Received: 1 December 2022 | Revised: 26 October 2023 | Accepted: 13 November 2023

DOI: 10.1111/geb.13792

Global Ecology and Biogeography | A Journal of Macroecology

WILEY

RESEARCH ARTICLE

Integrated species distribution models to account for sampling biases and improve range-wide occurrence predictions

Jussi Mäkinen[1,2,3] | Cory Merow[4] | Walter Jetz[2,3]

[1]Department of Ecology and Evolutionary Biology, Yale University, New Haven, Connecticut, USA

[2]Center for Biodiversity and Global Change, Yale University, New Haven, Connecticut, USA

[3]Research Center for Ecological Change, Faculty of Biological and Environmental Sciences, University of Helsinki, Helsinki, Finland

[4]Eversource Energy Center, University of Connecticut, Storrs, Connecticut, USA

Correspondence
Jussi Mäkinen, Research Center for Ecological Change, Faculty of Biological and Environmental Sciences, University of Helsinki, Helsinki, Finland.
Email: jussi.makinen@helsinki.fi

Funding information
National Aeronautics and Space Administration, Grant/Award Number: 80NSSC17K0282, 80NSSC18K0435 and 80NSSC22K0883; E.O. Wilson Biodiversity Foundation; National Science Foundation, Grant/Award Number: 2225078

Handling Editor: Antoine Guisan

Abstract

Aim: Species distribution models (SDMs) that integrate presence-only and presence–absence data offer a promising avenue to improve information on species' geographic distributions. The use of such 'integrated SDMs' on a species range-wide extent has been constrained by the often limited presence–absence data and by the heterogeneous sampling of the presence-only data. Here, we evaluate integrated SDMs for studying species ranges with a novel expert range map-based evaluation. We build new understanding about how integrated SDMs address issues of estimation accuracy and data deficiency and thereby offer advantages over traditional SDMs.

Location: South and Central America.

Time Period: 1979–2017.

Major Taxa Studied: Hummingbirds.

Methods: We build integrated SDMs by linking two observation models – one for each data type – to the same underlying spatial process. We validate SDMs with two schemes: (i) cross-validation with presence–absence data and (ii) comparison with respect to the species' whole range as defined with IUCN range maps. We also compare models relative to the estimated response curves and compute the association between the benefit of the data integration and the number of presence records in each data set.

Results: The integrated SDM accounting for the spatially varying sampling intensity of the presence-only data was one of the top performing models in both model validation schemes. Presence-only data alleviated overly large niche estimates, and data integration was beneficial compared to modelling solely presence-only data for species which had few presence points when predicting the species' whole range. On the community level, integrated models improved the species richness prediction.

Main Conclusions: Integrated SDMs combining presence-only and presence–absence data are successfully able to borrow strengths from both data types and offer improved predictions of species' ranges. Integrated SDMs can potentially alleviate the impacts of taxonomically and geographically uneven sampling and to leverage the detailed sampling information in presence–absence data.

KEYWORDS
citizen science, data integration, INLA, range prediction, sampling bias, spatial latent effect, Trochilidae

1 | INTRODUCTION

Information about species distributions is widely used for assessing species vulnerability to climate and land use change (Dawson et al., 2011; Jetz et al., 2007), and for optimizing species conservation efforts (Hannah et al., 2020; Jetz et al., 2022; Jung et al., 2021). Species distribution models (SDMs) provide such information through spatial predictions, but the accuracy and precision of the predictions depends on the quantity and quality of the underlying species occurrence data (Guillera-Arroita et al., 2015). Data to support SDMs from sources such as citizen science and new sensing technologies are growing rapidly (Amano et al., 2016; Wüest et al., 2019), but due to sampling biases these data alone are limited in how they can advance global species distribution knowledge (Oliver et al., 2021). Alleviating this deficiency and supporting improved spatial predictions has been key goal of more recent SDM approaches that integrate multiple types of species occurrence data (Dorazio, 2014; Fletcher et al., 2015). Such integrated SDMs have the potential to improve predictions of species' whole ranges, which are central to many basic and applied uses and underpin the species distribution 'Essential Biodiversity Variable' (Guisan et al., 2013; Jetz et al., 2019). If scalable to more species and larger geographic extents, there is thus a real potential for integrated SDMs to substantially improve the spatial biodiversity information base for under-sampled taxa and regions (Isaac et al., 2020). Hence, successful applications and validations of species range-wide integrated SDMs are urgently needed.

Data integration can be conducted in multiple ways (Fletcher et al., 2019), and in this study we focus on the joint likelihood approach (henceforth 'integrated models'), in which data sets are assigned separate likelihood models with shared parameters (Dorazio, 2014; Pacifici et al., 2017). Integrated SDMs of this sort have traditionally been fitted with systematically sampled presence–absence (PA) along with opportunistically sampled presence-only (PO) data (Fletcher et al., 2019) or coarser resolution PA data from species atlases (Chevalier et al., 2021, 2022). These models have been shown to improve all essential aspects of SDMs compared to single data-type models: estimation accuracy of model parameters (Chevalier et al., 2021; Koshkina et al., 2017), predictive accuracy on a hold-out data (Koshkina et al., 2017; Simmonds et al., 2020), model fit on the training data (Bowler et al., 2019; Martino et al., 2021) and model uncertainty (Farr et al., 2021; Rufener et al., 2021). So far, integrated SDMs have been shown to improve model fit regardless of the level of data deficiency (Simmonds et al., 2020), but associating gain in model performance from data integration with number of presence or sampling points has not been tested with real species. Traditionally, different integration methods, model structures and strategies to account for sampling biases have been validated at a spatially restricted set of testing points (Fletcher et al., 2019; Koshkina et al., 2017; Simmonds et al., 2020) or by a measure of information content (Martino et al., 2021). To our knowledge, no existing studies predict a species entire range with integrated SDMs. This is due to lack of data to evaluate SDMs on a spatial extent of a species' whole range and in a manner that does not propagate sampling biases in the validation (Isaac et al., 2020). We present a new validation scheme, in which we use expert range maps to validate models with respect to their skill to predict species' whole range correctly and relate predictive performance of models on two spatial extents: the spatially constrained extent where there are systematically monitored species data (traditional validation method) and the extent of the species' whole range (new validation method). To validate integrated SDMs for modelling data-deficient species, we study how the benefits of data integration are associated with the number of species presence points and species range size in both model validation schemes. In fact, the predictive validation of the species' whole range is missing for traditional single data-type SDMs as well and with this study we bridge a method gap in the whole field of species distribution studies and not just integrated SDMs.

The improvements for distribution prediction from data integration originates from better coverage of different environmental conditions (Chevalier et al., 2021, 2022) and the study area (Doser et al., 2022) by the combination of data sets compared with a single data set alone. This alleviates 'niche truncation', which describes inaccurately estimated covariate effects and overly narrow niche estimates due to poor environmental coverage of the species data (Chevalier et al., 2021; Robinson et al., 2020). Another potential benefit of integrated SDMs is in addressing the often strong sampling bias in opportunistically sampled PO data (Fithian et al., 2015) which may confound the estimates of the covariate effects and thus degrade the accuracy of the distribution predictions (Warton et al., 2013). However, several examples to date suggest limited success for integrated models to overcome strong sampling bias in PO data compared to models that use only PA data (Fletcher et al., 2019; Simmonds et al., 2020). Accounting for sampling biases requires additional spatial information about the sampling intensity or an explicit model parameterization, such as spatial latent effects, reflecting the sampling process (Dorazio, 2014; Koshkina et al., 2017). The potential of spatial latent effects in cases where there is no spatially explicit covariate data of sampling intensity has been demonstrated previously (Ahmad Suhaimi et al., 2021; Gelfand & Shirota, 2019; Renner et al., 2019; Simmonds et al., 2020), but the approach has as yet not been validated for the full extent of species geographical ranges (Isaac et al., 2020). Indeed, previous work suggested that spatial latent effects might create inaccurate and counter-intuitive covariate coefficient estimates due to inherent collinearity with

relevant covariates (Hawkins et al., 2007; Kim, 2021; Mäkinen et al., 2022). In earlier integrated SDMs which have focused on relatively narrow spatial extents, the risk of such spatial confounding was relatively small (Gelfand & Shirota, 2019; Simmonds et al., 2020), but a recent study showed that spatial confounding impacted the covariate effect estimates already on a country wide extent (Renner et al., 2019). We hypothesize that improving range-wide distribution predictions through integrated SDMs is conditional on alleviating the 'niche truncation' while rigorously treating sampling biases of each data set (Zulian et al., 2021) and avoiding spatial confounding.

Here, we create novel understanding of how integrated SDMs improve information about species distributions. We compare different strategies to account for sampling biases and spatial confounding in terms of models' predictive accuracy and covariate effect estimates. Our approach used PA data for detailed information about species–environment relationships for a limited set of sampling locations and PO data to deliver environmental signals across the species entire range. We modelled distributions of 71 hummingbird species in South and Central America and conducted a model comparison on two levels. First, we chose 25 species, which had high enough prevalence in the PA data set for meaningful spatial block-wise cross-validation and expert range map-based validation (Ridgely et al., 2003). We computed the correlation between the change in model performance when integrating data sets and number of species presence observations. Second, we predicted species richness with all 71 species and evaluated the models relative to the species richness map derived by stacking expert range maps of the 71 species. We assumed that systematic inaccuracies in single species SDMs from varying sampling intensity or spatial confounding propagate to community level predictions through stacked species distribution predictions (Pineda & Lobo, 2012). Given that species richness predictions are a common output from SDM studies (Schmitt et al., 2017), we considered this as an important output feature. Expert range maps can support an evaluation of model's predictions across the whole geographic range, albeit only at a coarser resolution, here assumed to be 2500 km^2 (Hurlbert & Jetz, 2007). To alleviate potential sampling biases of PO data or spatial confounding of a spatial latent effect, we formulated two additional integrated SDMs: one with an extra spatial latent effect to capture the varying sampling intensity in PO data set, as done previously (Gelfand & Shirota, 2019; Simmonds et al., 2020), and one with a spatially restricted latent effect (Hanks et al., 2015), which ensures that a spatial latent effect explains only residual spatial variation independent of the covariates and thereby avoids collinearity between covariates and a spatial latent effect. In total, we implemented five models (Table 1) to address the following questions:

1. Does an integrated SDM perform better compared to single data-type SDMs in terms of predictive accuracy in the spatially restricted extent and across the species' whole range?
2. Do integrated SDMs benefit from accounting for sampling bias of PO data or from avoiding spatial confounding?
3. What is the role of covariate effect estimates in the differences of model performance?
4. Are benefits of data integration associated with the levels of data deficiency measured with the number of presence observations and range size?

