Check for updates

Received: 24 July 2020 | Revised: 29 January 2021 | Accepted: 8 February 2021

DOI: 10.1111/ddi.13255

BIODIVERSITY METHODS

Diversity and Distributions WILEY

Integrated species distribution models: A comparison of approaches under different data quality scenarios

Siti Sarah Ahmad Suhaimi[1] | Gordon S. Blair[1,2] | Susan G. Jarvis[3]

[1]School of Computing and Communications, Lancaster University, Lancaster, UK

[2]Centre for Excellence in Environmental Data Science (CEEDS), Lancaster Environment Centre, Lancaster, UK

[3]UK Centre for Ecology & Hydrology, Lancaster Environment Centre, Lancaster, UK

Correspondence
Susan G. Jarvis, UK Centre for Ecology & Hydrology, Lancaster Environment Centre, Bailrigg, Lancaster, LA1 4AP, UK.
Email: susjar@ceh.ac.uk

Present address
Siti Sarah Ahmad Suhaimi, PPG Coatings (Malaysia), UOA Business Park, Shah Alam, Malaysia

Funding information
Engineering and Physical Sciences Research Council, Grant/Award Number: EP/P002285/1; Natural Environment Research Council, Grant/Award Number: NE/R016429/1

Editor: Boris Leroy

Abstract

Aim: Integrated species distribution modelling has emerged as a useful tool for ecologists to exploit the range of information available on where species occur. In particular, the ability to combine large numbers of ad hoc or presence-only (PO) records with more structured presence–absence (PA) data can allow ecologists to account for biases in PO data which often confound modelling efforts. A range of modelling techniques have been suggested to implement integrated species distribution models (IDMs) including joint likelihood models, including one dataset as a covariate or informative prior, and fitting a correlation structure between datasets. We aim to investigate the performance of different types of integrated models under realistic ecological data scenarios.

Innovation: We use a virtual ecologist approach to investigate which integrated model is most advantageous under varying levels of spatial bias in PO data, sample size of PA data and spatial overlap between datasets.

Main conclusions: Joint likelihood models were the best performing models when spatial bias in PO data was low, or could be modelled, but gave poor estimates when there were unknown biases in the data. Correlation models provided good model estimates even when there were unknown biases and when good quality PA data were spatially limited. Including PO data via an informative prior provided little improvement over modelling PA data alone and was inferior to using either the joint likelihood or correlation approach. Our results suggest that correlation models provide a robust alternative to joint likelihood models when covariates related to effort or detection in PO data are not available. Ecologists should be aware of the limitations of each approach and consider how well biases in the data can be modelled when deciding which type of IDM to use.

KEYWORDS

citizen science, distribution, informative prior, integrated model, joint likelihood, presence–absence, presence-only, simulation

Gordon S. Blair and Susan G. Jarvis should be considered joint senior author.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2021 The Authors. *Diversity and Distributions* published by John Wiley & Sons Ltd.

1 | INTRODUCTION

Integrated species distribution models (IDMs), which combine multiple data sources to model species distributions, are becoming increasingly common (Isaac et al., 2020). IDMs allow data of different types, such as structured samples from formal scientific research and opportunistic data from alternative sources like museums and citizen science programmes, to be combined in a single model (Isaac et al., 2020; Miller et al., 2019; Zipkin et al., 2019). Structured sampling is expensive and thus often spatially restricted while opportunistic data are often abundant and readily available. Combining both types of data can capitalize on the advantages of each data type and provide better predictions of species distributions and their drivers (Miller et al., 2019). Utilizing all available data is useful particularly for developing countries, which can face resource constraints hindering efforts for extensive conservation research (Bowler et al., 2019).

Studies have suggested different approaches to integrated distribution models. Fletcher et al. (2019) outline a broad range of integrated modelling techniques including data pooling, ensemble models, using auxiliary data and weighted joint likelihood. Data pooling is a simple method that is commonly used in practice where data are simply pooled by transforming or simplifying the more sophisticated dataset to accommodate the structure of a simpler second dataset, for example, by downgrading abundance data to presence–absence data prior to pooling, which ultimately results in some lost information. Data pooling is not always appropriate since it does not account for differences in the data sources like sampling biases, spatial and temporal variation in sampling effort, and differences in collection protocols.

