischof, R. B. O'Hara, T. T. Yu, T. Oo, M. Aung, and J. D. Linnell. 2019. Integrating data from different survey types for population monitoring of an endangered species: the case of the Eld's deer. *Scientific Reports* 9:1–14.
- Cameletti, M., F. Lindgren, D. Simpson, and H. Rue. 2013. Spatio-temporal modeling of particulate matter concentration through the SPDE approach. *ASTA Advances in Statistical Analysis* 97:109–131.
- Carver, E. 2015. Economic impact of waterfowl hunting in the United States: addendum to the 2011 National Survey of Fishing, Hunting, and Wildlife-associated Recreation. US Fish and Wildlife Service, Division of Economics, Falls Church, Virginia, USA.
- Chandler, M., et al. 2017. Contribution of citizen science towards international biodiversity monitoring. *Biological Conservation* 213:280–294.
- Devictor, V., R. J. Whittaker, and C. Beltrame. 2010. Beyond scarcity: citizen science programmes as useful tools for conservation biogeography. *Diversity and Distributions* 16:354–362.
- Dickinson, J. L., B. Zuckerberg, and D. N. Bonter. 2010. Citizen science as an ecological research tool: challenges and benefits. *Annual Review of Ecology, Evolution, and Systematics* 41:149–172.
- Doherty, K. E., J. S. Evans, J. Walker, J. H. Devries, and D. W. Howerter. 2015. Building the foundation for international conservation planning for breeding ducks across the US and Canadian border. *PLOS ONE* 10:e0116735.
- Dorazio, R. M. 2014. Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. *Global Ecology and Biogeography* 23:1472–1484.
- Ducks Unlimited. 2020. High-priority habitats: an overview of North America's most important waterfowl landscapes. Ducks Unlimited Inc, Memphis, Tennessee, USA.
- ECCC. 2020. List of birds protected by the Migratory Birds Convention Act in Canada, 1994. Environment and Climate Change Canada, Ottawa, Ontario, Canada.
- Fick, S. E., and R. J. Hijmans. 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology* 37:4302–4315.
- Fink, D., T. Auer, A. Johnston, V. Ruiz-Gutierrez, W. M. Hochachka, and S. Kelling. 2020. Modeling avian full annual cycle distribution and population trends with citizen science data. *Ecological Applications* 30:e02056.
- Fithian, W., J. Elith, T. Hastie, and D. A. Keith. 2015. Bias correction in species distribution models: pooling survey and collection data for multiple species. *Methods in Ecology and Evolution* 6:424–438.
- Fletcher, R. J., T. J. Hefley, E. P. Robertson, B. Zuckerberg, R. A. McCleery, and R. M. Dorazio. 2019. A practical guide for combining data to model species distributions. *Ecology* 100:e02710.
- Fournier, A. M. V., A. R. Sullivan, J. K. Bump, M. Perkins, M. C. Shieldcastle, and S. L. King. 2017. Combining citizen science species distribution models and stable isotopes reveals migratory connectivity in the secretive Virginia rail. *Journal of Applied Ecology* 54:618–627.
- Geldmann, J., J. Heilmann-Clausen, T. E. Holm, I. Levinsky, B. Markussen, K. Olsen, C. Rahbek, and A. P. Tøttrup. 2016. What determines spatial bias in citizen science? Exploring four recording schemes with different proficiency requirements. *Diversity and Distributions* 22:1139–1149.
- Gelman, A., J. Hwang, and A. Vehtari. 2014. Understanding predictive information criteria for Bayesian models. *Statistics and Computing* 24:997–1016.
- Humphreys, J. M., J. L. Murrow, J. D. Sullivan, and D. J. Prosser. 2019. Seasonal occurrence and abundance

- of dabbling ducks across the continental United States: joint spatio-temporal modelling for the Genus *Anas*. *Diversity and Distributions* 25:1497–1508.
- Illian, J. B., and D. F. R. P. Burslem. 2017. Improving the usability of spatial point process methodology: an interdisciplinary dialogue between statistics and ecology. *AStA Advances in Statistical Analysis* 101:495–520.