2 | MATERIALS AND METHODS

2.1 | Data sets

We obtained PO, PA and expert range map data for 71 hummingbird species from a previous data integration study (Ellis-Soto et al., 2021). Their data were accessible through the Map of Life (mol.org): PO observations are originally from GBIF (https://gbif.org), PA observations are a collection of long-term checklists in the northern Andes (https://mol.org/datasets/769f3b99-214e-4056-8c39-1200a6855943) and expert range maps were originally created by Ridgely et al. (2003) as then further amended in Jetz et al. (2012) (https://mol.org/datasets/d542e050-2ae5-457e-8476-027741538965). PO observations were restricted to 1979–2017 to match the climatic covariates used as explanatory variables. We thinned PO data set

TABLE 1 Models assessed in this study and their likelihoods and intensity functions.

Model	Definition	Likelihood	Intensity function
Presence-only	PO	$Y_{PO} \sim \text{Poisson}(\lambda_{PO})$	$\log(\lambda_{PO}) = \alpha + \beta x + g(s)$
Presence–absence	PA	$Y_{PA} \sim \text{Bernoulli}(f(\lambda_{PA}))$	$\log(\lambda_{PA}) = \alpha + \beta x + g(s)$
Integrated	PO + PA	$Y_{PO} \sim \text{Poisson}(\lambda_{PO})$ $Y_{PA} \sim \text{Bernoulli}(f(\lambda_{PA}))$	$\log(\lambda_{PO}) = \alpha_{PO} + \beta x + g(s)$ $\log(\lambda_{PA}) = \alpha_{PA} + \beta x + g(s)$
Integrated + additional sampling effect	PO + PA + samp.	$Y_{PO} \sim \text{Poisson}(\lambda_{PO})$ $Y_{PA} \sim \text{Bernoulli}(f(\lambda_{PA}))$	$\log(\lambda_{PO}) = \alpha_{PO} + \beta x + g(s) + h(s)$ $\log(\lambda_{PA}) = \alpha_{PA} + \beta x + g(s)$
Integrated + restricted spatial latent effect	PO + PA + restr.	$Y_{PO} \sim \text{Poisson}(\lambda_{PO})$ $Y_{PA} \sim \text{Bernoulli}(f(\lambda_{PA}))$	$\log(\lambda_{PO}) = \alpha_{PO} + \beta x + g(s)$ $\log(\lambda_{PA}) = \alpha_{PA} + \beta x + g(s)$ $\text{cor}(x, g(s)) = 0$

Note: PO and PA models were fitted with a single data type, and the integrated model PO + PA and its modifications were fitted with both data types. PO + PA + samp. introduces an additional spatial latent effect (h(s)) to explain the effect of varying sampling intensity. PO + PA + restr. restricts the spatial latent effect, g(s), by orthogonalizing it to the environmental covariates.

of each species by discarding duplicated observations having the exact same location. By removing duplicates, we changed the variable of interest from relative abundance to relative areal density of locations where species is present (also called relative occurrence rate). Sampling strategies of PO data may vary between surveys collected in GBIF (including eBbird) and the spatial distribution of PO points reflects the collective sampling intensity of the surveys. See the Tables S1.1–S1.3 in Appendix 1 for example species data and full lists of species and sample sizes, and Tables S1.4 and S1.5 in Appendix 1 for summary statistics of species sample and range sizes. PA observations have been sampled from areal units of size roughly $1\,km^2$. This is a standard survey unit area for avian monitoring. Still, some of the survey areas were not reported, but we think that the analysis is not sensitive to this missing information. Basically, if the size of the survey unit areas are independent of the environmental covariates and species intensity (which we can safely assume), the deviation of the average survey unit area from $1\,km^2$ would only affect the estimate of the model intercept of the presence–absence part of the integrated model and this parameter is not of ecological interest. Based on previous studies, $1\,km^2$ captures adequately the environmental niche of the species (Lu & Jetz, 2023).

For many study species, PO data cover a larger area than PA data and better capture the extent of the species range [see Figure S1.1 in Appendix 1 for example species, buff-winged starfrontlet (*Coeligena lutetiae*) and sword-billed hummingbird (*Ensifera ensifera*)]. Data sets also differ in the geometry of their sampling: PO data are point-wise observations without an areal measure of sampling, whereas PA observations are measures of species presence status in a fixed survey area (Gelfand & Shirota, 2019; Moraga et al., 2017). As such they are not comparable measures of either species presence probability or abundance, but they can be related to the same function of species intensity through different link functions (Section 2.2). PA locations do not follow a uniform or random sampling design, but we assume that the impact of preferential sampling is negligible compared to that in the PO data and does not impact the inference and the prediction accuracy.

Environmental conditions of our $1\,km^2$ analysis pixels were characterized by annual mean temperature, mean diurnal range, annual precipitation, precipitation seasonality, intra-annual variation of cloud coverage (Wilson & Jetz, 2016), enhanced vegetation index (EVI) (Didan, 2015) and topographic ruggedness index (TRI) (Amatulli et al., 2018). Climatic variables were derived from Chelsa bioclim data set (Karger et al., 2017a, 2017b). Models were built at $1\,km^2$ resolution following the grid system of Chelsa bioclim layers. Species and environmental data are accessible through Mäkinen et al. (2023).

2.2 | Integrated models

An overview of the five models assessed and their likelihood and intensity functions is given in Table 1. We defined the observation model for the PO data as an inhomogenous Poisson point process (IPPP) conditional on the intensity of the process (Renner et al., 2015). Log likelihood of a Poisson point process is defined as:

$$\log\left(l\left(\lambda;s_{PO}\right)\right)=\sum_{i=1}^{I}\log\left(\lambda\left(s_i\right)\right)-\int_{A}\lambda(s)ds, \quad (1)$$

where λ is the intensity of presence observations, s_i a spatial location of the ith presence observation, I the total number of the PO observations and A the study area. However, the integral over the study area cannot be computed and needs to be approximated, for example by discretizing the study area into quadrature points and assigning them weights (Warton & Shepherd, 2010). We sampled 20,000 quadrature points in a systematic grid which has been shown to be sufficient for convergence of the integral estimate in an IPPP (Renner et al., 2015) and assigned them the same quadrature weights, which equalled to the size of the grid cells. We did not apply any other sampling scheme or number of quadrature points, but relied on earlier studies showing the sufficiency of 20,000 points (Renner et al., 2015; Valavi et al., 2022). This saved us from high computational burden as well. Model convergence does not directly depend on the study extent but on the spatial autocorrelation of the environmental covariates. In our study, environmental variables vary smoothly over the study area and 20,000 quadrature points capture the environmental variation well enough to guarantee convergence.

The likelihood of the Bernoulli distributed PA observations can be linked to the species' intensity, λ, by computing the probability of a species presence with a complementary log–log link function of the intensity (Dorazio, 2014). The likelihood of the PA observations is defined as:

$$\log\left(l\left(\lambda;s_{PA}\right)\right)=\sum_{j=1}^{J}-y_j\log\left(1-\exp\left(-\lambda\left(s_j\right)\right)\right)+\left(1-y_j\right)\lambda\left(s_j\right), \quad (2)$$

where J is the total number of the PA observations and y_j is the jth PA observation. The joint log likelihood is computed as the sum of the PO and PA log likelihoods. The intensity of the presence points corresponds to the expected number of observations per $1\,km^2$ and as it matches with the sampling unit size of the PA data, we do not need to scale the intensity variable when computing the log likelihood for the PA data.

The logarithm of the intensity function was modelled as a function of first- and second-order polynomial effects of the environmental covariates (to allow modal responses) and the spatial latent effects. In Table 1, we present the single data-type and integrated models. The function of the log intensity is:

$$\log\left(\lambda\left(s_i\right)\right)=\alpha+\sum_{k=1}^{14}\beta_k x_k\left(s_i\right)+g\left(s_i\right), \quad (3)$$

where α is the average log intensity of the species over the study area, β_k is the coefficient related to the kth variable, $x_k\left(s_i\right)$ is the value of that variable in location s_i and $g\left(s_i\right)$ is the spatial latent effect shared between the Bernoulli and Poisson likelihoods (more information

about spatial latent effects is provided in section for model fitting 2.3). The likelihood functions for the two data sets have separate intercept parameters to account for the differing average species intensities (α_{PA}, α_{PO} for PO + PA and its varieties in the Table 1), partly arising from differing average sampling efforts in the data sets. Further details on the IPPP and log-intensity function have been discussed, e.g., in Matthiopoulos et al. (2020).

In two other integrated models we modified this intensity function: In the PO + PA + samp. model we added another spatial latent effect ($h(s_i)$), which was fitted only for the PO data and assigned to its likelihood. $h(s_i)$ was intended to capture the spatially varying sampling intensity in the PO data. A core assumption when fitting an IPPP for the PO data is that the presence locations are collected by randomly sampling the whole study area with a constant effort. However, we could relax this assumption in the PO + PA + samp. model (see Table 1) with $h(s_i)$ and by breaking the point process to two processes: one reflecting the species distribution ($\alpha + \sum_{k=1}^{14} \beta_k x_k(s_i) + g(s_i)$) and one reflecting the sampling effort ($h(s_i)$) (Diggle et al., 2010). This model is analogous to a marked point process model, where the point process refers to the sampling effort and the marks, here presence locations refer to the species distribution (Soriano-Redondo et al., 2019). The additional spatial latent effect $h(s)$ is not interesting for the species distribution and omitted when making spatial predictions of the species distribution. The same model structure and reasoning was used by Gelfand and Shirota (2019) and Simmonds et al. (2020). It is possible that $h(s_i)$ captures spatial variation from other sources as well, such as long-distance dispersal, but we assume that given $g(s_i)$ capturing spatial variation over the PO and PA data sets, $h(s_i)$ explains mostly the effect of sampling intensity. We validated this assumption implicitly in the model evaluation by leaving $h(s_i)$ out from the prediction of the species distribution. In case that $h(s_i)$ would capture ecologically essential variation of a species distribution, model's predictive accuracy should suffer compared to other integrated models.

