Predicting spatiotemporal abundance of breeding waterfowl across Canada: A Bayesian hierarchical modelling approach

Antoine Adde1 | Marcel Darveau1,2 | Nicole Barker3 | Steven Cumming1

1Department of Wood and Forest Sciences, Laval University, Quebec, QC, Canada
2Ducks Unlimited Canada, Quebec, QC, Canada
3Canadian Wildlife Service, Environment and Climate Change Canada, Edmonton, AB, Canada

Correspondence
Antoine Adde, Department of Wood and Forest Sciences, 2405 Rue de la Terrasse, Laval University, Quebec, QC, Canada G1V 0A6.
Email: antoine.adde.1@ulaval.ca

Funding information
Natural Sciences and Engineering Research Council of Canada, Grant/Award Number: 494135

Editor: Ana Benítez López

Abstract
Aim: Our aim was to develop predictive statistical models for mapping the abundance of 18 waterfowl species at a pan-Canadian level. We refined the previous generation of national waterfowl models by (a) developing new, more interpretable statistical models that (b) explicitly account for spatiotemporal variations in waterfowl abundance, while (c) testing for associations with an updated suite of habitat covariates.

Location: All of Canada, excluding the Northern Arctic ecozone.

Methods: Our response variables were annual species counts on 2,227 aerial-survey segments over a period of 25 years (1990–2015). Combining machine-learning and hierarchical regression modelling, we devised an innovative covariate selection strategy to select for each species the best subset of a panel of 232 candidate habitat covariates. With the selected covariates, we implemented hierarchical generalized linear models in a Bayesian framework, using the integrated nested Laplace approximation and stochastic partial differential equation approaches.

Results: On average, our models explained 47% of the observed variance for spatiotemporal predictions and 74% for temporally averaged spatial predictions. The 18 species models included 94 significant waterfowl-habitat associations involving 42 distinct habitat covariates, with an average of 5.3 covariates per model. Covariates for forest attributes were the most represented in our models. The proportional biomass of *Populus tremuloides* was the most frequently selected covariate (10/94 associations in 10/18 species). Model predictions generated spatial and spatiotemporal maps of species abundances over almost all of Canada.

Main conclusions: We showed that it is possible to efficiently combine machine-learning, variable selection and hierarchical Bayesian methods that exploit high-dimensional covariate spaces. Our approach yielded powerful and easily interpretable species distribution models with very few covariates, while accounting for residual autocorrelation. Possible applications of the resulting models and maps include the development of biodiversity indicators, the evaluation and execution of conservation planning strategies, and ecosystem services monitoring.
1 | INTRODUCTION

Since 1955, the Waterfowl Breeding Population and Habitat Survey (WBPHS) has produced annual counts of breeding waterfowl species over ≈30-km segments (Smith, 1995). The survey area now extends from the United States Prairie region to the Southern Arctic and from the Atlantic coast to Alaska. While the WBPHS was originally designed to inform hunting regulation (Nichols, Johnson, & Williams, 1995), the data from this survey have also widely served waterfowl ecological research. In particular, several studies have benefitted from the temporal and spatial coverage of the WBPHS to develop species distribution models (SDMs) (Barker, Cumming, & Darveau, 2014; Doherty, Evans, Walker, Devries, & Howarter, 2015).

By predicting species occurrence from habitat covariates (Elith & Leathwick, 2009b), SDMs can be used to prioritize waterfowl-habitat conservation (Barker et al., 2014), anticipate their responses to climate change (Stein, Skagen, & Noon, 2014) and evaluate their role in disease transmission (Belkhiri, Hijmans, Boyce, Crossley, & Martínez-López, 2018). Available North American waterfowl SDMs have been developed mostly at the regional level, in particular for the prairie biome (Doherty et al., 2015; Reese & Skagen, 2017; Stein et al., 2014). To date, only Barker et al. (2014) have used the WBPHS for constructing waterfowl SDMs at pan-Canadian extent. They used boosted regression trees to model the spatial abundances of 17 waterfowl species as a function of a suite of 78 biophysical and climatic covariates. The models constructed by Barker et al. (2014) have been applied to the spatial targeting of waterfowl habitats for conservation by Ducks Unlimited Canada and its partners (Prairie Habitat Joint Venture, 2016) and to ecosystem services monitoring by the Government of Canada (Malouin, Doyle, & Liss, 2013).

Developing practitioner-interpretable SDMs is important to identify actionable insights and facilitate model applications (Elith & Leathwick, 2009b). Although machine-learning (ML) methods, such as boosting regression trees, are known for fast and accurate predictions, their outputs can be problematic to interpret in terms of effect sizes, mechanisms or underlying ecological processes (Bucklin et al., 2015; Marmion, Parviainen, Boyce, Crossley, & Martínez-Minaya, Cameletti, Conesa, & Pennino, 2018; Wikle & Hooten, 2010). These methods have been applied to modelling waterfowl abundance in North America (Humphreys, Murrow, Sullivan, & Prosser, 2019; Ross, Hooten, DeVink, & Koons, 2015; Ross et al., 2012).

In the spirit of continuous improvement, our study aimed to revisit the Canadian breeding waterfowl models constructed by Barker et al. (2014) by (a) developing new, more interpretable statistical models that (b) explicitly account for spatiotemporal variations in waterfowl abundance, while (c) testing for associations with an updated suite of habitat covariates. Our objective was to implement hierarchical GLMs (HGLMs) in a Bayesian framework to explicitly predict and map the spatiotemporal pan-Canadian abundance of 18 breeding waterfowl species over a 25-year period, using a suite of 232 candidate habitat covariates. This high dimensionality required defining an innovative, robust and flexible modelling strategy to cope with computational requirements and ensure both parsimony and high predictive power in the final models.

2 | METHODS

2.1 | Study area

Our study area included all of Canada, excluding the Northern Arctic ecozone (CEC, 1997) and other adjacent Arctic islands, with a total area of ≈8.5 million km² (Figure 1).
2.2 | Waterfowl data

Each spring, aerial observers of the WBPHS count adults of common waterfowl species seen within 200 m transects subdivided into segments of ≈28.8 km, using fixed-wing aircraft (Smith, 1995). We retrieved segment-level annual counts of waterfowl species conducted over a period of 25 years (1990–2015) in the Canadian portion of the WBPHS. This period was defined to match the 1990 expansion of the survey area into eastern Canada and the availability of habitat covariates. The data from 2007 were excluded because of a deviation from the usual survey design (Barker et al., 2014). To identify a temporal signal and ensure representativeness, only the segments sampled for ≥5 years (n = 2,227 out of the 2,287 available segments; Figure 1) were considered in our analyses.

We extracted data for 14 waterfowl species and four species groups, each group including two or three species difficult to distinguish by the survey protocol (Table 1). These four groups are referred to as “species” hereafter. Based on the work of Barker et al. (2014), Johnsgard (2010) and Poole (2005), we classified the studied species into three nesting guilds: cavity, ground and overwater (Table 1). For a better characterisation of the breeding population and to avoid male-biased sex ratios, we transformed the raw waterfowl counts to total indicated pairs (TIP) based on species-specific life histories (Barker, 2015; Dzubin, 1969; Smith, 1995). To account for incomplete detection in estimating continental population sizes, visibility correction factors (VCFs) exist to adjust the WBPHS counts (Smith, 1995). However, these visibility adjustments are generated by the use of widely different methods at various frequencies and spatial scales across Canada (Barker, 2015; C. Roy, Environment and Climate Change Canada, 2020, personal communication). As explained by Barker et al. (2014), these differences preclude the use of VCFs to model detections effort at the segment scale.