Pacifici et al. (2017) outlined three approaches which move beyond data pooling towards true integrated species distribution models (IDMs) termed the joint likelihood, correlation and covariate models. The joint likelihood method simultaneously fits a likelihood to both data sources while accounting for different data types and other observation processes. The covariate method incorporates information from a second dataset via a fixed effect. For correlation modelling, the datasets are indirectly connected through a shared covariance matrix that captures similar patterns present in both data sources. Additional approaches, such as incorporating information from one dataset as an informative prior, have also been suggested (Miller et al., 2019).

In their study, Pacifici et al. (2017) found that integrated models consistently performed better than models fitted with single data sources. The relative performance between the three integrated methods tested depended on the information held by the unstructured data source. The joint likelihood method was found to be more sensitive to the quality of the unstructured data source compared to covariate and correlation modelling, but all performed relatively well. We may expect that joint likelihood models perform best when bias is low and data of both types are plentiful (Fletcher et al., 2019; Simmonds et al., 2020). It is not yet clear whether alternative approaches may outperform joint likelihood models when some data are spatially biased or low in quantity, or when information on spatial biases is lacking.

Opportunistic data are often biased towards areas of high human population density and to areas that are easily accessible for recording (Isaac & Pocock, 2015). In many cases, information on effort in unstructured data is lacking, although covariates, such as human population size or distance to roads, may be available (Fithian et al., 2015). However, in some cases the causes of spatial bias may be unknown or suitable covariates to explain spatial bias in recording may not be available. In these cases, researchers will need to consider whether spatially biased opportunistic data can still be used to model species distributions. Joint likelihood approaches to data integration have been shown to be very sensitive to unknown spatial biases (Simmonds et al., 2020), producing poorer results than single dataset analyses. However, integrated models that include a lower quality dataset via a correlation structure or covariate are hypothesized to be more robust to issues such as unknown spatial bias as the degree of information sharing is less than in a joint likelihood model (Pacifici et al., 2017).

Another challenge for integrating distribution data is lack of overlap in spatial extent of different data sources (Bowler et al., 2019). High-quality professional survey data are very expensive to collect, and so researchers may often be faced with a case where high-quality data are available for only a subset of the total area of interest, for example for a subset of countries across a larger region. Joint likelihood approaches have been shown to perform well when spatial extents are not the same under some conditions (Koshkina et al., 2017), but alternative approaches have not yet been tested. We might expect the correlation approach in particular to perform poorly under conditions of low spatial overlap as there is less area from which to estimate the correlation between datasets. A key factor determining model performance might be the degree to which a spatially restricted PA survey covers both environmental gradients and any gradients in PO data effort.

To assist ecologists in choosing the most appropriate modelling approach, we need to consider the performance of each method under a range of different data conditions. To do so, we use a virtual ecologist approach (Zurell et al., 2010), whereby we simulate a species distribution and sample from it under different scenarios. The strength of this approach is that we know the true distribution and can compare model performance to this known truth. In our simulations, we consider a scenario where a researcher is in possession of both "high-quality" presence–absence (PA) data and an unstructured citizen science type dataset with presence-only (PO) information and needs to make a decision about whether to integrate the two datasets, and if so which approach to use. We consider the potential impacts on model choice of variation in sample size of PA data, the degree to which the PO dataset is spatially biased (e.g. due to spatial variation in the number of observers) and the overlap in spatial extents between the two datasets. We also consider the importance of knowledge of spatial bias in PO data by either including or excluding a covariate explaining the bias to test whether alternatives to the joint likelihood approach are more robust to unknown biases in data

sources. We test three integrated modelling approaches under the restrictive conditions of a single focal species and absence of repeat visits.

2 | METHODS

2.1 | Data generation and sampling

To assess the performance of three integrated SDM methods, we constructed a simulation study. We use the approach for data generation and sampling described in Simmonds et al. (2020), with minor modifications. Briefly, we generate an intensity surface over a 100 by 100 grid, which represents the true spatial patterns of species distributions (where intensity is high we expect to observe more individuals; Figure 1). The intensity surface was generated as a log-Gaussian Cox process. The intensity function assumes that the species distribution is determined by both an environmental effect and a random spatial term, the latter meaning that locations closer together are more likely to have similar numbers of individuals. To simulate the locations of individuals (here assumed to be equivalent to points for simplicity), we produce a separate realization from the log-Gaussian Cox process for each simulated survey. By creating separate realizations, we assume that the underlying intensity is the same across surveys but it is unlikely that the same exact individuals are recorded in both surveys.