In the PO + PA + restr. model, we constrained the variation of $g(s_i)$ so that it did not correlate with the environmental covariates and thus captured such spatial variation, which was left over after we accounted for the covariate effects. Such restricted spatial regression methods have been introduced previously (Hanks et al., 2015; Reich et al., 2006) to avoid spatial confounding, namely high collinearity, between environmental covariates and a spatial latent effect. Such confounding distorts the inference and returns counter-intuitive estimates for the covariate effects but can be avoided with strict restrictions on the spatial latent effect (Hodges & Reich, 2010). A similar approach for single data-type SDMs has been earlier supported by Crase et al. (2012), who showed that a restricted spatial latent effect simultaneously returned accurate estimates for the covariate effects and captured the spatial autocorrelation of the response variable. However, if the estimates of the non-spatial model for the covariate effects are inaccurate due to collinearity between the covariate and the unobserved spatial process, a restricted spatial latent effect cannot improve the accuracy of the estimates, though it increases posterior variance of the covariate effects (Hanks et al., 2015; Hodges & Reich, 2010). In our study, the unobserved spatial process corresponded to e.g. sampling intensity, which could correlate with environmental covariates by chance or due to a common driving factor, such as a macroclimatic variable. Thus, spatial confounding could originate from the collinearity between the observed covariates and the unobserved processes. To our understanding, this model structure has not been earlier applied in integrated SDMs, and it provides an interesting opportunity to avoid confounding between the covariates and the spatial latent effects, though it assumes that sampling bias in PO data is independent of the environmental gradients and species population density which usually is not the case. We constrained the spatial latent effect by orthogonalizing it to the intercept and the environmental covariates, following the approach described previously (Hanks et al., 2015; Reich et al., 2006). See Appendix 2 for a more detailed definition of the constraint. By comparing different versions of the PO + PA models, we tested whether sampling bias (treated explicitly with the PO + PA + samp. model) or spatial confounding between model terms (avoided explicitly with the PO + PA + restr. model) has a greater impact on the predictive performance of the model.

2.3 | Model fitting

We fitted models for each species separately with integrated nested laplace integration (INLA) through the R-INLA package and applied stochastic partial differential equations (SPDE) for fitting the spatial latent effects (Bakka et al., 2018; Lindgren & Rue, 2015; Rue et al., 2017). We assigned Matérn-3/2 type covariance function for both spatial latent effects ($g(s), h(s)$) and assigned penalized complexity priors for the parameters of the covariance functions (Fuglstad et al., 2018) (magnitude of variation and spatial autocorrelation range of covariance). We put more prior probability for locally varying (with probability 0.95 that spatial autocorrelation range is less than 200 km) and relatively weak (with probability 0.99 that magnitude of variation is less than 2) spatial latent effects compared to the spatial structure and effect of the environmental covariates. Moreover, we assigned slightly informative priors for the model intercepts and covariate effects (Gaussian distribution with mean zero and standard deviation 10). Spatial latent effects ($g(s), h(s)$) were fitted along with the covariate effects, but only $g(s)$ was used for predicting as well.

2.4 | Model validation

We validated models with fourfold block-wise cross-validation on the PA data set, and the PO data set was used only for model training. In cross-validation, folds were formed by splitting the PA sites into 20 spatially distinct blocks and grouping blocks into four folds, which provide the most even species prevalence among the

folds. The presence points of the PO data set that were located less than 30 km from the testing points were discarded to improve the independence of the testing data set. We kept the quadrature points in their original locations since we wanted to keep the area where validation points are located comparable to any part of the study area where a species may occur, but no species records exist, as when predicting species whole ranges. Discarding the PO points decreases the effect of spatial latent effect in the prediction and impacts the predicted values of the model around the spatial block of the validation points compared to a full model. This reflects a scenario where a model is used for predicting the species distribution in an area, where there are not any kind of observational data. This allows us to validate models better with respect to how accurately covariate effects were estimated (Roberts et al., 2017; Wenger & Olden, 2012). We validated models with AUC and Cohen's kappa. For kappa, we set a threshold for transforming the predicted intensities into binary presence–absence prediction by choosing the log-intensity value that provided the highest sum of specificity and sensitivity in the training PA locations. This was done for the PO model as well, though the training PA data set was not used in model fitting. In the model validation, we followed a Bayesian approach by taking 500 samples from the predictive posterior distribution and computing the mean predictive accuracy over the samples. We further computed the standard deviation of the predictions in the cross-validation to test for predictive precision.

We restricted the model validation to 25 species, which had at least three presence observations in each testing fold creating on average a prevalence of 0.145 for testing folds. Prevalence is the proportion of sites with a presence observation to all sampling sites. We conducted cross-validation and species' whole-range validation for this set of species to keep the metrics comparable.

We conducted an additional, complementary assessment of range-wide predictions with the expert range maps. To do this, we first predicted the species log intensity throughout the study domain [the larger of a 750-km buffer around the (independent) expert range maps ranges or the PO locations]. On the extent of 1000 s of kilometres, expert range maps for groups like birds can offer a sufficiently reliable source of presence and absences at spatial grains of ca. 100 km upward, with 50 km justifiable for well-studied systems (Graham & Hijmans, 2006; Hurlbert & Jetz, 2007). Moreover, expert range maps provide a reliable estimate of species absences (high specificity) regardless of the spatial resolution and thus inform about the extent of occurrence (Graham & Hijmans, 2006; Hurlbert & Jetz, 2007). Most importantly, we can assume that the errors in expert range maps are randomly distributed across the species ranges and independent of the SDMs' error sources, such as the spatially varying sampling intensity of the PO data. We transformed the expert range maps into presence–absence data in 50 km resolution (2500 km^2) by treating each coarse resolution cell as presence if any 1 km^2 cell was present inside it. We set the threshold for the range-wide prediction in the same way as in the cross-validation and used AUC and kappa as the validation metrics. Additionally, we conducted an expert range map-based validation for all 71 species to make sure that the results are not sensitive to the sub-sample of species. We conducted another expert range map-based validation by computing the sum of species intensities over the study domain and calculated the proportion of the intensities inside the expert range map. This validates models with respect to how much they over predict species intensities outside of the expected range so that the best-performing model would predict 100% of the population inside the expected range. An evaluation of integrated models was previously applied by Schank et al. (2017) for predicted population size, but we are not aware of earlier studies using expert range maps for validating predictions of a species range or population distribution.

We compared the models in terms of covariate effect estimates which denote the breadth of conditions that a species can tolerate. We assumed that the PA data covers a small part of the full range for a few species, and hence not a complete sample of environmental gradients, which are accessible to those species. Hence, the PA model could suffer from 'niche truncation' and provide overly narrow niche estimates (Chevalier et al., 2021). Integrating PO data from across the species whole range allows to sample the whole environmental gradients and could alleviate 'niche truncation'. Without the true data generating values, we compared models only in terms of relative 'niche truncation' or 'inflation' (overly large niche estimates) among the models. We assessed the role of niche estimates in species range predictions by comparing models in terms of covariate effect estimates and range prediction accuracy. Moreover, we tested whether the changes in covariate effect estimates are related to the spatial autocorrelation range of the covariates. We assume that accurate effect estimates for a relatively coarse scale covariate (long spatial autocorrelation range) would require the observational data to fully cover the study area to sample whole environmental gradients. Here, we computed the strength of the spatial autocorrelation of the covariates with Moran's index using the Moran function from raster package in R (v. 4.2.1). Moran's index measures the overall strength of correlation between neighbouring cells across the study area. Values close to one indicate strong spatial clustering and smooth spatial appearance of the covariate. To study the benefits of data integration relative to data deficiency we computed Pearson's correlation coefficients and statistical significance of the correlation between the change in predictive performance when comparing single data-type and integrated SDMs and the number of presence points in the PO and PA data sets and the species range size (from the expert range map).

We computed a species richness prediction over 71 species by stacking single species predictions and compared the species richness prediction to a species richness map from stacked expert range maps. We compared the predicted species richness with the number of the PO records on the cell level to study the dependence of the richness prediction on the sampling intensity. We did both comparisons with Pearson's correlation test. This was

conducted in the 50 km resolution to reflect the relatively coarse resolution implicit in the expert range maps.

3 | RESULTS

In cross-validation, the PO + PA + samp. model performed better than the PO model but slightly worse than the PA model (Figure 1: first row). For most species, the PA model performed the best. The other integrated models (PO + PA and PO + PA + restr.) provided only slight improvement in cross-validation compared to the PO model but were inferior to the PA and PO + PA + samp. models in terms of AUC and kappa metrics (Figure 1: first row).