2.3 | Habitat covariates

Given the ecological diversity of North American waterfowl species and the wide range of habitat covariates that are potentially relevant for modelling their abundance and distribution, we adopted an exploratory multistage variable selection approach (see Section 2.5). We derived a suite of 232 candidate covariates from eight distinct habitat datasets (Table 2). Among the 232 covariates, 128 were spatial and 104 were spatiotemporal. The covariates were classified as “Geoclimatic” (92/232), “Hydrologic” (42/232), “Land cover” (63/232) and “Forest attributes” (35/232). We adopted a covariate naming convention of metric–variable combinations (e.g., mean [metric] elevation [variable]).

As we had no information on the exact location of individual waterfowl occurrences within the 28.8 km segments, we had to characterize the waterfowl–environment associations operating at the scale of the area surveyed within each segment. Although the survey protocol records birds detected within 200 m of either side of the flight line, the flight line itself may not be precisely as mapped and may vary slightly among years, due to navigational error. We sampled segment-scale covariates in a 1 km buffer surrounding the mapped segment centre lines. Preliminary correlation tests between waterfowl abundance and habitat covariates measured at seven different radii ranging from 500 m to 10 km provided similar results, suggesting relative scale insensitivity over the range of buffer widths evaluated. The probable underlying reason is that the environmental...
TABLE 1 Eighteen waterfowl species or species groups modelled in this study

| Scientific name               | Vernacular name | Abbreviation | Nesting guild |
|-------------------------------|-----------------|--------------|---------------|
| Bucephala albeola             | Bufflehead      | BUFF         | Overwater     |
| Bucephala clangula            | Goldeneye spp.  | GOLD         | Cavity        |
| Bucephala islandica           |                 |              |               |
| Lophodytes cucullatus         | Merganser spp.  | MERG         | Cavity        |
| Mergus merganser              |                 |              |               |
| Mergus serrator               |                 |              |               |
| Anas acuta                    | Northern pintail| NOPI         | Ground        |
| Anas crecca                   | Green-winged teal| AGWT       | Ground        |
| Anas platyrhynchos            | Mallard         | MALL         | Ground        |
| Anas rubripes                 | American black duck| ABDU    | Ground        |
| Branta canadensis             | Canada goose    | CAGO         | Ground        |
| Mareca americana              | American wigeon | AMWI         | Ground        |
| Mareca strepera               | Gadwall         | GADW         | Ground        |
| Melanitta americana           | Scoter spp.     | SCOT         | Ground        |
| Melanitta fusca               |                 |              |               |
| Melanitta perspicillata       |                 |              |               |
| Spatula clypeata              | Northern shovel | NSHO        | Ground        |
| Spatula discors               | Blue-winged teal| BWTE        | Ground        |
| Aythya affinis                | Scaup spp.      | SCAU         | Overwater     |
| Aythya marila                 |                 |              |               |
| Aythya americana              | Redhead         | REDH         | Overwater     |
| Aythya collaris               | Ring-necked duck| RNDU        | Overwater     |
| Aythya valisineria            | Canavasback     | CANV         | Overwater     |
| Oxyura jamaicensis            | Ruddy duck      | RUDU         | Overwater     |

2.4 Hierarchical spatiotemporal modelling

To predict the spatiotemporal abundance of 18 waterfowl species at a pan-Canadian level, we fit HGLMs in a Bayesian framework using the integrated nested Laplace approximations (INLA; Rue, Martino, & Chopin, 2009), a paradigm for Bayesian inference in latent Gaussian models. The main benefit of using INLA instead of the classical Markov Chain Monte Carlo techniques is computational. Being deterministic, the INLA approach needs no sampling, thus allowing faster estimation, even for complex spatiotemporal models with high dimensionality. We fit models using the R package "inlabru" (Bachl, Lindgren, Borchers, & Illian, 2019) that is built on the widely used "INLA" package (Lindgren & Rue, 2015).

For each individual species, our response variable TIP(s, t) was the TIP recorded for segment s in the year t. Segment centroids were used as model-point locations. The TIP counts exhibited overdispersion and excess zeros. Accordingly, we modelled TIP(s, t) using a zero-inflated negative binomial (ZINB) model, which can be described as a mixture of a point mass at zero with probability p and a negative binomial distribution with probability 1 − p; see Greene (1994) for details. Adopting the ZINB R-INLA type 2 likelihood (Breivik, Storvik, & Nedreaas, 2017; Cosandey-Godin, Krainski, Worm, & Flemming, 2014; Hu et al., 2016), we modelled TIP(s, t) as follows:

\[ P \{ \text{TIP}(s, t) = n \} = p(s, t) I_{n=0} + [1 - p(s, t)] \times \text{NB}[n, k, \theta(s, t)] \]  

(1)

where \( p(s, t) \) is an additional probability for zero; \( I_{n=0} \) is an indicator function taking the value one, if true, and zero otherwise; and \( \text{NB}[n, k, \theta(s, t)] \) is the negative binomial distribution with the dispersion parameter k.

The expectation \( \mu(s, t) \) of the negative binomial distribution in (Equation 1) is linked to the linear predictor \( \eta(s, t) \) by \( \mu(s, t) = \exp \{ \eta(s, t) \} \) where

\[ \eta(s, t) = X(s, t) \beta + u(s, t) \]  

(2)

with \( X(s, t) \) a design matrix with fixed covariates, \( \beta = [\beta_0, \ldots, \beta_k] \) is the vector of regression coefficients, and \( u(s, t) \) is a spatiotemporal random effect. Under a ZINB R-INLA type 2 likelihood, the additional probability for zero \( p(s, t) \) in Equation (1) is

\[ p(s, t) = 1 - \left( \frac{\exp \{ \eta(s, t) \}}{1 + \exp \{ \eta(s, t) \}} \right)^* \]  

(3)
where the parameter $\alpha$ adjusts for a zero-probability change. This probability is inversely proportional to the linear predictor $\eta(s, t)$.