- Illian, J. B., S. Martino, S. H. Sørbye, J. B. Gallego-Fernández, M. Zunzunegui, M. P. Esquivias, and J. M. J. Travis. 2013. Fitting complex ecological point process models with integrated nested Laplace approximation. *Methods in Ecology and Evolution* 4:305–315.
- Isaac, N. J. B., et al. 2020. Data integration for large-scale models of species distributions. *Trends in Ecology & Evolution* 35:56–67.
- Isaac, N. J. B., A. J. van Strien, T. A. August, M. P. de Zeeuw, and D. B. Roy. 2014. Statistics for citizen science: extracting signals of change from noisy ecological data. *Methods in Ecology and Evolution* 5:1052–1060.
- Johnsgard, P. A. 2010. Waterfowl of North America: hunting and recreational values of North American waterfowl. Pages 23–30 *in* Waterfowl of North America, Revised Edition (2010). Univ. of Nebraska Press, Lincoln, Nebraska, USA.
- Kéry, M., and J. A. Royle. 2016. Applied hierarchical modeling in ecology: analysis of distribution, abundance and species richness in R and BUGS: Volume 1: prelude and static models. Academic Press.
- Krainski, E. T., V. Gómez-Rubio, H. Bakka, A. Lenzi, D. Castro-Camilo, D. Simpson, F. Lindgren, and H. Rue. 2019. Advanced spatial modeling with stochastic partial differential equations using R and INLA. Chapman & Hall/CRC Press, Boca Raton, Florida, USA.
- Latifovic, R., D. Pouliot, and I. Olthof. 2017. Circa 2010 land cover of Canada: local optimization methodology and product development. *Remote Sensing* 9:1098.
- Lindgren, F., and H. Rue. 2015. Bayesian spatial modelling with R-INLA. *Journal of Statistical Software* 63:1–25.
- Lindgren, F., H. Rue, and J. Lindström. 2011. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73:423–498.
- Mack, G., and D. Morrison. 2006. Waterfowl of the boreal forest. Alberta Pacific Forest Industries Inc., St. Albert, Alberta, Canada.
- Mattsson, B. J., J. A. Dubovsky, W. E. Thogmartin, K. J. Bagstad, J. H. Goldstein, J. B. Loomis, J. E. Diffendorfer, D. J. Semmens, R. Wiederholt, and L. López-Hoffman. 2018. Recreation economics to inform migratory species conservation: case study of the northern pintail. *Journal of Environmental Management* 206:971–979.
- McKinley, D. C., et al. 2017. Citizen science can improve conservation science, natural resource management, and environmental protection. *Biological Conservation* 208:15–28.
- Miller, D. A. W., K. Pacifici, J. S. Sanderlin, and B. J. Reich. 2019. The recent past and promising future for data integration methods to estimate species' distributions. *Methods in Ecology and Evolution* 10:22–37.
- Møller, J., A. R. Syversveen, and R. P. Waagepetersen. 1998. Log gaussian cox processes. *Scandinavian Journal of Statistics* 25:451–482.
- NABCI. 2014. Bird conservation regions - bird studies Canada on behalf of the North American Bird Conservation Initiative. NABCI, Port Rowan, Ontario, Canada. <http://www.birdscanada.org/research/gislab/index.jsp?targetpg=bcr>
- Nichols, J. D., and B. K. Williams. 2006. Monitoring for conservation. *Trends in Ecology & Evolution* 21:668–673.
- Opitz, T., F. Bonneu, and E. Gabriel. 2020. Point-process based Bayesian modeling of space-time structures of forest fire occurrences in Mediterranean France. *Spatial Statistics* 40:100429.
- Pacifici, K., B. J. Reich, D. A. Miller, B. Gardner, G. Stauffer, S. Singh, A. McKerrow, and J. A. Collazo. 2017. Integrating multiple data sources in species distribution modeling: a framework for data fusion. *Ecology* 98:840–850.
- Pagel, J., B. J. Anderson, R. B. O'Hara, W. Cramer, R. Fox, F. Jeltsch, D. B. Roy, C. D. Thomas, and F. M. Schurr. 2014. Quantifying range-wide variation in population trends from local abundance surveys and widespread opportunistic occurrence records. *Methods in Ecology and Evolution* 5:751–760.