When predicting species' whole range with all the data, the PO + PA + samp. model performed on average better than the PA model and as well as the PO and other integrated models (Figure 1: second row). This pattern was consistent across both validation metrics in the binary range-wide prediction (AUC and kappa) and the proportion

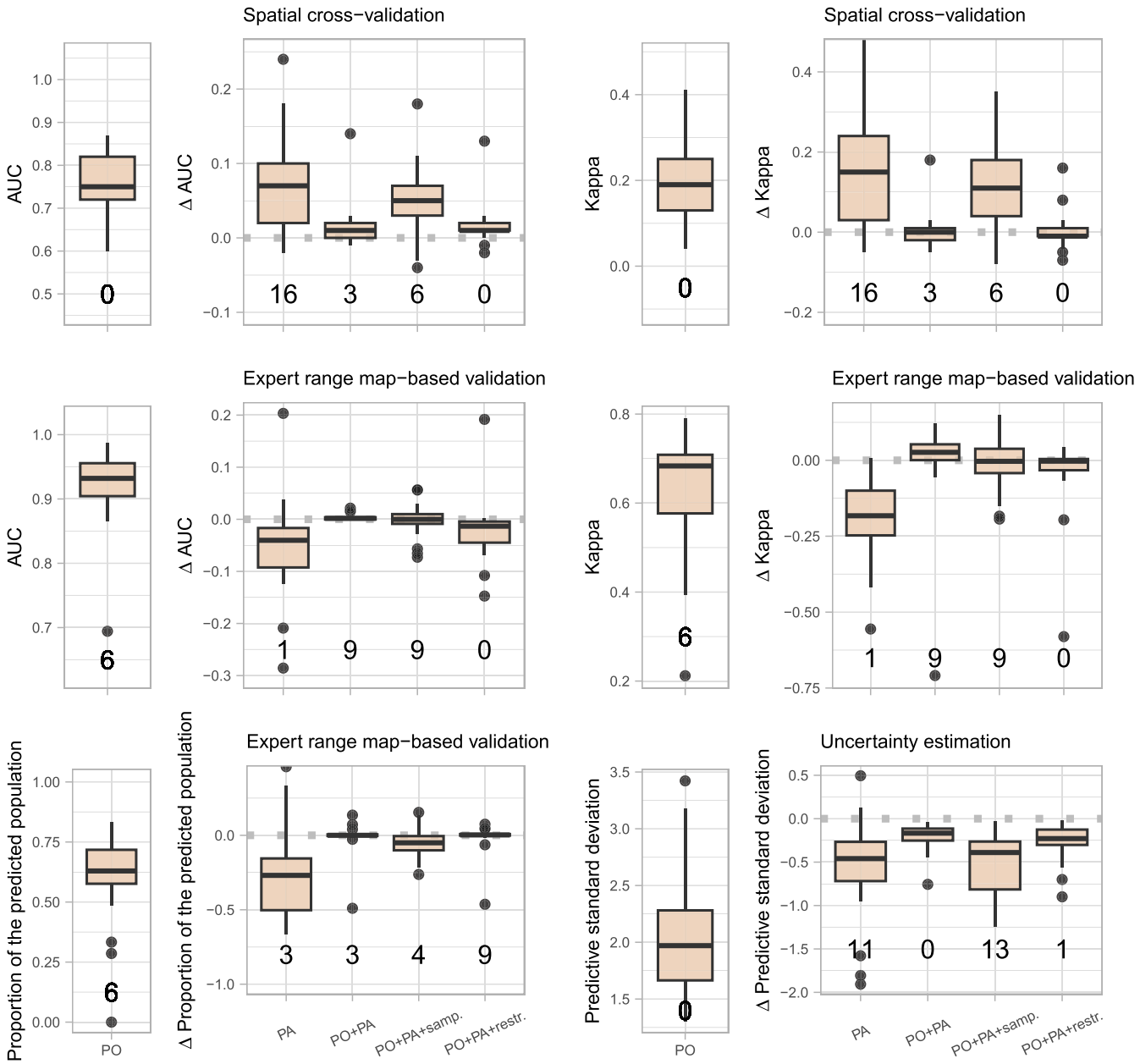

FIGURE 1 Model performance comparison for the 25 hummingbird species with sufficient data for spatial cross-validation. Panels in the first and third columns show the distribution of the absolute values of PO model and the second and fourth columns show distribution of the pairwise differences in performance to PO model. First row: cross-validation. Second row: range-wide prediction. Third row: range-wide prediction and predictive precision. In the range map-based validation, values on the *y*-axis denote the proportion of the predicted population inside the expert range map. In the uncertainty estimation, values on the *y*-axis denote the average predictive standard deviation over the validation points of all cross-validation folds. The number under each box denotes the number of species for which the respective model performed the best.

of the population predicted inside the expert range (Figure 1: second and third rows). Overall, there were only four species, for which a single model performed the best in cross-validation and expert range map-based validation in terms of AUC (three with the PO + PA + samp. model and one with the PO + PA model). The expert range map-based validation results for all 71 species were fully in line with the validation results for the 25 species examined with the cross-validation (see Figure S1.2 in Appendix 1). We also assessed the predictive precision in the cross-validation as a measure of predictive uncertainty. Here the PA model provided the highest precision followed by the PO + PA + samp. model (Figure 1: third row). See Figure S1.1 in Appendix 1 for an illustration of the model comparison for two example species, the buff-winged starfrontlet, a narrow-ranged species with relatively few PO observations, and the sword-billed hummingbird, a wide-ranged species with relatively many PO observations.

The PO and integrated models estimated on average stronger quadratic effects for temperature (on average −9) and precipitation (on average −4) than the PA model did (average quadratic effect −5 for temperature and −1 for precipitation) (Figure 2). The change in the covariate effect estimates was smaller for other covariates decreasing on average close to zero for EVI. Here we highlight estimates for a few covariates, but see Figures S1.5 and S1.6 in Appendix 1 for the same comparison for all covariates for the 25 cross-validated species and for all 71 species. There were only small differences among the PO and integrated models in terms of covariate effect estimates. The environmental covariates had similar levels of spatial autocorrelation

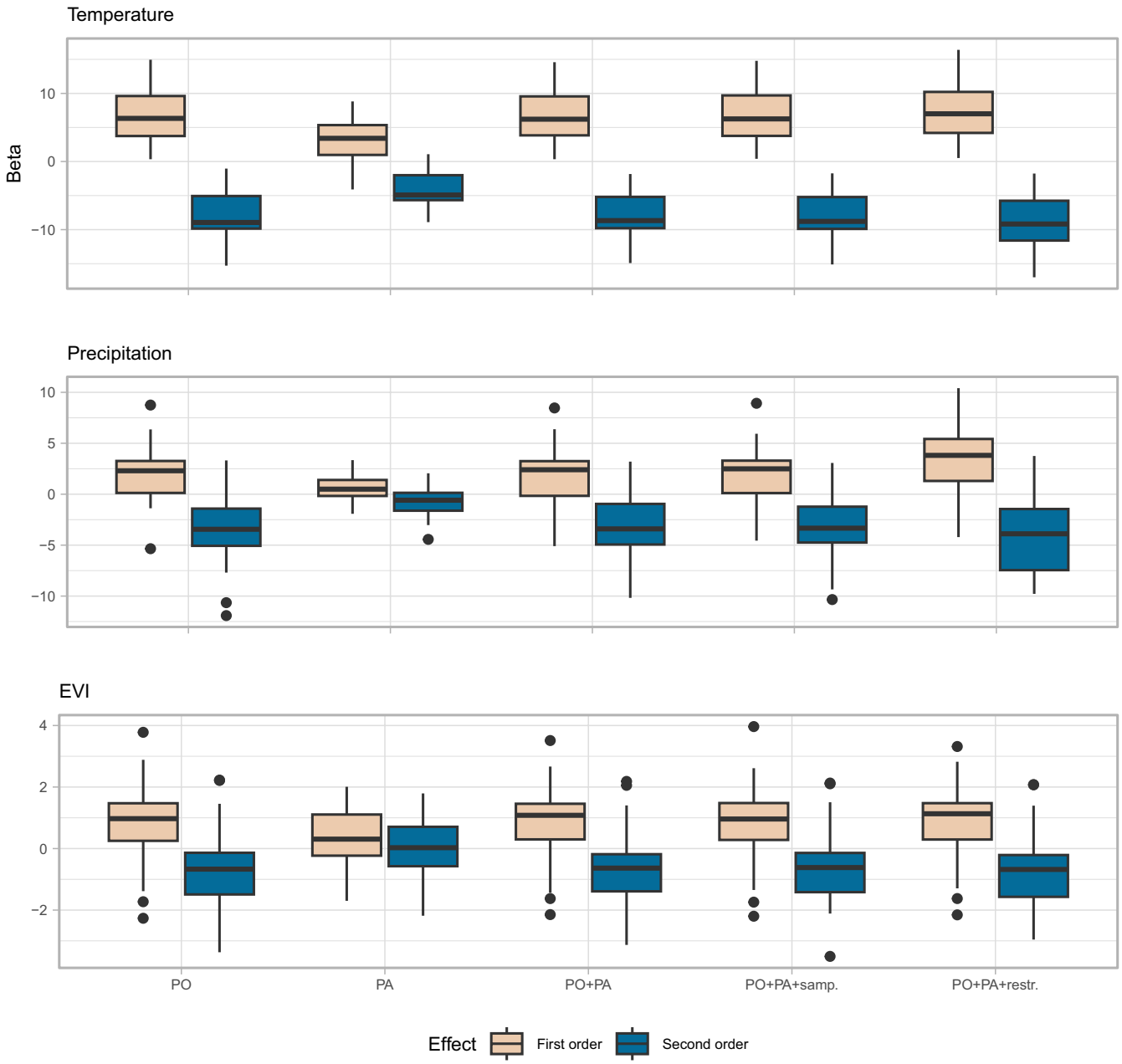

FIGURE 2 Distributions of the estimates for the effects of temperature, precipitation and EVI for the 25 hummingbirds.

measured with the Moran's *I*, and thus the changes in the covariate effect estimates were not associated with the spatial structure of the covariates. See Table S1.6 in Appendix 1 for a table of Moran's *I* of each covariate raster.

The change in the model performance in cross-validation between the PO and PO + PA + samp. models was not associated with number of presence points in the PO or PA data sets or species range size (Table 2). However, there was a negative association between the model improvement in expert range map-based validation between the PO and PO + PA + samp. models and the number of PO presence points (Table 2). We found the same association in cross-validation between the performance difference of the PA and PO + PA + samp. models. The same comparison over all 71 species showed only weak and statistically insignificant associations (see Table S1.7 in Appendix 1).

TABLE 2 Pearson's correlation coefficient and statistical significance (*$p < 0.05$) for how the performance difference between the PO + PA + samp. and the PO or PA models is associated with the number of species presences in PO and PA data sets and the species range size (estimated from the expert range map).

	Cross-validation		Expert range map	
	Model		Model	
	PO	PA	PO	PA
Presences (PO)	−0.35	−0.51*	−0.41*	0.04
Presences (PA)	−0.12	−0.16	−0.28	−0.09
Range size	0.14	0.1	0.13	−0.09

Note: Model performance is the AUC in cross-validation and expert range map-based validation for 25 test species.

TABLE 3 Correlation between species richness derived from expert range maps versus stacked SDM predictions and species richness, assessed at 50 × 50 km grain ($N = 11{,}291$).