The spatiotemporal random effect $u(s, t)$ represents other processes not explained by the covariates, including spatial autocorrelation. It was modelled as a spatially continuous Gaussian random field with yearly realisations (dependent replicates with year as a grouping structure), using the stochastic partial differential equation (SPDE) approach (Lindgren, Rue, & Lindström, 2011). Detailed and applied examples of the INLA-SPDE approach can be found elsewhere (e.g., Blangiardo & Cameletti, 2015; Cameletti, Lindgren, Simpson, & Rue, 2013; Krainski, Lindgren, Simpson, & Rue, 2019). Briefly, the SPDE approach evaluates spatial covariance

| Class (no. of covariates) | Dataset (source) | Dimensions | Variable | Metric |
|---------------------------|------------------|------------|----------|--------|
| Geoclimatic (n = 92/232)  | Canada's digital elevation model (NRCAN, 2013) | Spatial $y = 2000s$ $r = 50 m$ | Elevation | Mean; SD |
|                           | Climate-ANUSPLIN (McKenney et al., 2011) | Spatiotemporal $y = 1990$–2015 $r = 10 km$ | Temperature (min; max) Precipitations (sum) Climate moisture index | $y$-mean; $y$-SD; mon-mean; cu.mon-mean |
|                           | GIMMS-NDVI (Tucker et al., 2005) | Spatiotemporal $y = 1990$–2015 $r = 8 km$ | NDVI | $y$-mean; $y$-SD; mon-mean; cu.mon-mean |
| Hydrologic (n = 42/232)   | Canada's Wetland map (Amani et al., 2019) | Spatial $y = 2010s$ $r = 30 m$ | Wetland and water cover: 6 classes | Class level<sup>a</sup>: edge density; mean patch area; patch density; % of landscape |
|                           | Land Cover of Canada (Latifovic, Pouliot, & Olthof, 2017) | Spatial $y = 2010$ $r = 30 m$ | Wetland and water cover: 2 classes | Class level<sup>a</sup>: edge density; mean patch area; patch density; % of landscape |
|                           | National Hydro Network (NRCAN, 2004) | Spatial $y = 2000s$ $r = 1:50,000$ | Hydrological linear features: 10 classes | Class level: total length |
| Land cover (n = 63/232)   | Land Cover of Canada (Latifovic et al., 2017) | Spatial $y = 2010$ $r = 30 m$ | Land cover (non-wetland/water): 13 classes | Class level<sup>a</sup>: edge density; mean patch area; patch density; % of landscape |
|                           | kNN-Canada's forest attributes (Beaudoin et al., 2017) | Spatial $y = 2011$ $r = 250 m$ | Landscape fragmentation | Landscape level<sup>a</sup>: patch density; patch richness density; edge density |
|                           | National Road Network (Statistics Canada, 2012) | Spatial $y = 2010s$ $r = 1:50,000$ | Vegetation cover: 4 classes | Class level: % of landscape |
| Forest attributes (n = 35/232) | kNN-Canada's forest attributes (Beaudoin et al., 2017) | Spatial $y = 2011$ $r = 250 m$ | Tree species composition: 24 species or groups | Species level: % of aboveground tree biomass (AGTB) |
|                           | | | Forest biomass: 6 classes | Class level: mean |
|                           | | | Forest volume: 2 classes | Forest age |
|                           | | | Forest height | Mean |
|                           | | | Crown closure | % of landscape |

Note: $n$, number of covariates per class; $y$, period or year of data acquisition; $r$, spatial resolution; SD, standard deviation. The metrics “$y$-mean,” "mon-mean" and "cu.mon-mean" referred to the three temporal schemes used for spatiotemporal covariate extraction, including (a) annual means; (b) monthly means for the January–May period; and (c) cumulative 1-month interval means for the January–May period (i.e., past climate from May to April, May to March, May to February and May to January), respectively.

<sup>a</sup>For details on the classes/variables, see Table S1.1.

<sup>b</sup>For details on the class and landscape metrics, see Table S1.2.
by modelling the spatial variation as a Gaussian Markov random field (GMRF) with zero mean and covariance defined by a Matérn correlation function. The GMRF was modelled and projected across our entire study area based on a spatial mesh created by a constrained refined Delaunay triangulation. The field values at model-point locations were interpolated using the weighted average of the three closest mesh nodes. The triangle sizes were small in the areas covered by the WBPHS but grew larger as the density of observations decreased (Figure S1.1). To avoid edge effects, the mesh was extended outside the domain of interest. The mesh design was exploratory, where we sought a trade-off between the computational costs and precision of the spatial field. The priors for model hyperparameters were set to the R-INLA “non-informative” defaults (see Text S1.1 for details). As in a frequentist approach, this strategy allows the data to predominate in calculating posterior distributions.

2.5 | Four-step habitat covariate selection

To optimize the predictive abilities, interpretability and parsimony of the final model, as well as to handle the computational challenges related to the high dimensionality of our dataset, we defined a four-step habitat covariate selection strategy based on (1) correlation tests and collinearity diagnosis; (2) component-wise gradient boosting with a stability selection procedure; (3) forward covariate selection; and (4) structured HGLM. This strategy was applied individually to each waterfowl species.

2.6 | Step 1: Correlation tests

As a pre-processing step, we performed bivariate Pearson’s correlation tests between TIP and each of the 232 candidate habitat covariates. The covariates were evaluated sequentially in decreasing order of the absolute values of the correlation coefficients. The covariates were retained if they were “non-collinear” with any previously added covariate. The collinearity threshold was defined as a Pearson correlation coefficient of 0.70 in absolute value. A maximum of 50 covariates were retained.

2.7 | Step 2: Component-wise gradient descent boosting

To identify the most influential habitat covariates among the 50 selected in Step 1, we fit GLMs using an iterative component-wise gradient descent boosting (CGDB) algorithm with a stability selection procedure (Hofner, Boccuto, & Göker, 2015; Hothorn, Bühlmann, Kneib, Schmid, & Hofner, 2010; Mayr & Hofner, 2018) as implemented in the R package “mboost” (Hothorn, Bühlmann, Kneib, Schmid, & Hofner, 2013). CGDB is a ML method that generates prediction rules that have the same interpretation as common statistical models. For a schematic overview of the CGDB algorithm, see box 1 in the paper of Mayr, Hofner, and Schmid (2012). In short, the first step of the process involves computing the negative gradient of a pre-selected loss function; we used a negative binomial log-likelihood as the loss function. In the following step, the covariates were fitted separately to a negative gradient vector using their corresponding base learners (i.e., GLMs). Once all potential base learners had been tested, the best-fitting one was added to the current model, and the negative gradient was re-evaluated at the updated model fit. This procedure was repeated for a species-specific number of iterations, which was determined by an initial stability selection procedure to identify the covariates that were most commonly selected after applying the CGDB algorithm on 100 random subsamples of half of the available TIP data. For details on this procedure, see Hofner et al. (2015). We set the number of selected base learners (i.e., the number of covariates) per boosting run to 15, and the upper bound of the per-family error rate to 2 with unimodality assumption. Given these specifications, the covariates selected in at least 90 of the 100 subsamples were considered stable and were retained for further analysis in Step 3.

2.8 | Step 3: Forward covariate selection

A forward covariate selection (FCS) in conjunction with a modified version of the HGLM described in Section 2.4 was conducted to search for the best combination of the remaining covariates. To reduce the computational costs of FCS, we used a spatially and temporally unstructured version of the HGLM, with random intercepts for segment and year. The combinations were ranked by the deviance information criteria (DIC), which measure the compromise between the fit and parsimony of the model (Rue et al., 2009; Spiegelhalter, Best, Carlin, & Linde, 2002). In FCS, all covariate pairs were tested, and the best-performing pair with the lowest DIC was retained. The number of covariates was iteratively increased, maximizing the decrease in DIC until no remaining covariate decreased the model DIC. The covariates included in the model with the lowest DIC were retained for Step 4.

2.9 | Step 4: Structured HGLM

The structured HGLM of Section 2.4 was fit using the covariates from Step 3. We examined the resulting 95% credible intervals of the covariate coefficients and identified those in which the coverage included 0. If there were any, a final model was fit without any of the “non-significant” covariates.