Two different observation processes were simulated: a PA dataset simulating a professional survey and a PO dataset simulating unstructured citizen science data. Firstly, to sample the PA data from the generated true data, the whole field with dimension (100 × 100) was divided into 25 (20 × 20) squares. A stratified sampling scheme was simulated, as this is often used to ensure representative coverage in real-world surveys. A preset number of "samples" of size 1 × 1 was divided equally across strata, so that, for example, for 100 total samples, 4 samples would be taken from each stratum. The samples were recorded as presences if they intersected with a point in the point pattern and absences if they did not. We assumed that each location was visited only once and that there was perfect detection in PA samples. This is a strong assumption for most mobile taxa where perfect detection is unlikely. However, for easily identified sessile taxa, for example many flowering plants, the assumption of perfect detection in PA data may be a reasonable approximation.

Although some unstructured data may contain information on absences, the citizen science data simulated here were assumed to be presence-only data since much opportunistic data only hold presence information. PO data often come with some form of bias due to

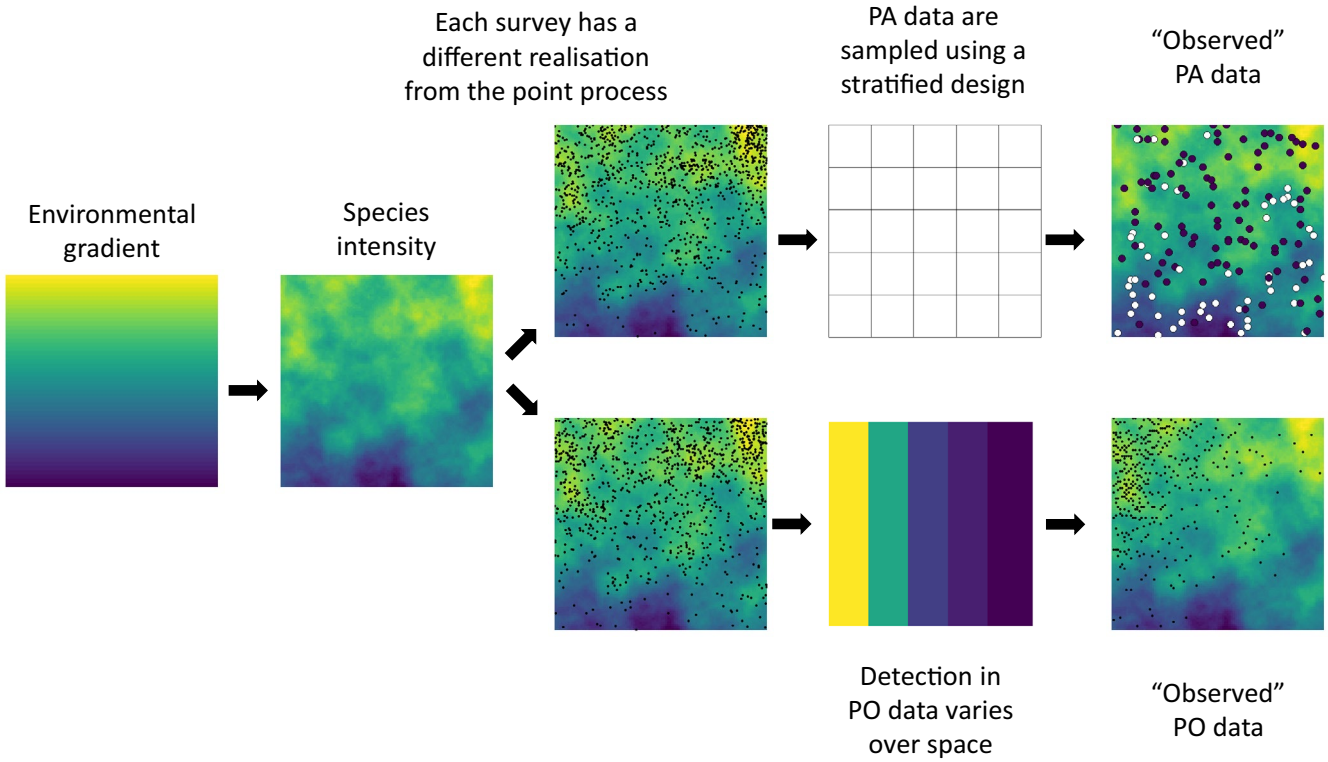

FIGURE 1 Visualization of the simulation methodology. Yellow indicates higher values, and blue indicates lower values throughout. The true species intensity is determined by an environmental gradient plus a random spatial term. From this intensity, a realization of individuals (small black points) is produced for each survey type. For PA data (top row), sampling locations are determined by a stratified sampling design with equal coverage across the domain. Observed PA data are represented by large black points for presences and large white points for absences. For PO data (bottom row), detection is made to vary across the sampling area with higher detection towards the left. Observed PO data are thinned by the detection probability so that the spatial distribution of PO data is influenced by both the true intensity and the detection surface

sampling effort, sampling frequency, coverage area and detectability (Isaac et al., 2014). For example, more sightings are expected near roads and towns where accessibility is high making it convenient for citizens to visit.

To incorporate bias in the PO data in the simulation, the strata were also used to vary detection probability across the domain setting it to be higher towards the left side, assuming higher concentration of opportunistic surveying in that area (Figure 1). The bias pattern simulated had a pattern perpendicular to the environmental covariate which should allow these processes to be separated (Fithian et al., 2015). Presence-only data generated from a realization of the log-Gaussian Cox process were then thinned using these detection