	Correlation	
Model	Expert range map	Sampling intensity
PO	0.82	0.7
PA	0.85	0.62
PO + PA	0.83	0.71
PO + PA + samp.	0.87	0.66
PO + PA + restr.	0.84	0.69

Note: The correlation between expert range map species richness and sampling intensity is 0.56. Correlation is estimated with Pearson's correlation coefficient.

We next extended our predictions to the full set of 71 hummingbird species and compared the species richness predictions from stacked model's predictions with the species richness estimate from the stacked expert range maps. Results showed that prediction of the PO + PA + samp. model had the highest correlation with the estimated species richness from the expert range maps and the second lowest correlation with the sampling intensity after the PA model (Table 3). The PO model predicted high species richness in a few areas, which also correspond to the areas of high sampling intensity (Figure 3), whereas the PA model predicted high species richness across the northern Andean Mountain range. The prediction of the PO + PA + samp. model was less impacted by the sampling intensity than the other integrated models were, and the predicted hot spot areas matched better with those of the expert range map estimate.

4 | DISCUSSION

We tested integration of the opportunistically sampled PO data and the species checklist-based PA data for fitting integrated SDMs for 71 hummingbird species on the extent of the species' whole ranges. Of the different integration methods (tested, e.g. in Ahmad Suhaimi et al., 2021; Pacifici et al., 2017; Simmonds et al., 2020), we focused on the joint likelihood approach since it allowed us to analytically link different types of survey data sets to a function of environmental covariates and spatial latent effects. We found that the integrated SDM with the additional PO data-specific spatial latent effect (PO + PA + samp.) predicted the presence–absence status of the

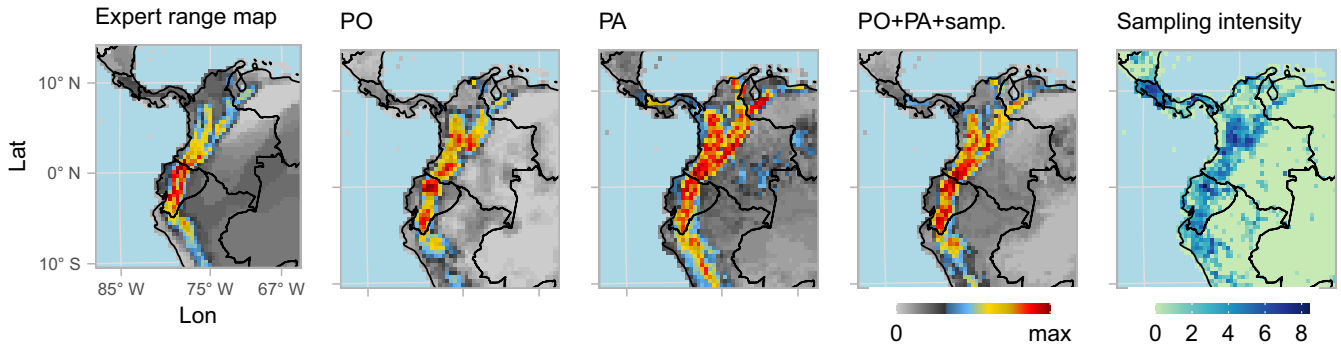

FIGURE 3 Predicted species richness of the three presented models in relation to expert range map richness and sampling intensity (number of samples are shown on log scale), all represented for grid cells of size 50 × 50 km. Model-based richness is based on the stacked thresholded (binary) distributions of the 71 assessed species. The rightmost panel shows the sampling intensity as the count of presence records of assessed species in each grid cell.

species in cross-validation more accurately and precisely than the PO or other integrated models did. Moreover, the PO + PA + samp. model improved the accuracy of the species' whole-range predictions compared to the PA model. In all model comparisons, the PO + PA + samp. model balanced between the PA, PO and other integrated models by providing relatively good predictions but not necessarily being the top performing model. Compared to the PO model, the PO + PA + samp. model improved the range prediction especially for data-deficient species. The PO and integrated models had stronger effects of the environmental covariates, especially temperature and precipitation, than the PA model had. The aggregated species richness predictions of the PO + PA + samp. model was less severely impacted by sampling bias of the PO data set than other integrated models were and based on a comparison to stacked expert range maps, the PO + PA + samp. model provided more accurate prediction for the areal species richness than the other models did. The sampling component of this model seems to succeed in capturing redundant spatial variation in the PO data set and address the effects of the varying sampling intensity.

4.1 | Model comparison

Our results show that integrated SDMs can help address limited amounts of the occurrence data but the improvements are conditional on accounting for different biases of the data sets, such as the sampling bias of the PO data, as can be expected based on previous studies (Dorazio, 2014; Renner et al., 2019). However, the method applied here using the additional spatial latent effect has not previously been validated on a spatial extent representing a species entire range. Previous studies have also found support for integrated models over the PA or PO models (Ahmad Suhaimi et al., 2021; Koshkina et al., 2017; Pacifici et al., 2017; Simmonds et al., 2020), and our results support those findings with the addition that the improvement of the integrated modelling depends on the prediction domain of the model: over the PA domain, the PA model performed better in cross-validation than the PO + PA + samp. model, but underperformed the PO + PA + samp. model when predicting species' whole range. Although three studies have tested the integrated SDMs in scenarios where the PA data covers only a part of the study area (Ahmad Suhaimi et al., 2021; Robinson et al., 2020), they have not validated the models with independent range estimates and thus our approach shows that modelling the PA data alone may give inaccurate predictions of species' whole ranges.

We can relate the differences in the performance of the PA, PO and PO + PA + samp. models to the estimates of the environmental covariate effects and of the hyperparameters controlling the spatial latent effects. First, although all models supported modal responses of the species population to temperature and precipitation, the PO and integrated models supported steeper decline of the intensity function around the optimal temperature conditions than what the PA model supported. Second, the PO model had the highest estimate for the standard deviation of the spatial latent effect followed by the PO + PA + samp. (estimate for the standard deviation of the shared spatial latent effect) and the PA model with the lowest standard deviation estimate (see Figure S1.3 in Appendix 1). We also suspect that, due to relatively small estimates (on average one third of the estimate for the PO model) for the standard deviation of the spatial latent effect, the PA model's predictions in the cross-validation were more informed by the environmental covariates compared with the PO model's predictions. On the other hand, the PO and PO + PA + samp. models predicted species' whole range on average more accurately than the PA model did due to the stronger effects of the temperature and precipitation and that their spatial latent effects reduced the predicted species intensity outside of the species range, where there were no presence observations in either of the PO or PA data sets. The PA model over-predicted ranges by predicting species distribution to areas, which may be unsuitable or inaccessible for the species or where the species may have been outcompeted. The PA model's predictions had a smaller range of the predicted values compared with the predictions of the PO + PA + samp. model. For example, for sword-billed hummingbird, the prediction of the PA model varied between −8 and 0 and the prediction of the PO + PA + samp. model varied between −15 and 0 (see Figure S1.1 in Appendix 1). This conclusion is supported by the finding that the PA model predicted on average larger ranges than what the expert range maps indicated and larger than what other models predicted (Figure S1.4 in Appendix 1). We can assume that the expert range maps are overestimates of the true ranges since they do not always capture holes in species ranges. Only the PO + PA + samp. model predicted on average slightly smaller ranges than what the expert range maps indicated. In the worst case, an overestimated range size may underestimate the species vulnerability status and exclude the species from conservation targets (Hurlbert & Jetz, 2007). From an ecological standpoint, small covariate effect estimates of the PA model characterize species more as generalists compared with the other models, which return tighter response curves and predict smaller environmentally suitable and geographically accessible space for the species. The results showed that the PO + PA + samp. model could leverage information about the species environmental niche and other range constrains from species' whole range and improve the range prediction compared with the PA model. Due to accounting for the additional spatial variation in the PO data, it could still improve cross-validation compared with the PO model. To our knowledge, integrated models have not been earlier validated with two contrasting spatial extents (cross-validation vs. species' whole range) with a link to the covariate effect estimates, and our results show the magnitude and driver of improvements of the integrated models compared with the PO and PA models.

Comparison of the different integrated models showed that the PO + PA + samp. model performed better than other model candidates. Large estimates for the standard deviation and small estimates for the spatial autocorrelation range of the PO data-specific spatial latent effect ($h(s)$) of the PO + PA + samp. model (see Figure S1.3 in Appendix 1) indicate that a high proportion of the variation of the PO data was likely related to the spatially

varying sampling intensity. This feature of the data generating process was not accounted for by the PO or other integrated models which penalized the performance of this group of models in the cross-validation. However, the estimates for the environmental covariates did not change much between the PO and integrated models, which indicates that the sampling bias impacted more the spatial latent effect than the environmental covariate effects. These findings address the importance of accounting for the sampling bias of the PO data explicitly in the model structure. Similar conclusions were previously reported (Gelfand & Shirota, 2019; Simmonds et al., 2020), and this matches well with the ideas presented (Renner et al., 2019) to find a shared spatial process between the data sets and simultaneously account for the data set-specific sampling processes. To our surprise, we did not detect strong spatial confounding between the environmental covariates and spatial latent effects and confounding seemed to take place in the spatial latent effect in the PO and PO + PA models so that predictions of these effects mostly reflected the spatially varying sampling intensity. This conclusion was supported by the smaller estimates for the standard deviation of the shared spatial latent effect of the PO + PA + samp. model compared to that of the PO + PA model. We did not test whether leaving a spatial latent effect out when predicting with the PO + PA or PO + PA + restr. models would have returned more accurate results. However, a problem inherent to the PO, PA, PO + PA or PO + PA + restr. models (and like to many other SDMs) is that there is no instrument to separate out data set-specific variation from the model that is used for predicting. A more extreme option would have been to leave spatial latent effects out also from the model fitting and explain variation in species' log intensity solely with environmental variables. However, such non-spatial model would have returned overly optimistic uncertainties to model estimates and not necessarily improved estimates' accuracy (Mäkinen et al., 2022). Moreover, the model structure chosen here corresponds to state-of-the-art SDMs, introduced e.g. in Chakraborty et al. (2011) and Latimer et al. (2009), applied e.g. in Mäkinen and Vanhatalo (2018) and Soriano-Redondo et al. (2019) and provides opportunities to explain different sorts of redundant spatial variation in the data compared with non-spatial models.