2.10 | Habitat covariate selection summary

We summarized the representation of the four habitat covariate classes ("Geoclimatic," "Hydrologic," “Land cover” and “Forest
attributes in the final models by calculating the percentage of the total covariates that belonged to a given class weighted by the number of non-collinear candidate covariates in that class. We set the collinearity threshold to an absolute Pearson correlation coefficient of 0.70, as before. This statistic was calculated for the 18 species individually as well as by nesting guild (Table 1). Because covariates were classified into four classes, the weighted percentage for each covariate class should have been approximately 25%, if covariates were selected in proportion to their availability. We graphically reported individual associations included in the final models, with information on the magnitude and sign of the associations based on the estimated regression coefficients.

2.11 | Model checking

We interpreted the model inferential properties using the means, standard deviations (SD) and 95% credible intervals of the posterior distributions of model parameters. For all 18 species, we provided a model summary table of these statistics for each model parameter. We also reported the nominal range, or the distance at which the spatial correlation modelled by the GMRF declines to 0.1 (Krainski et al., 2018); high nominal ranges would indicate that there is important residual spatial autocorrelation.

Model goodness of fit was assessed by computing the coefficient of determination ($R^2$), that is the proportion of variance in recorded TIP explained by the model, using the median of the posterior predictive distribution (i.e., the distribution of the predicted data). The $R^2$ values were computed for both spatiotemporal ($n = 47,020$ observations) and temporally averaged spatial (by segment; $n = 2,227$ observations) predictions. To evaluate the specific contribution of the random effect in model predictions, we re-computed the $R^2$ values using the predicted GMRF mean values. Finally, we computed the value of $p$ between the temporally averaged observed and predicted values to allow comparison with the results of Barker et al. (2014), who used this goodness-of-fit metric. A two-sample t test was used to compare the differences in $p$ between the two studies.

INLA provides leave-one-out cross-validation model checks as a part of the fitting process (Held, Schrödle, & Rue, 2010), including the computation of conditional predictive ordinates (CPO; Pettit, 1990). CPO analysis, which computes a leave-one-out predictive distribution for each observation, is a computationally less-demanding alternative to traditional K-fold cross-validation, which requires refitting each model K times to obtain a complete set of out-of-sample predictions. Details on the CPO analysis and its results are available in Supplementary material S5.1.

2.12 | Predictive mapping

A regular grid of 300-arcsecond resolution (cell area ranging from 26.6 to 63.9 km$^2$) covering the study area (Figure 1) was used for the predictions. Estimated predictions from fitted models were computed via the "predict" function of the "inlabru" R library (Bachl et al., 2019) that works by repeatedly drawing samples from the posterior distributions of the model parameters. Mapped predictions were obtained as function of both environmental covariates and the estimated GMRF. We provided maps of the yearly median in the units of TIP/km$^2$ and coefficient of variation (CV; SD/mean) of the predictions for each species in each grid cell. We also provided the annual maps of the GMRF.

3 | RESULTS

3.1 | Model building and parameter estimates

Our models were successfully fit for all 18 species (see Tables S2.1–S2.18 for model summaries). The tables presenting habitat covariates retained at the first three selection steps (1: correlation tests; 2: CGDB; and 3: FCS) for the 18 species are given in Appendix S3. On average, $50 \pm 0.0$ habitat covariates were retained after Step 1, $11.6 \pm 1.4$ after Step 2 and $6.9 \pm 2.0$ after Step 3. The mean number of habitat covariates included in the final models (after Step 4) was $5.3 \pm 2.2$, ranging from eight for the goldeneye, northern pintail and ring-necked duck to one for the ruddy duck (Tables S2.1–S2.18).

We illustrate with a model of average performance obtained for bufflehead (Table 3), a cavity-nesting species with continuous distribution across Canada. The final bufflehead model included five habitat covariates from three datasets, with an $R^2$ value of .42. The proportions of the aboveground biomass of white spruce (Picea glauca) and balsam poplar (Populus balsamifera), as well as of the standard deviation of annual NDVI, were positively associated with bufflehead abundance. Conversely, bufflehead abundance was negatively associated with the mean patch areas of the tundra-like vegetation classes.

In average, for all 18 species, the spatial autocorrelation was high, with a mean nominal range of $8.23 \pm 3.32$. We obtained a maximum nominal range value of 19.73 for American black duck and a minimum of 5.45 for merganser. Significant overdispersion and zero inflation were found for all 18 species. The means of the overdispersion parameter $k$ and zero-inflation parameter $\alpha$ were $1.90 \pm 0.67$ and $0.51 \pm 0.24$, respectively.

3.2 | Goodness of fit

Table 4 and Figures S4.1–S4.2 provide the goodness-of-fit assessments of the 18 models for spatiotemporal (ST) and temporally averaged spatial (TAS) predictions. For ST predictions, the mean $R^2$ value between the recorded TIPs and the predicted median was $0.47 \pm 0.16$. The mean $R^2$ value computed with only GMRF values dropped to $0.17 \pm 0.07$, demonstrating the contribution of both random and fixed effects in the predictive performances of our models. The TAS predictions had a higher mean $R^2$ value of $0.74 \pm 0.12$ (Table 4).
### TABLE 3

Summaries of the posterior distribution of the parameters (mean, standard deviation (SD) and quantiles) estimated for the final bufflehead model

| Parameter | Mean   | SD    | $Q_{0.025}$ | $Q_{0.50}$ | $Q_{0.975}$ |
|-----------|--------|-------|-------------|------------|-------------|
| **Covariates [dataset]** |        |       |             |            |             |
| Intercept | -1.75  | 0.68  | -3.10       | -1.75      | -0.41       |
| Prop._AGTB Picea glauca [kNN-CFA] | 0.15   | 0.01  | 0.13        | 0.15       | 0.18        |
| Prop._AGTB Populus balsamifera [kNN-CFA] | 0.13   | 0.01  | 0.11        | 0.13       | 0.15        |
| Mean patch area_sub-polar or polar grassland-lichen-mod [LCC] | -0.15  | 0.03  | -0.21       | -0.15      | -0.10       |
| Mean patch area_sub-polar or polar barren-lichen-moss [LCC] | -0.08  | 0.03  | -0.14       | -0.08      | -0.02       |
| sd_yearly NDVI [NDVI] | 0.29   | 0.02  | 0.26        | 0.29       | 0.33        |
| **Hyperparameters** |        |       |             |            |             |
| Overdispersion parameter | 1.76   | 0.06  | 1.64        | 1.75       | 1.88        |
| Zero-inflation parameter | 0.41   | 0.02  | 0.36        | 0.41       | 0.45        |
| GMRF nominal range | 9.35   | 8.17  | 10.61       |            |             |

Note: See Tables S2.1–S2.18 for the final models of all species.

*prop._AGTB Picea glauca*: the proportions of the aboveground biomass of white spruce (*Picea glauca*).

*prop._AGTB Populus balsamifera*: the proportions of the aboveground biomass of balsam poplar (*Populus balsamifera*).

“sd_yearly NDVI”: The standard deviation of annual NDVI.

Abbreviation: GMRF, Gaussian Markov random field.