probabilities. The number of PO samples varied for each simulation because both the intensity and thinning were stochastic processes. In some cases, a modeller may be able to explain varying detection probability in PO data with a covariate, for example representing survey effort or observer density. To simulate a covariate the modeller may be available the probabilities were transformed into a smooth pattern that varied from 1 to 0 along the x axis, simulating a good but not perfect covariate for spatial variation in detection.

2.2 | Joint likelihood model

Perhaps the most intuitive approach when faced with two datasets that we wish to integrate into a single model is to use a joint likelihood. In our simulation study, models for each dataset can be defined as follows. The PA data, which was recorded as either present or absent, followed the Bernoulli distribution.

$$Y_i \sim Bernoulli\left(p_i\right), i = 1, 2, \ldots, n \quad (1)$$

$$\log\left(-\log\left(1-p_i\right)\right) = \alpha_1 + \beta_1 x_{1,i} + f(s)$$

where Y_i is the response variable (presence or absence status) in sample i, p_i is the probability of success (presence), α_1 represents the intercept, β_1 the coefficient of environmental covariate x_1 and $f(s)$ a random spatial term. The random spatial effect is modelled as a latent Gaussian field approximated using stochastic partial differential equation (SPDE) with a Matérn covariance. Integrated nested Laplace approximation (INLA) can be used to estimate the solution of the SPDE (Lindgren et al., 2011). The linear predictor was linked to the success probability by the complementary log-log link (cloglog) function log(−log(1 − p)) (see Kery & Royle, 2016).

For the PO model, the dataset was generated from a log-Gaussian Cox process and we can model the expected number of individuals in an area with a Poisson distribution.

$$N(A) \sim Poisson\left(\int_A \lambda(s)\, d(s)\right) \quad (2)$$

$$\log\left(\lambda(s)\right) = \alpha_2 + \beta_1 x_{1,(s)} + \beta_2 x_{2,(s)} + f(s)$$

where N is the expected number of observations in an area A, λ is the mean intensity function, α_2 the intercept, β_1 the parameter coefficient for environmental covariate x_1, β_2 the parameter coefficient for bias covariate x_2 and $f(s)$ is the random spatial effect. Note that as in Simmonds et al. (2020) the thinning probability of the PO data is not explicitly estimated, but the covariate x_2 is used instead to explain variation in detection probability in PO data. In some models, covariate x_2 was not available to simulate situations where information on the spatial bias in PO data is missing.

To fit the PO data as a Poisson process, information on covariate values at integration points is also required. The method of Simpson et al. (2016) was used to derive integration points and set weights in the likelihood. This method is suitable when models are fit using INLA and requires a suitable triangulation (mesh) of the domain to be defined.