In addition to the higher predictive accuracy, the PO + PA + samp. model had consistently lower predictive uncertainty than the PO or other integrated models (Figure 1: Uncertainty estimation). Similar results have been reported also by Rufener et al. (2021) when fitting integrated SDMs for fisheries. Uncertainty is an important feature for the applicability of the model's predictions in decision-making problems, for example when assessing the effectiveness of management actions (Regan et al., 2005; Williams & Hooten, 2016). For most species the PO + PA + samp. model combined relatively high predictive accuracy on the extents of spatially constrained monitoring data and species' whole ranges with relatively low model uncertainty. Given that this may not necessarily be the case for other species presenting different data features, a multi-model inference framework would help choose the best model among a set of models. We recommend including the PO + PA + samp. model in future data integration studies and for wider and applied use creating information about biodiversity across spatial scales. Moreover, the improvements in species richness predictions show that integrated modelling can refine macroecological studies and provide new findings in the analysis of richness gradients.

4.2 | Studying data-deficient species and areas

The general assumption that data integration benefits data-deficient species, which are scarcely observed due to low population abundance or low detectability, was not fully supported by our study. Our results showed that the PA model performed in the cross-validation better than the PO + PA + samp. model for species which had relatively many PO observations. This indicates that in the spatially restricted cross-validation, the PO data more likely confounded the prediction of the PO + PA + samp. model than improved it possibly due to that the PA model was estimated with data only from relatively close to the validation points, whereas the PO model (especially for data abundant species) was estimated with data from across the whole continent. Locally estimated models, such as the PA model, can perform relatively well in cross-validation in a small area, but poorly in predicting across large spatial extents, known as poor model transferability (Yates et al., 2018). This has been shown for Ebird data when comparing locally adjusted spatio-temporal models to fixed models at different spatial extents (Fink et al., 2010). Contrary, when predicting the species' whole range, the PO + PA + samp. model performed better than the PO model when there were only few PO observations. Although, on average, over the study species the PO and PO + PA + samp. models performed equally well, supporting the model with the PA data improved the range prediction for relatively scarcely sampled species. Range size was not associated with the change in model performance between the models. Similar to the model evaluation, also the associations between data or species characteristics and the benefits of data integration depended on the spatial extent of the validation data, as data integration was shown to be more beneficial for data-deficient species on the extent of species' whole range. However, we think that it is not only the number of data points that controls the possible gain of predictive accuracy from data integration, but that the predictive improvement from data integration depends also on how well the data sets cover the environmental gradients that are accessible to the species and the level of preferential sampling of the PO data. We see this as an interesting question for future studies.

From a geographical point of view, the PA model suffered from over-predicting a species population outside of the species range and the PO model suffered from over-predicting species richness in the area of high sampling intensity (the latter was also found by Ellis-Soto et al., 2021). These errors were corrected to some extent by the PO + PA + samp. model supporting its use for studying unevenly sampled areas, such as remote areas in the tropics (Oliver et al., 2021) or marine regions (Menegotto & Rangel, 2018). Previous

efforts in integrated SDMs, including this study, have examined charismatic species, which are commonly sampled by public surveyors and have long-term well-established systematic survey protocols, such as birds (Meyer et al., 2015). Integrated modelling approach can be used to improve spatial information for less commonly surveyed species groups, like invertebrates (Meyer et al., 2015), which may be surveyed systematically only in a small part of the species geographical range.

4.3 | Limitations of the study and future perspectives

By involving different types of species information in the model inference and validation propagates uncertainties and biases in the results. We considered expert range maps as fixed truth of species occurrence status despite expert range maps have been shown to suffer from inaccuracies (Graham & Hijmans, 2006). However, we could alleviate the inaccuracies by coarsening the resolution of the model validation (see Hurlbert & Jetz, 2007). For all species, this may not have been an adequate treatment and validation would have required an even coarser resolution. The fact that cross-validation supported the PA model over all other models, whereas expert range map-based validation gave the opposite result, may have been partly due to the inaccuracies of the range maps. Still, many results support the conclusion that the PA models are inaccurate as providing overly small estimates for the covariate effects and hence predicting larger species ranges than what other models predict, or expert range maps present (see Figure S1.4 in Appendix 1). For hummingbirds, expert range maps can be considered a credible information source, but we cannot make this assumption for all taxa since for many species groups range maps may suffer from larger inaccuracies or simply, they do not exist.

New approaches to account for the sampling bias of the PO data are currently being developed (Chauvier et al., 2021; Nolan et al., 2021). Here, we used the approach from previous studies (Gelfand & Shirota, 2019; Simmonds et al., 2020) and modelled the sampling intensity of the PO data with a spatial latent effect. We used presence records of only the focal species for estimating the spatial latent effect, but given the general interest of the surveyors to all the study species, we could have fitted it over presence records of all species. A similar approach for borrowing information between species has been used to constrain likelihood estimation only to locations, where any of the study species has been sampled (target group sampling) (Chakraborty et al., 2011; Phillips et al., 2009). Estimating the sampling intensity with a spatial latent effect over all sampling locations and adding it to the function of species log intensity would provide a probabilistic alternative to fixing background sites to observation locations of any species and assuming a constant survey effort over the locations.

Observational survey data are still only one of the information sources for species distributions and environmental niches. Previous studies have shown that integrating PO data and spatial information from expert range maps or elevational tolerance limits of the species provides good results too (Ellis-Soto et al., 2021; Merow et al., 2017). On the other hand, physical performance experiments provide detailed information about individual level responses to environmental variables, and estimated responses from the experiments can be further used to project a species distribution to a landscape (Muñoz et al., 2021). New types of integrated SDMs (Kotta et al., 2019; Talluto et al., 2016) provide interesting approaches for integrating experiments and observational data in a single model through a shared latent process, like the structure used in the PO + PA + samp. model. These information sources should be considered in future studies and flexible methods for modelling jointly such discrepant observations are highly needed.

Our models did not address the fact that hummingbirds' distributions are constrained by mountain chains and dispersal barriers (Weinstein et al., 2014). Although spatial latent effects assumingly capture dispersal processes, we modelled the decay of covariance in Euclidean space and did not account for the impact of dispersal barriers on the covariance. Adding dispersal constrains in the covariance structure has been shown by Bakka et al. (2019), and for hummingbirds dispersal barriers could be derived from elevational tolerance limits, which were used for restricting species distribution predictions (Ellis-Soto et al., 2021). This approach could be used to improve the identifiability of different processes impacting the geographic distribution of a species population and we see it as an interesting research question for future studies.

5 | CONCLUSION

We have shown that compared to traditional models using a single data type, integrated SDMs provide advantages for predicting a species entire range. Specifically, integrated SDMs offered predictions of greater accuracy and lower uncertainty. This predictive improvement is conditional on explicitly modelling the sampling intensity of the opportunistically sampled data. This approach simultaneously solves the problems of small covariate effect estimates, which are common in models fitted with ecological surveys covering a relatively small spatial extent and sampling bias, which is common in heterogeneous and opportunistic species data covering a species' whole range. The improvement in the accuracy of single species range predictions propagates also to community level species richness predictions and has the potential to refine macro-level richness mappings. Integrated modelling can alleviate the problem of taxonomic and geographic data gaps and biases by combining the respective strengths of the PO and PA data and thus support and improve spatial information basis for large-scale biodiversity science and conservation.

ACKNOWLEDGEMENTS

The authors thank the Map of Life (mol.org) team for their essential work on managing the global database for environment and species. The authors are grateful to all members of the Jetzlab for creating a friendly and inspiring workspace.

FUNDING INFORMATION

This work was supported by the E.O. Wilson Biodiversity Foundation in association with the Half-Earth Project. CM was supported by grants from NSF (2225078) and NASA (80NSSC22K0883) and WJ was supported by grants from NASA (80NSSC17K0282 and 80NSSC18K0435).

CONFLICT OF INTEREST STATEMENT

The authors do not have a conflict of interest influencing their objectivity regarding this study.

DATA AVAILABILITY STATEMENT

Species and environmental data and scripts for running the analysis are stored in https://doi.org/10.5061/dryad.k98sf7mdg (Mäkinen et al., 2023).

ORCID

Jussi Mäkinen https://orcid.org/0000-0001-6599-8279
Cory Merow https://orcid.org/0000-0003-0561-053X

REFERENCES

Ahmad Suhaimi, S. S., Blair, G. S., Jarvis, S. G., & Leroy, B. (2021). Integrated species distribution models: A comparison of approaches under different data quality scenarios. *Diversity and Distributions*, *27*(6), 1066–1075.

Amano, T., Lamming, J. D., & Sutherland, W. J. (2016). Spatial gaps in global biodiversity information and the role of citizen science. *Bioscience*, *66*(5), 393–400.

Amatulli, G., Domisch, S., Tuanmu, M. N., Parmentier, B., Ranipeta, A., Malczyk, J., & Jetz, W. (2018). A suite of global, cross-scale topographic variables for environmental and biodiversity modeling. *Scientific Data*, *5*(1), 1–15.

Bakka, H., Rue, H., Fuglstad, G. A., Riebler, A., Bolin, D., Illian, J., Krainski, E., Simpson, D., & Lindgren, F. (2018). Spatial modeling with R-INLA: A review. *WIREs Computational Statistics*, *10*(6), e1443.

Bakka, H., Vanhatalo, J., Illian, J. B., Simpson, D., & Rue, H. (2019). Non-stationary Gaussian models with physical barriers. *Spatial Statistics*, *29*, 268–288.

Bowler, D. E., Nilsen, E. B., Bischof, R., O'Hara, R. B., Yu, T. T., Oo, T., Aung, M., & Linnell, J. D. C. (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), 7766.

Chakraborty, A., Gelfand, A. E., Wilson, A. M., Latimer, A. M., & Silander, J. A. (2011). Point pattern modelling for degraded presence-only data over large regions. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, *60*(5), 757–776.

Chauvier, Y., Zimmermann, N. E., Poggiato, G., Bystrova, D., Brun, P., Thuiller, W., & Qiao, H. (2021). Novel methods to correct for observer and sampling bias in presence-only species distribution models. *Global Ecology and Biogeography*, *30*(11), 2312–2325.

Chevalier, M., Broennimann, O., Cornault, J., & Guisan, A. (2021). Data integration methods to account for spatial niche truncation effects in regional projections of species distribution. *Ecological Applications*, *31*(7), e02427.