### TABLE 4

Goodness-of-fit model evaluation metrics for the 18 species and comparison with the study of Barker et al. (2014), who modelled all 17 of our 18 species, excepting only Canada goose

| Species | ST $R^2$ Value | TAS $R^2$ Value | ST Spearman $\rho$ | TAS Spearman $\rho$ | Barker et al. (2014) TAS $R^2$ Value | Barker et al. (2014) SPEARMAN $\rho$ |
|---------|----------------|-----------------|-------------------|--------------------|--------------------------------------|-------------------------------------|
| SCOT    | .76            | .94             | 0.79              | 0.75               |                                      |                                     |
| MALL    | .69            | .85             | 0.96              | 0.95               |                                      |                                     |
| BWTE    | .66            | .85             | 0.87              | 0.82               |                                      |                                     |
| NOPI    | .64            | .83             | 0.88              | 0.82               |                                      |                                     |
| NSHO    | .64            | .82             | 0.88              | 0.85               |                                      |                                     |
| GADW    | .61            | .82             | 0.85              | 0.80               |                                      |                                     |
| AMWI    | .50            | .82             | 0.90              | 0.87               |                                      |                                     |
| CAGO    | .48            | .77             | 0.90              | NA                 |                                      |                                     |
| SCAU    | .44            | .71             | 0.89              | 0.82               |                                      |                                     |
| BUFF    | .42            | .68             | 0.88              | 0.83               |                                      |                                     |
| AGWT    | .40            | .71             | 0.82              | 0.72               |                                      |                                     |
| CANV    | .39            | .69             | 0.82              | 0.79               |                                      |                                     |
| REDH    | .39            | .62             | 0.78              | 0.72               |                                      |                                     |
| MERG    | .37            | .76             | 0.90              | 0.86               |                                      |                                     |
| ABDU    | .32            | .69             | 0.86              | 0.82               |                                      |                                     |
| GOLD    | .29            | .65             | 0.81              | 0.75               |                                      |                                     |
| RNDU    | .27            | .61             | 0.77              | 0.68               |                                      |                                     |
| RUDU    | .18            | .45             | 0.64              | 0.57               |                                      |                                     |

Note: n, sample size in segment-years or segments.

Abbreviations: ST, spatiotemporal predictions; TAS, temporally averaged spatial predictions.
Furthermore, we calculated Spearman’s correlation \( (\rho) \) between the recorded TIPs and TAS predictions for comparing our models with those of Barker et al. (2014), who reported this statistic (Table 4). A significant difference in \( \rho \) was observed between our models \((0.84 \pm 0.07)\) and Barker et al. (2014) \((0.79 \pm 0.09)\); \( t(16) = 9.4, p < .01 \). The correlation between our two sets of Spearman’s correlations was \( \rho = 0.96, p < .01 \), indicating that the rank orders of model performance among species was very nearly identical in the two studies.

### 3.3 | Habitat covariate selection

Using guild-level stratification, Figure 2 shows the representation of the four covariate classes in the 18 models compared to the number of non-collinear candidate covariates in each class. Only “Forest attribute” covariates (37.90%) were selected more frequently than expected (>25%) given their availability. The “Hydrologic” (22.29%), “Geoclimatic” (21.23%) and “Land cover” (18.58%) covariates were selected less frequently than expected. “Forest attribute” was the single most highly represented covariate class for ground (40.96%) nesters, but not for overwater (31.93%) and cavity-nesting (36.90%) species. The covariates from the “Hydrologic” class were the most frequently selected in cavity nester models (42.18%). For overwater nesters, the “Geoclimatic” class was the most represented one (32.58%).

The 18 models included a total of 94 significant waterfowl-habitat associations involving 42 distinct habitat covariates (Figure 3). The “Forest attribute” class accounted for 34/94 associations, of which 23 were positive. The proportional biomass of *Populus tremuloides* was the most frequently selected covariate with 10/94 positive associations in 10/18 species (Figure 3). The second most selected “Forest attribute” covariate was the proportional biomass of *Picea mariana* accounting for 5/32 associations, all of which were negative (Figure 3).

Thirteen covariates of the “Land cover” class accounted for 24/94 associations, of which 17 were negative. Among the 13 selected “Land cover” covariates, only three were reported as associations for more than one species (Figure 3). These three covariates included “edge density of cropland” (4/5 positive associations and 1/5 negative association with merganser), “patch density of sub-polar taiga needle-leaf forest” (4/4 negative associations) and “proportion of treed land cover” (3/3 negative associations).

The “Hydrologic” class contributed 20/94 associations (Figure 3), of which 19 were positive. The most commonly selected covariates in this class were “patch density of water” (4/20), “proportion of water” (3/20) and “proportion of shallow water” (3/20). Only 1/20 associations, namely “edge density of deep water,” were negative, for green-winged teal (Figure 3).

The covariates of the “Geoclimatic” class accounted for 16/94 associations (Figure 3), of which nine were negative and seven were positive (Figure 3). All 16 of these associations were either NDVI (9/16) or elevation-related (7/16) covariates. The two most commonly selected “Geoclimatic” covariates included the standard deviation of elevation, which was negatively associated with four species, and the standard deviation of monthly NDVI within years, which was positively associated with four species. No direct climatic covariates, such as temperature or precipitation, were selected in any model.

### 3.4 | Predictive mapping

For each species, the maps of the predicted TIP and GMRF are available in Appendix S5.1. In addition, the spatial data objects (TIF files) containing predictions with technical information to import them into an R workspace are provided in Appendix S5.2. We illustrate with temporally averaged sample maps for American black duck, bufflehead and canvasback (Figure 4). The recorded TIP (Figure 4, first row) visibly matches with the spatial patterns of the predicted TIP (Figure 4, second row). The CV values (Figure 4, third row) were high in British Columbia, Yukon, Nunavut and northern Quebec, where the survey data were absent. Within the surveyed distribution area of the species, the CV values were low and homogeneous. The general spatial patterns of GMRF (Figure 4, fourth row) were close to the predicted TIP, but with a coarser level of details. The annual fluctuations in the spatial patterns of the predicted TIP were low (see spatiotemporal sequences provided in Appendix S5.1).

### 4 | DISCUSSION

#### 4.1 | Modelling approach

We demonstrated the capacity of our Bayesian hierarchical modelling approach to predict spatiotemporal variations in the abundance...
of 18 waterfowl species across Canada. On average, the model variance explained was 47% for spatiotemporal predictions and 74% for TAS predictions. We took advantage of the strengths of both ML and GLM paradigms to construct our models. We exploited the rapidity and power of the ML techniques, in particular of gradient-descent boosting (Hothorn et al., 2010), to select highly informative subsets of an initial set of 232 candidate covariates before fitting HGLMs. Considering the high dimensionality of our dataset, using a HGLM approach only would have been computationally infeasible. This strategy led to robust, parsimonious and interpretable models that exceeded the predictive power of the pure ML models of Barker et al. (2014). We recognized the need for caution when comparing results among different modelling techniques. Our sample size was 40% larger than theirs (2,227 vs. 1,591), and it has been shown that GLMs improve their predictive accuracy as the sample size increases and might even surpass ML models for sufficiently large samples (Aguirre-Gutiérrez et al., 2013). The improved performance of our models might also be related to our richer suite of candidate covariates and/or to the explicit consideration of latent spatiotemporal processes via random fields included in the models.