The joint likelihood was derived by multiplying the likelihoods of each dataset. To enable this, two parameters are shared between the individual dataset models: the coefficient β_1 and the random spatial effect $f(s)$. By using the combination of a log link for the Poisson model and a cloglog link for the binomial model both responses can be interpreted on the same scale, allowing for these parameters to be directly shared between the models (Kery & Royle, 2016).

2.3 | Informed prior model

Covariate models as defined by Pacifici et al. (2017) share information between datasets by including information from one dataset, usually the lower quality dataset, in the model of a second dataset via a fixed effect. One disadvantage of this approach using PO data as the first dataset is that a decision has to be made about the spatial scale at which PO data should be aggregated to produce a suitable covariate. Here we use a different approach whereby the first dataset influences the second via informative priors rather than a fixed effect (Miller et al., 2019). To do this, we sequentially model one dataset first and use the parameter estimates to derive informative priors for the second dataset model. The first dataset therefore contributes to the estimates in the second model, but the information shared is controlled by the influence of the prior. If the first model produces imprecise estimates of the shared parameters, then the priors will be less informative than if very precise estimates are obtained. We assumed that the PA dataset was a better quality dataset than the PO dataset as it was spatially unbiased and had perfect detection. Therefore, we modelled the PO dataset first and used the estimates to derive informative priors for the PA data model. The informative prior should influence the likelihood towards better inference because some knowledge about the species distribution can be derived from the abundant PO data within the same spatial domain.

Firstly, the PO data were modelled as in Equation 2 with uninformative priors. The posterior distributions of β_1 and $f(s)$ were then extracted from the PO model to include in the PA model as informative

priors. To extract parameters describing *f(s)*, the internal parameterization of the variance and spatial scale from the Matérn covariance, θ_1 and θ_2, were used. Priors on the intercepts in both models were kept as uninformative uniform distributions as intercepts were not shared between models.

To assess the impact of providing a suitable covariate for spatial bias in the PO model, the first model was fit both with and without knowledge of the bias covariate x_2. Even though there was no bias covariate in the second, PA, model, whether or not it was estimated in the PO model would influence the estimates of the environmental covariate and spatial field and therefore the informative priors used in the second model. These two scenarios reflect the fact that the researcher may or may not have knowledge about the processes that determine spatial bias in the PO data.

2.4 | Correlation model

The correlation method assumes that there is some spatial correlation between different data sources which can be estimated (Pacifici et al., 2017). Where a species is abundant in one data source, then, naturally, it should also be abundant in the other data source when both sources cover the same spatial domain.

To fit this model using INLA, we allowed the spatial random effects to be correlated between datasets using the "group" function. This approach is commonly used in spatio-temporal models, where there is a spatial correlation pattern which may be correlated over time (Blangiardo et al., 2013). Here, we estimated $f_1(s)$ for the PA data and $f_2(s)$ for PO data using SPDE (Equations 3 and 4) and estimated the correlation ρ between data sources using an exchangeable correlation structure whereby $f_1(s) = \rho f_2(s) + \varepsilon$. This relationship between $f_1(s)$ and $f_2(s)$ is controlled by the correlation coefficient ρ and also includes some error ε. Note that although the spatial fields were correlated rather than jointly estimated, the environmental covariate was estimated via joint likelihood, that is β_1 is shared across both models. This model therefore assumes that while the spatial pattern in observations may not be identical between datasets, for example due to unknown biases, the effect of the environmental covariate is still shared.

$$\log\left(-\log\left(1-p_i\right)\right) = \alpha_1 + \beta_1 x_{1,i} + f_1(s) \quad (3)$$

$$\log\left(\lambda(s)\right) = \alpha_2 + \beta_1 x_{1,(s)} + \beta_2 x_{2,(s)} + f_2(s) \quad (4)$$

Since this type of integrated species distribution model only assumes a spatial correlation function exists between data sources, the outcome of the model fitting will be two separate spatial fields. So, to make predictions one must choose which spatial field to use. With the same reasoning as before, the perfect detection in PA data made it the better choice to be used to create predictions from the correlation model. If the correlation between the datasets is high, the choice of grouping for predictions would not matter greatly and both groupings would perform similarly. The strength of the correlation indicates how much information is shared across the data sources. Models were fit with and without covariate x_2 to assess how sensitive models were to information on spatial bias in PO data.