Chevalier, M., Zarzo-Arias, A., Guélat, J., Mateo, R. G., & Guisan, A. (2022). Accounting for niche truncation to improve spatial and temporal predictions of species distributions. *Frontiers in Ecology and Evolution*, *10*, 944116.

Crase, B., Liedloff, A. C., & Wintle, B. A. (2012). A new method for dealing with residual spatial autocorrelation in species distribution models. *Ecography*, *35*(10), 879–888.

Dawson, T. P., Jackson, S. T., House, J. I., Prentice, I. C., & Mace, G. M. (2011). Beyond predictions: Biodiversity conservation in a changing climate. *Science*, *332*(6025), 53–58.

Didan, K. (2015). MOD13Q1 MODIS/Terra vegetation indices 16-day L3 global 250m SIN grid V006. *NASA EOSDIS Land Processes DAAC*, *10*, 415.

Diggle, P. J., Menezes, R., & Su, T. L. (2010). Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, *59*(2), 191–232.

Dorazio, R. M. (2014). Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. *Global Ecology and Biogeography*, *23*(12), 1472–1484.

Doser, J. W., Leuenberger, W., Sillett, T. S., Hallworth, M. T., & Zipkin, E. F. (2022). Integrated community occupancy models: A framework to assess occurrence and biodiversity dynamics using multiple data sources. *Methods in Ecology and Evolution*, *13*(4), 919–932.

Ellis-Soto, D., Merow, C., Amatulli, G., Parra, J. L., & Jetz, W. (2021). Continental-scale 1 km hummingbird diversity derived from fusing point records with lateral and elevational expert information. *Ecography*, *44*(4), 640–652.

Farr, M. T., Green, D. S., Holekamp, K. E., & Zipkin, E. F. (2021). Integrating distance sampling and presence-only data to estimate species abundance. *Ecology*, *102*(1), e03204.

Fink, D., Hochachka, W. M., Zuckerberg, B., Winkler, D. W., Shaby, B., Munson, M. A., Hooker, G., Riedewald, M., Sheldon, D., & Kelling, S. (2010). Spatiotemporal exploratory models for broad-scale survey data. *Ecological Applications*, *20*(8), 2131–2147.

Fithian, W., Elith, J., Hastie, T., & Keith, D. A. (2015). Bias correction in species distribution models: Pooling survey and collection data for multiple species. *Methods in Ecology and Evolution*, *6*(4), 424–438.

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.

Fletcher, R. J., McCleery, R. A., Greene, D. U., & Tye, C. A. (2015). Integrated models that unite local and regional data reveal larger-scale environmental relationships and improve predictions of species distributions. *Landscape Ecology*, *31*(6), 1369–1382.

Fuglstad, G.-A., Simpson, D., Lindgren, F., & Rue, H. (2018). Constructing priors that penalize the complexity of Gaussian random fields. *Journal of the American Statistical Association*, *114*(525), 445–452.

Gelfand, A. E., & Shirota, S. (2019). Preferential sampling for presence/absence data and for fusion of presence/absence data with presence-only data. *Ecological Monographs*, *89*(3), e01372.

Graham, C. H., & Hijmans, R. J. (2006). A comparison of methods for mapping species ranges and species richness. *Global Ecology and Biogeography*, *15*(6), 578–587.

Guillera-Arroita, G., Lahoz-Monfort, J. J., Elith, J., Gordon, A., Kujala, H., Lentini, P. E., McCarthy, M. A., Tingley, R., & Wintle, B. A. (2015). Is my species distribution model fit for purpose? Matching data and models to applications. *Global Ecology and Biogeography*, *24*(3), 276–292.

Guisan, A., Tingley, R., Baumgartner, J. B., Naujokaitis-Lewis, I., Sutcliffe, P. R., Tulloch, A. I., Regan, T. J., Brotons, L., McDonald-Madden, E., Mantyka-Pringle, C., Martin, T. G., Rhodes, J. R., Maggini, R., Setterfield, S. A., Elith, J., Schwartz, M. W., Wintle, B. A., Broennimann, O., Austin, M., … Buckley, Y. M. (2013). Predicting species distributions for conservation decisions. *Ecology Letters*, *16*(12), 1424–1435.

Hanks, E. M., Schliep, E. M., Hooten, M. B., & Hoeting, J. A. (2015). Restricted spatial regression in practice: Geostatistical models, confounding, and robustness under model misspecification. *Environmetrics*, *26*(4), 243–254.

Hannah, L., Roehrdanz, P. R., Marquet, P. A., Enquist, B. J., Midgley, G., Foden, W., Lovett, J. C., Corlett, R. T., Corcoran, D., Butchart, S. H. M., Boyle, B., Feng, X., Maitner, B., Fajardo, J., McGill, B. J., Merow, C., Morueta-Holme, N., Newman, E. A., Park, D. S., … Svenning, J. C. (2020). 30% land conservation and climate action reduces tropical extinction risk by more than 50%. *Ecography*, *43*(7), 943–953.

Hawkins, B. A., Diniz-Filho, J. A. F., Bini, L. M., De Marco, P., & Blackburn, T. M. (2007). Red herrings revisited: Spatial autocorrelation and parameter estimation in geographical ecology. *Ecography*, *30*(3), 375–384.

Hodges, J. S., & Reich, B. J. (2010). Adding spatially-correlated errors can mess up the fixed effect you love. *The American Statistician*, *64*(4), 325–334.

Hurlbert, A. H., & Jetz, W. (2007). Species richness, hotspots and the scale dependence of range maps in ecology and conservation. *Proceedings of the National Academy of Sciences*, *104*(33), 13384–13389.

Isaac, N. J. B., Jarzyna, M. A., Keil, P., Dambly, L. I., Boersch-Supan, P. H., Browning, E., Freeman, S. N., Golding, N., Guillera-Arroita, G., Henrys, P. A., Jarvis, S., Lahoz-Monfort, J., Pagel, J., Pescott, O. L., Schmucki, R., Simmonds, E. G., & O'Hara, R. B. (2020). Data integration for large-scale models of species distributions. *Trends in Ecology & Evolution*, *35*(1), 56–67.

Jetz, W., McGeoch, M. A., Guralnick, R., Ferrier, S., Beck, J., Costello, M. J., Fernandez, M., Geller, G. N., Keil, P., Merow, C., Meyer, C., Muller-Karger, F. E., Pereira, H. M., Regan, E. C., Schmeller, D. S., & Turak, E. (2019). Essential biodiversity variables for mapping and monitoring species populations. *Nature Ecology & Evolution*, *3*(4), 539–551.

Jetz, W., McGowan, J., Rinnan, D. S., Possingham, H. P., Visconti, P., O'Donnell, B., & Londono-Murcia, M. C. (2022). Include biodiversity representation indicators in area-based conservation targets. *Nature Ecology & Evolution*, *6*(2), 123–126.

Jetz, W., Thomas, G. H., Joy, J. B., Hartmann, K., & Mooers, A. O. (2012). The global diversity of birds in space and time. *Nature*, *491*(7424), 444–448.

Jetz, W., Wilcove, D. S., & Dobson, A. P. (2007). Projected impacts of climate and land-use change on the global diversity of birds. *PLoS Biology*, *5*(6), e157.

Jung, M., Arnell, A., de Lamo, X., Garcia-Rangel, S., Lewis, M., Mark, J., Merow, C., Miles, L., Ondo, I., Pironon, S., Ravilious, C., Rivers, M., Schepaschenko, D., Tallowin, O., van Soesbergen, A., Govaerts, R., Boyle, B. L., Enquist, B. J., Feng, X., … Visconti, P. (2021). Areas of global importance for conserving terrestrial biodiversity, carbon and water. *Nature Ecology & Evolution*, *5*(11), 1499–1509.

Karger, D. N., Conrad, O., Bohner, J., Kawohl, T., Kreft, H., Soria-Auza, R., Zimmermann, N. E., Linder, H. P., & Kessler, M. (2017a). Data from: Climatologies at high resolution for the earth's land surface areas [Dataset]. Dryad. https://doi.org/10.5061/dryad.kd1d4

Karger, D. N., Conrad, O., Bohner, J., Kawohl, T., Kreft, H., Soria-Auza, R. W., Zimmermann, N. E., Linder, H. P., & Kessler, M. (2017b). Climatologies at high resolution for the earth's land surface areas. *Scientific Data*, *4*(1), 1–20.

Kim, D. (2021). Predicting the magnitude of residual spatial autocorrelation in geographical ecology. *Ecography*, *44*(7), 1121–1130.

Koshkina, V., Wang, Y., Gordon, A., Dorazio, R. M., White, M., Stone, L., & Warton, D. (2017). Integrated species distribution models: Combining presence-background data and site-occupancy data with imperfect detection. *Methods in Ecology and Evolution*, *8*(4), 420–430.

Kotta, J., Vanhatalo, J., Jänes, H., Orav-Kotta, H., Rugiu, L., Jormalainen, V., Bobsien, I., Viitasalo, M., Virtanen, E., Sandman, A. N., Isaeus, M., Leidenberger, S., Jonsson, P. R., & Johannesson, K. (2019). Integrating experimental and distribution data to predict future species patterns. *Scientific Reports*, *9*(1), 1–14.

Latimer, A. M., Banerjee, S., Sang, H., Jr., Mosher, E. S., & Silander, J. A., Jr. (2009). Hierarchical models facilitate spatial analysis of large data sets: A case study on invasive plant species in the northeastern United States. *Ecology Letters*, *12*(2), 144–154.

Lindgren, F., & Rue, H. (2015). Bayesian spatial modelling with R-INLA. *Journal of Statistical Software*, *63*(19), 1–25.

Lu, M., & Jetz, W. (2023). Scale-sensitivity in the measurement and interpretation of environmental niches. *Trends in Ecology & Evolution*, *38*, 554–567.

Mäkinen, J., Merow, C., & Jetz, W. (2023). Data from "Integrated species distribution models to account for sampling biases and improve range wide occurrence predictions" [Data set]. Dryad. https://doi.org/10.5061/dryad.k98sf7mdg

Mäkinen, J., Numminen, E., Niittynen, P., Luoto, M., & Vanhatalo, J. (2022). Spatial confounding in Bayesian species distribution modeling. *Ecography*, *2022*(11), e06183.