Our waterfowl dataset was statistically challenging in being overdispersed and zero-inflated count data with discontinuous temporal and spatial domains. Taking advantage of the powerful INLA-SPDE approach (Lindgren & Rue, 2015), our analysis allowed the estimation of waterfowl associations with habitat covariates and spatiotemporal random effects in the form of a GMRF. The Gaussian random fields were able to capture the much otherwise-unexplained heterogeneity. These structures presumably represent large-scale processes that were independent of the model covariates. The random effects revealed a marked spatial autocorrelation in the distributions of many species, especially those with restricted core distribution areas such as the American black duck.

Ignoring spatial and temporal correlations could have resulted in incorrect estimates of covariates’ effects and erroneous predictions. In our study, we used three indirect ways for disentangling the GMRF processes from the effects of environmental covariates. First,
as shown in Figure 4, the visual inspection of the GMRF revealed spatial structures indicating the presence of residual autocorrelation that could not be explained by the environmental covariates alone. Second, to evaluate the specific contribution of the GMRF in model predictions, we re-computed model the $R^2$ values, our goodness-of-fit metric, using only the spatial–temporal random effects. With covariates excluded, the mean $R^2$ value dropped from 0.47 to 0.17. Third, in the fourth step of our covariate selection process, we evaluated the possibility that covariates selected within an unstructured version (neither spatially nor temporally explicit) of the HGLM were “wrongfully” identified as significant because autocorrelation was not explicitly modelled. This was done by confronting the covariates with an alternative hypothesis via a spatiotemporally explicit version of the HGLM that includes the GMRF. On average, for the 18 waterfowl species, 1.6 covariates were dropped in this step, as they were no longer significant when autocorrelation was accounted for.

**FIGURE 4** Maps of temporally averaged (1990–2015) spatial records and predictions for American black duck, bufflehead and canvasback. Records: recorded total indicated pairs (TIP). Predictions: median of the posterior distribution. CV: coefficient of variation of the posterior distribution. GMRF: Mean Gaussian Markov random field (random effect). See Appendix S5.1 for the spatial record and prediction maps of all species.
4.2 | Waterfowl-habitat associations

Synthesis of duck-habitat associations extracted from the 18 models provided new insights into the environmental factors driving the distribution and abundance of breeding waterfowl in Canada. Searching for causal explanations for each of the 94 waterfowl-habitat covariate associations identified in the final models would be infeasible. However, thanks to our explicit consideration of spatial autocorrelation, it is possible to state that the identified habitat covariates were not simply proxies for biome-level characteristics; had that been the case, the relationships would have been subsumed within the random effect. We believe it is reasonable to conclude that all covariates retained in the final models do indicate habitat selection of species within their respective ranges. This must reflect, in at least some cases, underlying processes related to feeding and/or nesting behaviour. For example, the positive association we found between cropland abundance and mallard TIPs has previously been related to their foraging preferences (Forcey, Linz, Thogmartin, & Bleier, 2007; Lieske et al., 2018; Newbold & Eadie, 2004). Similarly, the positive association between deadwood biomass and goldeneye TIPs is easily interpreted in light of their cavity-nesting behaviour, although not, to our knowledge, previously demonstrated.

Forest-related covariates were the group most frequently selected in the final models, relative to their availability in the initial set. In contrast, no direct climatic covariates (e.g., covariates derived from measured precipitation or temperature) were selected in any model. Vegetation and forest patterns are driven by multi-scale factors, including regional climate, local topography, soil characteristics, hydrology and disturbance regime (Grime, 2006; Perry, Oren, & Hart, 2008). Thus, at the spatial resolution of our analysis (11.2 km² segments), the forest-related covariates derived from 6.25 ha imagery (Beaudoin, Bernier, Villesmaire, Guindon, & Guo, 2017) were “super-proxies” for possibly many underlying, interacting environmental conditions in which the waterfowl species select their breeding sites. This is well illustrated by the covariates related to the aboveground tree biomass of Populus species, which were included in 14/18 models. As with many waterfowl species, the trees of this genus are abundant along the prairie-boreal ecotone in Canada (Beaudoin et al., 2017) and in riparian habitats within the continuous boreal forest (Harper & Macdonald, 2001).

To our knowledge, ours was the first species distribution modelling study to use the new, detailed pan-Canadian wetland map (Amani et al., 2019), which was developed upon the five classes of the Canadian Wetland Classification System (i.e., bog, fen, marsh, swamp and shallow water; National Wetlands Working Group, 1997). This new product enabled us to go beyond the traditional finding that “Ducks like water” (Pimm, 1994) and to reveal the details of species-specific wetland-habitat preferences. For example, we found that two cavity-nesting species (bufflehead and merganser) were positively associated with the patch density of bogs, whereas northern pintail was positively associated with the patch density of fens. We hope our demonstration of this new wetland map will encourage its broader use and continued development.

4.3 | Limitations

Considerable efforts have been made over recent decades in developing national- or continental-level environmental layers relevant to species distribution modelling in general and to waterfowl ecology in particular (topography, climate, land cover including wetlands and forest cover attributes such as tree species biomass). Many of these products are derived from remote-sensed imagery by means of ML or other statistical approaches. They cannot be validated over their whole extent. In addition, many of these datasets are based on the imagery captured in a single year, which forced us to assume that several habitat covariates were static over the study years. Although we took great pains to acquire the best available data, we recognize that our ability to model spatially varying temporal trends remains limited.

SDMs depend on the assumption that species distribution and abundance patterns are, at least, partially determined by measurable environmental conditions and, thus, may be modelled using habitat covariates (Austin, 2007). Because the studied waterfowl species are migratory birds, they are presumably affected by environmental conditions outside as well as on their breeding grounds. By focusing on the breeding habitats only, our models missed a part of the interannual variability explained by varying conditions on migratory routes and wintering grounds (Franklin, 2010; McPherson & Jetz, 2007). The covariates indirectly related to the conditions outside the breeding season might have accounted for a part of this variability. We also acknowledge the importance of biotic factors, such as dispersal, competition and predation, as well as historical factors (Stralberg et al., 2017). Such factors are rarely incorporated into SDMs (Austin, 2007).

Despite the standardized WBPHS protocol that has been followed for more than 60 years, the data remain subject to imperfect detection due to spatial-temporal variation among observers, in environmental conditions and in species’ behaviours. This is generally the case with species count datasets (Kéry & Schmidt, 2008). This could affect our findings in terms of both predicted abundances and habitat associations identified. To account for incomplete detection, VCFs are estimated as part of the survey protocol to adjust the WBPHS counts (Smith, 1995). As VCFs are calculated in widely different ways throughout the survey, we believe that using them will not correct the predictions but rather will introduce a different source of bias in the results (Barker et al., 2014; C. Roy, Environment and Climate Change Canada, 2020, personal communication).

Finally, we acknowledge issues related to the segment-level measurement of our response variable. Segments are organized sequentially along transects within strata. Accordingly, treating segments as independent samples would have resulted in pseudo-replication. However, we believe that our spatial autocorrelation analysis substantially accounts for this spatial structure as well as for the repeated measurements. In addition, we recognized that representing segment locations by their centroids led to a loss in spatial accuracy in the sense that waterfowl abundances vary spatially within segments. However, the locations of individual observations within
segments are not available. We believe that the assumption of internal homogeneity was appropriate to our modelling approach.