In addition to the three integrated models described above, models were also fit to the PA and PO datasets separately for comparison (Equations 1 and 2).

2.5 | Validation

In each set of simulations, predictions were made for an equally spaced sample of 400 locations across the domain. This coarser resolution for prediction is advantageous as it speeds up model fitting. Three metrics were assessed. Firstly, the accuracy was calculated using the mean absolute eError (MAE) between the predicted log intensity and the true log intensity. This metric was chosen because it is less sensitive to outliers than the mean squared error. Smaller errors indicate that the predicted values are close to the truth and that the model has a good fit. Note that the MAE was not calculated relative to mean predictions (as in Simmonds et al., 2020) so reflects the degree to which the absolute intensity can be returned.

The Pearson correlation between predicted log intensity and actual log intensity was also calculated to capture the similarity in predicted spatial distributions. This metric was included because models with high MAE could still capture relative spatial patterns fairly well, which can be assessed by the correlation metric. Positive correlation means that the predicted log intensity increases with the true value, whereas negative correlation means prediction pattern moves in the opposite direction of the truth. The closer the correlation coefficient is to one, the closer the spatial match between predicted and actual values.

Lastly, bias in the estimate of the environmental covariate coefficient was assessed since the true parameter value was known. A mean estimate close to the true value indicates a well-fitting model and small credible intervals of the posterior distribution represents a precise model.

2.6 | Scenarios

We assessed model performance in two simple simulation studies. For the first study (Table 1), the first scenario tested was to study the effect of the sample size of PA dataset on the performance of integrated models. In reality, PA sampling is expensive, sample size is often restricted and the researcher may need to decide how many samples to collect in future surveys. Hence, the simulation study would indicate how sensitive the models are to the sample size and how much effect the PO data have on the integrated model given the size of the PA sample. The number of samples tested was 100 and 200.

The second parameter tested was the degree of bias in the PO sample. To analyse the effect of spatial bias, the detection probabilities that vary horizontally across the field as in Figure 1 were

changed to control the degree of thinning of the point process. Two new sets of probabilities were specified to represent an unbiased situation and a very biased situation. A constant detection probability of 0.2 was set across the whole field dimension to represent the unbiased PO sampling and for the very biased dataset, a set of probabilities with larger range was formed; (0.5, 0.4, 0.1, 0.01, 0.001), where points are more clustered towards the left side of the field compared to the default bias pattern (0.5, 0.3, 0.1, 0.05, 0.01). The detection probabilities for the unbiased and very biased PO data were chosen carefully to avoid too large of a difference in the total number of PO observations between the scenarios to make them comparable.

TABLE 1 Details of the parameters used in the scenarios investigating sample size of PA data and bias in PO data. The probability of observing PO data is given as five values, each associated with a stratum (see Figure 1), forming a gradient from high to low probability. In the Unbiased scenario, the probability of observing PO data is constant

Scenario	Number of PA samples	Probability of observing PO data
Default	100	(0.5, 0.3, 0.1, 0.05, 0.01)
Large sample	200	(0.5, 0.3, 0.1, 0.05, 0.01)
Biased	100	(0.5, 0.4, 0.1, 0.01, 0.001)
Unbiased	100	0.2

A second simulation study was conducted to assess the impact of partial coverage of the domain by one dataset. Due to limited resources, running a research sampling survey sometimes may only cover a limited area while PO sampling often covers a larger proportion of the surrounding area especially where accessibility is high. Hence, one of the objectives of this simulation study was to understand the effects of the size of PA survey area on the integrated model when combined with a PO sample with larger area of study. The second objective was to vary the location of the PA survey in relation to the spatial patterns of bias and environmental covariates.