- Pocock, M. J., J. C. Tweddle, J. Savage, L. D. Robinson, and H. E. Roy. 2017. The diversity and evolution of ecological and environmental citizen science. *PLOS ONE* 12:e0172579.
- Renner, I. W., J. Elith, A. Baddeley, W. Fithian, T. Hastie, S. J. Phillips, G. Popovic, and D. I. Warton. 2015. Point process models for presence-only analysis. *Methods in Ecology and Evolution* 6:366–379.
- Robinson, O. J., V. Ruiz-Gutierrez, and D. Fink. 2018. Correcting for bias in distribution modelling for rare species using citizen science data. *Diversity and Distributions* 24:460–472.

- Rue, H., S. Martino, and N. Chopin. 2009. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71:319–392.
- Sicacha-Parada, J., I. Steinsland, B. Cretois, and J. Borgelt. 2020. Accounting for spatial varying sampling effort due to accessibility in Citizen Science data: a case study of moose in Norway. *Spatial Statistics* 42:100446.
- Silverman, B. W. 1986. *Density estimation for statistics and data analysis*. CRC Press, Boca Raton, Florida, USA.
- Simpson, D., J. B. Illian, F. Lindgren, S. H. Sørbye, and H. Rue. 2016. Going off grid: computationally efficient inference for log-Gaussian Cox processes. *Biometrika* 103:49–70.
- Simpson, D., H. Rue, A. Riebler, T. G. Martins, and S. H. Sørbye. 2017. Penalising model component complexity: a principled, practical approach to constructing priors. *Statistical Science* 32:1–28.
- Smith, G. W. 1995. A critical review of the aerial and ground surveys of breeding waterfowl in North America. U.S. Fish and Wildlife Service, Laurel, Maryland, USA.
- Sorenson, L. G., R. Goldberg, T. L. Root, and M. G. Anderson. 1998. Potential effects of global warming on waterfowl populations breeding in the northern Great Plains. *Climatic Change* 40:343–369.
- Soriano-Redondo, A., C. M. Jones-Todd, S. Bearhop, G. M. Hilton, L. Lock, A. Stanbury, S. C. Votier, and J. B. Illian. 2019. Understanding species distribution in dynamic populations: a new approach using spatio-temporal point process models. *Ecography* 42:1092–1102.
- Statistics Canada. 2012. National Road Network feature catalogue segmented view. Statistics Canada Statistical Geomatics Centre, Ottawa, Ontario, Canada.
- Steen, V. A., C. S. Elphick, and M. W. Tingley. 2019. An evaluation of stringent filtering to improve species distribution models from citizen science data. *Diversity and Distributions* 25:1857–1869.
- Strimas-Mackey, M., W. M. Hochachka, V. Ruiz-Gutierrez, O. J. Robinson, E. T. Miller, T. Auer, S. Kelling, D. Fink, and A. Johnston. 2020. Best practices for using eBird data. Version 1.0. Cornell Lab of Ornithology, Ithaca, New York, USA.
- Sullivan, B. L., et al. 2014. The eBird enterprise: an integrated approach to development and application of citizen science. *Biological Conservation* 169:31–40.
- Title, P. O., and J. B. Bemmels. 2018. ENVIREM: an expanded set of bioclimatic and topographic variables increases flexibility and improves performance of ecological niche modeling. *Ecography* 41:291–307.
- USFWS. 2019. Waterfowl population status, 2019. U.S. Department of the Interior, Washington, DC USA.
- USFWS. 2020. List of migratory birds protected by the Migratory Bird Treaty Act. U.S. Fish and Wildlife Service, Falls Church, Virginia, USA.
- Warton, D. I., I. W. Renner, and D. Ramp. 2013. Model-based control of observer bias for the analysis of presence-only data in ecology. *PLOS ONE* 8: e79168.
- Watanabe, S. 2010. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research* 11:3571–3594.
- Weiss, D. J., et al. 2018. A global map of travel time to cities to assess inequalities in accessibility in 2015. *Nature* 553:333–336.

SUPPORTING INFORMATION

Additional Supporting Information may be found online at: <http://onlinelibrary.wiley.com/doi/10.1002/ecs2.3790/full>