4.4 Future directions and applications

Our model building strategy combines ML techniques for covariate selection and Bayesian hierarchical modelling to fit final models that are both parsimonious and interpretable. We believe this approach could be applicable to many future species distribution modelling studies that deal with similar high-dimensionality challenges posed by a vast array of potential covariates that are now available. Another benefit of the Bayesian hierarchical approach lies in the possibility of retrieving the uncertainty associated with each model parameter. Accounting for and reporting these uncertainties, such that transparent predictions of waterfowl abundance can be obtained, is essential, if the results are to be used in decision-making (Pearson et al., 2006). The extrapolation of our models allowed the generation of predictive annual and TAS maps of waterfowl abundance that could be used to define a variety of biodiversity indicators for a continuous spatial domain as extensive as Canada.

Individual waterfowl-habitat associations extracted from the 18 models could be useful for guiding future waterfowl modelling studies in the choice of potentially relevant candidate covariates. Indeed, if this work represents an important modelling effort at a national scale, refining individual waterfowl-habitat relationships at finer or regional scales might be necessary for local applications. Such work would be facilitated by building on the associations presented in this study, and if our models are proved wrong by future data, a specific knowledge will have been learned. Because the number of covariates in each model was small, each parameter estimate could be developed into formal ecological hypothesis about factors affecting individual species.

SDMs require the assumption that species are in equilibrium with their environment and that environmental gradients are homogenously sampled within the geographic space (Elith & Leathwick, 2009a). Our study does not fully satisfy, at least, the second requirement because in predicting at the national level, we considerably overstepped the spatial extent of the survey data. Consequently, the areas excluded from the coverage of the WBPHS, particularly those targeting Yukon and British Columbia, were correctly identified by the models as having high prediction uncertainty. Predictions made for these areas must not be used without reference to their uncertainty estimates. Nevertheless, we believe these predictions remain valuable for the spatial targeting of conservation areas, and, importantly, for pointing out locations where additional sampling efforts are needed and/or integration of complementary databases would be helpful. Indeed, other datasets with qualitative (e.g. Atlas of the Breeding Birds of British Columbia; Bird Studies Canada, 2019) and quantitative (e.g. cooperative roadside waterfowl breeding population surveys in Yukon; Hawkins & Hughes, 2008) information on breeding waterfowl exist for these areas. These datasets could be combined with the WBPHS through the use of integrated SDMs (Fletcher et al., 2019; Isaac et al., 2020; Miller, Pacifici, Sanderlin, & Reich, 2019). In addition, the recent incorporation of waterfowl data from citizen science within advanced spatiotemporal modelling framework has shown promising results (Humphreys et al., 2019).

ACKNOWLEDGEMENTS

This work was supported by a Natural Sciences and Engineering Research Council of Canada Strategic Partnership Grant for Projects (NSERC STGP 494135) and by the Boreal Avian Modelling Project (www.borealbirds.ca). Antoine Adde benefited of 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 waterfowl data, and we acknowledge their support regarding the database. We thank Dr. Meisam Amani (John Wood Group PLC) and Dr. Brian Brisco (Natural Resources Canada) for giving us access to the Canadian wetland map.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

PEER REVIEW

The peer review history for this article is available at https://publoverview.com/publon/10.1111/ddi.13129.

DATA AVAILABILITY STATEMENT

All data employed in the present work are public and can be downloaded from the original sources. A script that allows the reproduction of the formulation of our final models and predictions is provided in Appendix S6. The ZIP file containing the maps and spatial data objects can be downloaded at https://figshare.com/s/v9f45a6198c4b35c7910.

ORCID

Antoine Adde https://orcid.org/0000-0003-4388-0880
Marcel Darveau https://orcid.org/0000-0002-0143-9272
Nicole Barker https://orcid.org/0000-0002-7556-800X

REFERENCES

Afton, A. D., & Anderson, M. G. (2001). Declining scaup populations: A retrospective analysis of long-term population and harvest survey data. The Journal of Wildlife Management, 65(4), 781–796. https://doi.org/10.2307/3803028
Aguirre-Gutiérrez, J., Carvalheiro, L. G., Polce, C., van Loon, E. E., Raes, N., Reemer, M., & Biesmeijer, J. C. (2013). Fit-for-purpose: Species distribution model performance depends on evaluation criteria–Dutch hoverflies as a case study. PLoS One, 8(5), e63708. https://doi.org/10.1371/journal.pone.0063708
Amani, M., Mahdavi, S., Afshar, M., Brisco, B., Huang, W., Mohammad Javad Mirzadeh, S., ... Hopkinson, C. (2019). Canadian Wetland Inventory using Google Earth Engine: The first map and preliminary results. Remote Sensing, 11(7), 842. https://doi.org/10.3390/rs11070842
Austin, M. (2007). Species distribution models and ecological theory: A critical assessment and some possible new approaches. Ecological Modelling, 200(1-2), 1–19. https://doi.org/10.1016/j.ecolmodel.2006.07.005
Bachl, F. E., Lindgren, F., Borchers, D. L., & Illian, J. B. (2019). inlabru: An R package for Bayesian spatial modelling from ecological survey data. *Methods in Ecology and Evolution*, 10(6), 760–766. https://doi.org/10.1111/2041-210X.13168

Barker, N. (2015). *Modelling waterfowl abundance and distribution to inform conservation planning in Canada*. (PhD thesis). Laval University, Quebec, Canada.

Barker, N., Cumming, S., & Darveau, M. (2014). Models to predict the distribution and abundance of breeding ducks in Canada. *Avian Conservation and Ecology*, 9(2), 7. https://doi.org/10.5751/ACE-00669-090207

Beaudoin, A., Bernier, P. Y., Vallemaire, P., Guindon, L., & Guo, X. J. (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(1), 85–93.

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

Bird Studies Canada (2019). *Atlas of the breeding birds of British Columbia*. Retrieved from https://www.birdatlas.bc.ca

Blangiardo, M., & Cameletti, M. (2015). *Spatial and spatio-temporal Bayesian models with R-INLA*. Chichester, UK: John Wiley & Sons.

Breivik, O. N., Storvik, G., & Nedreass, K. (2017). Latent Gaussian models to predict historical bycatch in commercial fishery. *Fisheries Research*, 185, 62–72. https://doi.org/10.1016/j.fishres.2016.09.033

Browsencombe, J. W., Griffin, L. P., Gagne, T. O., Haak, C. R., Cooke, S. J., Finn, J. T., & Danylychuk, A. J. (2019). Environmental drivers of habitat use by a marine fish on a heterogeneous and dynamic reef flat. *Marine Biology*, 166(2), 18. https://doi.org/10.1007/s00227-018-3464-2

Bucklin, D. N., Basile, M., Bencsort, A. M., Brandt, L. A., Mazzotti, F. J., Romañach, S. S., ... Watling, J. I. (2015). Comparing species distribution models constructed with different subsets of environmental predictors. *Diversity and Distributions*, 21(1), 23–35. https://doi.org/10.1111/ddi.12247

Cameletti, M., Lindgren, F., Simpson, D., & Rue, H. (2013). Spatio-temporal modeling of particulate matter concentration through the SPDE approach. *Advances in Statistical Analysis*, 97(2), 109–131. https://doi.org/10.1007/s10182-012-0196-3

Commission for Environmental Cooperation. (1997). *Ecological regions of North America: Toward a common perspective*. Montreal, QC: Commission for Environmental Cooperation (Map updated 2006).