In data generation, the spatial variation in detection and environmental covariate were assumed to be perpendicular to each other (Figure 1). So, a few aspects were tested that include the overlap of the PA sample with areas where there were large amounts

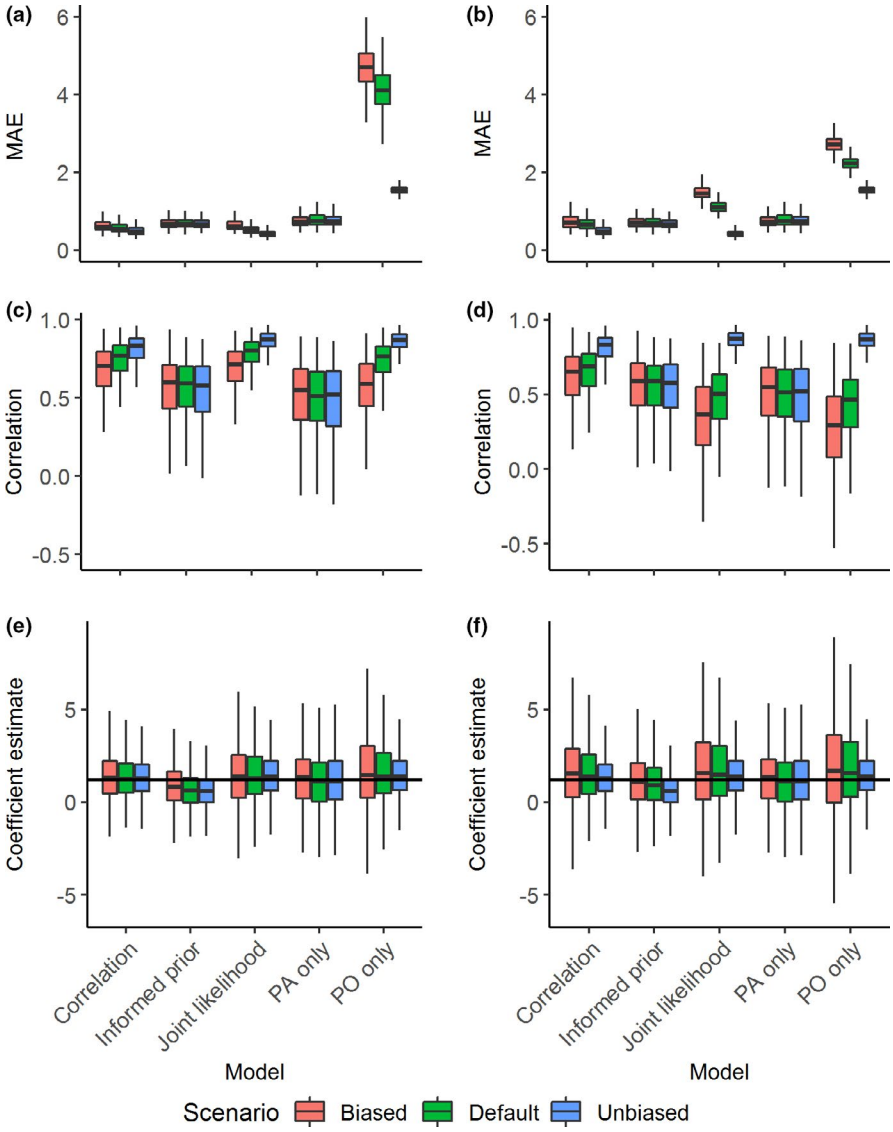

FIGURE 2 Performance of three integrated models and single dataset models in terms of (a) and (b) mean absolute error (MAE), (c) and (d) correlation with true intensity, and (e) and (f) estimation of the environmental coefficient under biased, default and unbiased scenarios. Panels (a), (c) and (e) show models including bias covariate x_2, panels (b), (d) and (f) show models fit without this covariate. The true value of the environmental coefficient is shown by a solid horizontal line in panels (e) and (f)

of opportunistic data (at the left of the area: see Appendix S1 in Supporting Information) against an area where PO data were sparse (right). The overlap of PA data in an area with high values of the environmental covariate (top) or low values (bottom) was also compared. The coverage area was also varied, and the number of PA samples differed according to the size of area. The total number of PA samples for the whole field was set to 250 so that 1/5th of the total area had 50 samples and 3/5ths had 150 samples. The bias in the PO observations was calculated in the same way as the default scenario in Table 1. The models were also fitted using PA data with full coverage area for comparison. All the new scenarios formed for the PA dataset were integrated with the PO data that covered the whole extent and performance was analysed.