Mäkinen, J., & Vanhatalo, J. (2018). Hierarchical Bayesian model reveals the distributional shifts of Arctic marine mammals. *Diversity and Distributions*, *24*(10), 1381–1394.

Martino, S., Pace, D. S., Moro, S., Casoli, E., Ventura, D., Frachea, A., Silvestri, M., Arcangeli, A., Giacomini, G., Ardizzone, G., & Jona Lasinio, G. (2021). Integration of presence-only data from several sources: A case study on dolphins' spatial distribution. *Ecography*, *44*(10), 1533–1543.

Matthiopoulos, J., Fieberg, J., & Aarts, G. (2020). *Species-Habitat Associations: Spatial data, predictive models, and ecological insights*. University of Minnesota Libraries Publishing.

Menegotto, A., & Rangel, T. F. (2018). Mapping knowledge gaps in marine diversity reveals a latitudinal gradient of missing species richness. *Nature Communications*, *9*(1), 1–6.

Merow, C., Wilson, A. M., & Jetz, W. (2017). Integrating occurrence data and expert maps for improved species range predictions. *Global Ecology and Biogeography*, *26*(2), 243–258.

Meyer, C., Kreft, H., Guralnick, R., & Jetz, W. (2015). Global priorities for an effective information basis of biodiversity distributions. *Nature Communications*, *6*(1), 1–8.

Moraga, P., Cramb, S. M., Mengersen, K. L., & Pagano, M. (2017). A geostatistical model for combined analysis of point-level and area-level data using INLA and SPDE. *Spatial Statistics*, *21*, 27–41.

Muñoz, M. M., Feeley, K. J., Martin, P. H., & Farallo, V. R. (2021). The multidimensional (and contrasting) effects of environmental warming on a group of montane tropical lizards. *Functional Ecology*, *36*(2), 419–431.

Nolan, V., Gilbert, F., & Reader, T. (2021). Solving sampling bias problems in presence–absence or presence-only species data using zero-inflated models. *Journal of Biogeography*, *49*(1), 215–232.

Oliver, R. Y., Meyer, C., Ranipeta, A., Winner, K., & Jetz, W. (2021). Global and national trends, gaps, and opportunities in documenting and monitoring species distributions. *PLoS Biology*, *19*(8), e3001336.

Pacifici, K., Reich, B. J., Miller, D. A. W., Gardner, B., Stauffer, G., Singh, S., McKerrow, A., & Collazo, J. A. (2017). Integrating multiple data sources in species distribution modeling: A framework for data fusion. *Ecology*, *98*(3), 840–850.

Phillips, S. J., Dudík, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J., & Ferrier, S. (2009). Sample selection bias and presence-only distribution models: Implications for background and pseudo-absence data. *Ecological Applications*, *19*(1), 181–197.

Pineda, E., & Lobo, J. M. (2012). The performance of range maps and species distribution models representing the geographic variation of species richness at different resolutions. *Global Ecology and Biogeography*, *21*(9), 935–944.

Regan, H. M., Ben-Haim, Y., Langford, B., Wilson, W. G., Lundberg, P., Andelman, S. J., & Burgman, M. A. (2005). Robust decision-making under severe uncertainty for conservation management. *Ecological Applications*, *15*(4), 1471–1477.

Reich, B. J., Hodges, J. S., & Zadnik, V. (2006). Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models. *Biometrics*, *62*(4), 1197–1206.

Renner, I. W., Elith, J., Baddeley, A., Fithian, W., Hastie, T., Phillips, S. J., Popovic, G., Warton, D. I., & O'Hara, R. B. (2015). Point process models for presence-only analysis. *Methods in Ecology and Evolution*, *6*(4), 366–379.

Renner, I. W., Louvrier, J., & Gimenez, O. (2019). Combining multiple data sources in species distribution models while accounting for spatial dependence and overfitting with combined penalized likelihood maximization. *Methods in Ecology and Evolution*, *10*(12), 2118–2128.

Ridgely, R., Allnutt, T., Brooks, T., McNicol, D., Mehlman, D., Young, B., & Zook, J. (2003). *Digital distribution maps of the birds of the Western hemisphere*. Version.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., & Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, *40*(8), 913–929.

Robinson, O. J., Ruiz-Gutierrez, V., Reynolds, M. D., Golet, G. H., Strimas-Mackey, M., Fink, D., & Maiorano, L. (2020). Integrating citizen science data with expert surveys increases accuracy and spatial extent of species distribution models. *Diversity and Distributions*, *26*(8), 976–986.

Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P., & Lindgren, F. K. (2017). Bayesian computing with INLA: A review. *Annual Review of Statistics and its Application*, *4*(1), 395–421.

Rufener, M. C., Kristensen, K., Nielsen, J. R., & Bastardie, F. (2021). Bridging the gap between commercial fisheries and survey data to model the spatiotemporal dynamics of marine species. *Ecological Applications*, *31*(8), e02453.

Schank, C. J., Cove, M. V., Kelly, M. J., Mendoza, E., O'Farrill, G., Reyna-Hurtado, R., Meyer, N., Jordan, C. A., González-Maya, J. F., Lizcano, D. J., Moreno, R., Dobbins, M. T., Montalvo, V., Sáenz-Bolaños, C., Jimenez, E. C., Estrada, N., Díaz, J. C. C., Saenz, J., Spínola, M., & Miller, J. A. (2017). Using a novel model approach to assess the distribution and conservation status of the endangered Baird's tapir. *Diversity and Distributions*, *23*(12), 1459–1471.

Schmitt, S., Pouteau, R., Justeau, D., De Boissieu, F., & Birnbaum, P. (2017). ssdm: An r package to predict distribution of species richness and composition based on stacked species distribution models. *Methods in Ecology and Evolution*, *8*(12), 1795–1803.

Simmonds, E. G., Jarvis, S. G., Henrys, P. A., Isaac, N. J. B., & O'Hara, R. B. (2020). Is more data always better? A simulation study of benefits and limitations of integrated distribution models. *Ecography*, *43*(10), 1413–1422.

Soriano-Redondo, A., Jones-Todd, C. M., Bearhop, S., Hilton, G. M., Lock, L., Stanbury, A., Votier, S. C., & Illian, J. B. (2019). Understanding species distribution in dynamic populations: A new approach using spatio-temporal point process models. *Ecography*, *42*(6), 1092–1102.

Talluto, M. V., Boulangeat, I., Ameztegui, A., Aubin, I., Berteaux, D., Butler, A., Doyon, F., Drever, C. R., Fortin, M.-J., Franceschini, T., Liénard, J., McKenney, D., Solarik, K. A., Strigul, N., Thuiller, W., & Gravel, D. (2016). Cross-scale integration of knowledge for predicting species ranges: A metamodelling framework. *Global Ecology and Biogeography*, *25*(2), 238–249.

Valavi, R., Guillera-Arroita, G., Lahoz-Monfort, J. J., & Elith, J. (2022). Predictive performance of presence-only species distribution models: A benchmark study with reproducible code. *Ecological Monographs*, *92*(1), e01486.

Warton, D. I., Renner, I. W., & Ramp, D. (2013). Model-based control of observer bias for the analysis of presence-only data in ecology. *PLoS One*, *8*(11), e79168.

Warton, D. I., & Shepherd, L. C. (2010). Poisson point process models solve the "pseudo-absence problem" for presence-only data in ecology. *The Annals of Applied Statistics*, *4*(3), 1383–1402.

Weinstein, B. G., Tinoco, B., Parra, J. L., Brown, L. M., McGuire, J. A., Stiles, F. G., & Graham, C. H. (2014). Taxonomic, phylogenetic, and trait beta diversity in south American hummingbirds. *The American Naturalist*, *184*(2), 211–224.

Wenger, S. J., & Olden, J. D. (2012). Assessing transferability of ecological models: An underappreciated aspect of statistical validation. *Methods in Ecology and Evolution*, *3*(2), 260–267.

Williams, P. J., & Hooten, M. B. (2016). Combining statistical inference and decisions in ecology. *Ecological Applications*, *26*(6), 1930–1942.

Wilson, A. M., & Jetz, W. (2016). Remotely sensed high-resolution global cloud dynamics for predicting ecosystem and biodiversity distributions. *PLoS Biology*, *14*(3), e1002415.

Wüest, R. O., Zimmermann, N. E., Zurell, D., Alexander, J. M., Fritz, S. A., Hof, C., Kreft, H., Normand, S., Cabral, J. S., Szekely, E., Thuiller, W., Wikelski, M., & Karger, D. N. (2019). Macroecology in the age of big data—Where to go from here? *Journal of Biogeography*, *47*(1), 1–12.

Yates, K. L., Bouchet, P. J., Caley, M. J., Mengersen, K., Randin, C. F., Parnell, S., Fielding, A. H., Bamford, A. J., Ban, S., Barbosa, A. M., Dormann, C. F., Elith, J., Embling, C. B., Ervin, G. N., Fisher, R., Gould, S., Graf, R. F., Gregr, E. J., Halpin, P. N., … Sequeira, A. M. M. (2018). Outstanding challenges in the transferability of ecological models. *Trends in Ecology & Evolution*, *33*(10), 790–802.

Zulian, V., Miller, D. A. W., Ferraz, G., & Jung, M. (2021). Integrating citizen-science and planned-survey data improves species distribution estimates. *Diversity and Distributions*, *27*(12), 2498–2509.

BIOSKETCH

Jussi Mäkinen focuses on statistical ecology to study species populations and processes impacting them. His work integrates theories, methods and data from different sources and organizational levels.

Cory Merow is a quantitative ecologist interested in forecasting biological responses to global change. He likes pina coladas and getting caught in the rain.

Walter Jetz' work addresses patterns and mechanisms of changing biodiversity distribution and the resulting implications on conservation and environmental management. His research combines remote sensing, phylogenetic, functional and spatio-temporal biodiversity data with new modelling approaches and informatics tools.

SUPPORTING INFORMATION

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

How to cite this article: Mäkinen, J., Merow, C., & Jetz, W. (2024). Integrated species distribution models to account for sampling biases and improve range-wide occurrence predictions. *Global Ecology and Biogeography*, *33*, 356–370. https://doi.org/10.1111/geb.13792