Cosandey-Godin, A., Krainski, E. T., Worm, B., & Flemming, J. M. (2014). Commission for Environmental Cooperation (Map updated 2006).

Cotter, J. I., Rutherford, S. M., Sain, S., & Vats, D. (2013). Stochastic search variable selection for approximate Bayesian computation. *Journal of Computational and Graphical Statistics*, 22(1), 77–99. https://doi.org/10.1080/10618600.2012.701793

Doherty, T. J., Hefley, T. J., Robertson, E. P., Zuckerberg, B., McCleery, R. A., & Dorazio, R. M. (2019). A practical guide for combining data to model species distributions. *Ecology*, 100(6), e02710. https://doi.org/10.1111/1466-8238.2019.01340.1

Fletcher, R. J., Hefley, T. J., Robertson, E. P., Zuckerberg, B., McCleery, R. A., & Dorazio, R. M. (2019). A practical guide for combining data to model species distributions. *Ecology*, 100(6), e02710. https://doi.org/10.1111/1466-8238.2019.01340.1

Fletcher, R. J., Hefley, T. J., Robertson, E. P., Zuckerberg, B., McCleery, R. A., & Dorazio, R. M. (2019). A practical guide for combining data to model species distributions. *Ecology*, 100(6), e02710. https://doi.org/10.1111/1466-8238.2019.01340.1

Forney, G. M., Linz, G. M., Thommartin, W. E., & Bleier, W. J. (2007). Influence of land use and climate on wetland breeding birds in the Prairie Pothole region of Canada. *Canadian Journal of Zoology*, 85(3), 421–436. https://doi.org/10.1139/Z07-005

Franklin, J. (2010). Moving beyond static species distribution models in support of conservation biogeography. *Diversity and Distributions*, 16(3), 321–330. https://doi.org/10.1111/j.1424-1345.2010.00110.x

Greene, W. H. (1994). Accounting for excess zeros and sample selection in poisson and negative binomial regression models. *Department of Economics Working Paper No. EC-94-10*. New York, NY: New York University.

Guisan, A., Tingley, R., Baumgartner, J. B., Naujokaitis-Lewis, I., Sutcliffe, P. R., Tulloch, A. I. T., ... Buckley, Y. M. (2013). Predicting species distributions for conservation decisions. *Ecology Letters*, 16(12), 1424–1435. https://doi.org/10.1111/ele.12189

Harper, K. A., & Macdonald, S. E. (2001). Structure and composition of riparian boreal forest: New methods for analyzing edge influence. *Ecology*, 82(3), 649–659. https://doi.org/10.1890/0012-9658.2001(082)0649:SACOB2.0.CO;2

Hawkins, J., & Hughes, N. (2008). Cooperative roadside waterfowl breeding population surveys in the southern Yukon Territory 2008 report. Canadian Wildlife Service.

Hefley, T. J., Broms, K. M., Brost, B. M., Buderman, F. E., Kay, S. L., Scharf, H. R., ... Hooten, M. B. (2017). The basis function approach for modeling autocorrelation in ecological data. *Ecology*, 98(3), 632–646. https://doi.org/10.1002/ecy.1674

Hefley, T. J., & Hooten, M. B. (2016). Hierarchical species distribution models. *Current Landscape Ecology Reports*, 1(2), 87–97. https://doi.org/10.1007/s40823-016-0008-7

Held, L., Schrödle, B., & Rue, H. (2010). Posterior and cross-validatory predictive checks: A comparison of MCMC and INLA. T. Kneib & G. Tutz In *Statistical modelling and regression structures*, (pp. 91–110). Heidelberg: Physica-Verlag HD.

Hofner, B., Boccuto, L., & Göker, M. (2015). Controlling false discoveries in high-dimensional situations: Boosting with stability selection. *BMCO Biometrics*, 16(1), 144. https://doi.org/10.1186/s12859-015-0575-3

Holloway, P., & Miller, J. A. (2015). Exploring spatial scale, autocorrelation and nonstationarity of bird species richness patterns. *ISPRS International Journal of Geo-Information*, 4(2), 783–798. https://doi.org/10.3390/ijgi4020087

Hothon, T., Bühlmann, P., Kneib, T., Schmid, M., & Hofner, B. (2010). Model-based boosting 2.0. *Journal of Machine Learning Research*, 11, 2109–2113.

Hothon, T., Bühlmann, P., Kneib, T., Schmid, M., & Hofner, B. (2013). mboost: Model-based boosting (pp. 1–54). Retrieved from http://r-forge.r-project.org/projects/mboost/

Hu, Y. I., Ward, M. P., Xia, C., Li, R., Sun, L., Lynn, H., ... Jiang, Q. (2016). Monitoring schistosomiasis risk in East China over space and time using a Bayesian hierarchical modeling approach. *Scientific Reports*, 6, 24173. https://doi.org/10.1038/srep24173
Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Linde, A. V. D. (2002). Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64(4), 583–639. https://doi.org/10.1111/1467-9868.00353

Statistics Canada (2012). National Road Network feature catalogue segmented view. Ottawa, ON: Statistics Canada, Statistical Geomatics Centre.

Steen, V., Skagen, S. K., & Noon, B. R. (2014). Vulnerability of breeding waterbirds to climate change in the Prairie Pothole Region, USA. PLoS One, 9(6), e96747. https://doi.org/10.1371/journal.pone.0096747

Stralberg, D., Matsuoka, S. M., Handel, C. M., Schmiegelow, F. K., Hamann, A., & Bayne, E. M. (2017). Biogeography of boreal passerine range dynamics in western North America: Past, present, and future. Ecography, 40(9), 1050–1066. https://doi.org/10.1111/ecog.02393

Tucker, C. J., Pinzon, J. E., Brown, M. E., Slayback, D. A., Pak, E. W., Mahoney, R., ... El Saleous, N. (2005). An extended AVHRR 8-km NDVI dataset compatible with MODIS and SPOT vegetation NDVI data. International Journal of Remote Sensing, 26(20), 4485–4498. https://doi.org/10.1080/01431160500168686

Wikle, C. K., & Hooten, M. B. (2010). A general science-based framework for dynamical spatio-temporal models. Test, 19(3), 417–451. https://doi.org/10.1007/s11749-010-0209-z

BIOSKETCH

Antoine Adde is interested in the statistical modelling of species distribution and abundance with an emphasis on ecological niche modelling. This paper is part of his PhD dissertation at Laval University, Québec, Canada.

Author contributions: S.C., M.D., N.B. and A.A. conceived the study; A.A. planned and conducted the analyses; A.A. wrote the original manuscript; all authors made substantial contributions to the interpretation of results and the editing of the manuscript.

How to cite this article: Adde A, Darveau M, Barker N, Cumming S. Predicting spatiotemporal abundance of breeding waterfowl across Canada: A Bayesian hierarchical modelling approach. Divers Distr. 2020;26:1248–1263. https://doi.org/10.1111/ddi.13129

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.