For each simulation, all the implemented models were fitted and performance measurements were analysed and compared. Models include the single PA and PO models and the integrated joint likelihood, informed prior and correlation models. Additionally, an extra covariate x_2 was included in the PO, joint, informed prior and correlation models to account for bias in the PO data. Covariate x_2 could be either a variable strongly related to effort (e.g. human population density) or auxiliary information on effort provided alongside PO data. The models fitted without this covariate represent a situation where the data may be spatially biased but there is no information to explain it, that is there is no auxiliary information and other suitable covariates are not available. This is often the situation faced by researchers aiming to use ad hoc observations for modelling where sources of spatial bias can be complex and not always easily approximated by available covariates (Johnston et al., 2020).

All models were implemented in R-INLA (Rue et al., 2009). Code to generate the data and run the simulations is available at https://github.com/NERC-CEH/IDM_comparisons.

3 | RESULTS

Increasing the size of the PA data increased performance, as measured in MAE and correlation, for the PA-only model and all integrated models, with a relatively consistent effect regardless of integrated model type (see Appendix S2).

When spatial bias in PO data was increased (Biased scenario) or decreased (Unbiased scenario), differences in model performance were observed between integrated model types. The degree of difference depended on whether or not a covariate was available to explain bias in the PO data. If a covariate was available (Figure 2a,c,e), then all three integrated model types performed well, with similar or lower MAE than the PA-only model (Figure 2a). Both joint likelihood and correlation models also showed higher correlations than the PA-only model (Figure 2c), indicating the integrated models have the benefits of both low error of the PA data and high coverage of the PO data. The informed prior model provided the least benefit over the PA-only model. High bias in the PO data in the biased scenario reduced the performance in terms of MAE and correlation for joint likelihood, correlation and PO-only models, reflecting the fact that even when the covariate x_2 was available it was not a perfect descriptor of bias.

If a covariate to explain the bias was not available (Figure 2b,d,f), then the degree of bias in PO data had a much larger impact on the joint likelihood model than on the informed prior or correlation models (Figure 2b). The joint likelihood model had lowest MAE of all models in the unbiased scenario and highest in the biased scenario while the informed prior and correlation models were relatively unaffected. However, the informed prior model also showed poor performance in all scenarios in relation to the best performing IDMs, suggesting again that this model provided little improvement over the PA-only model. Surprisingly, the PO-only models had higher error when a bias covariate was available, suggesting that β_2 was poorly estimated when no PA data was available.

All models estimated the environmental coefficient poorly (Figure 2e,f). The informed prior IDM showed an indication of underestimation of the environmental covariate effect in the unbiased scenario regardless of whether x_2 was available, possibly indicating poor estimation of this term in the prior informed by the PO data. In all other models, the mean estimates were unbiased but the range of estimates was very high, suggesting the environmental coefficient was poorly estimated on average.

For the partial coverage simulations, only the results with the smallest coverage of PA data (1/5th of the total area) are shown in Figure 3. Increasing coverage to 3/5th of the area improved performance of all models (see Appendix S3). In all simulations, the PO data covered the entire domain.

As in the first simulation, the results differed depending on how well bias in PO data could be explained. If a covariate to explain bias was available, then adding a small amount of PA data in any part of the domain improved performance over PA-only and PO-only models in terms of MAE (Figure 3a). Adding a small coverage of PA data also improved correlation with the true intensity for joint likelihood and correlation IDMs (Figure 3c).

If bias could not be explained by a covariate, then the placement of the PA data determined how well the joint likelihood model performed in terms of MAE. If the PA data were located so that it covered the gradient of sampling bias in the PO data (in the "top" or "bottom" positions), then the joint likelihood model performed better than the PA-only model (Figure 3b). However, if the PA data did not cover the bias gradient (i.e. was in the "left" or "right" positions) then the joint likelihood model had higher MAE than the PA-only model. The informed prior and correlation models were much less influenced by the placement of the PA-only data. Both informed prior and correlation models outperformed the PA-only model when PA data were spatially restricted, in contrast to the results with full extent of PA data where the informed prior model provided no benefit over modelling PA data alone.

Again, all models estimated the environmental coefficient with high uncertainty(Figure 3e,f), but the correlation and joint likelihood IDMs showed better estimation of the covariate than PA data alone, regardless of placement of the PA data or if a bias covariate was available. The informed prior model underestimated the coefficient

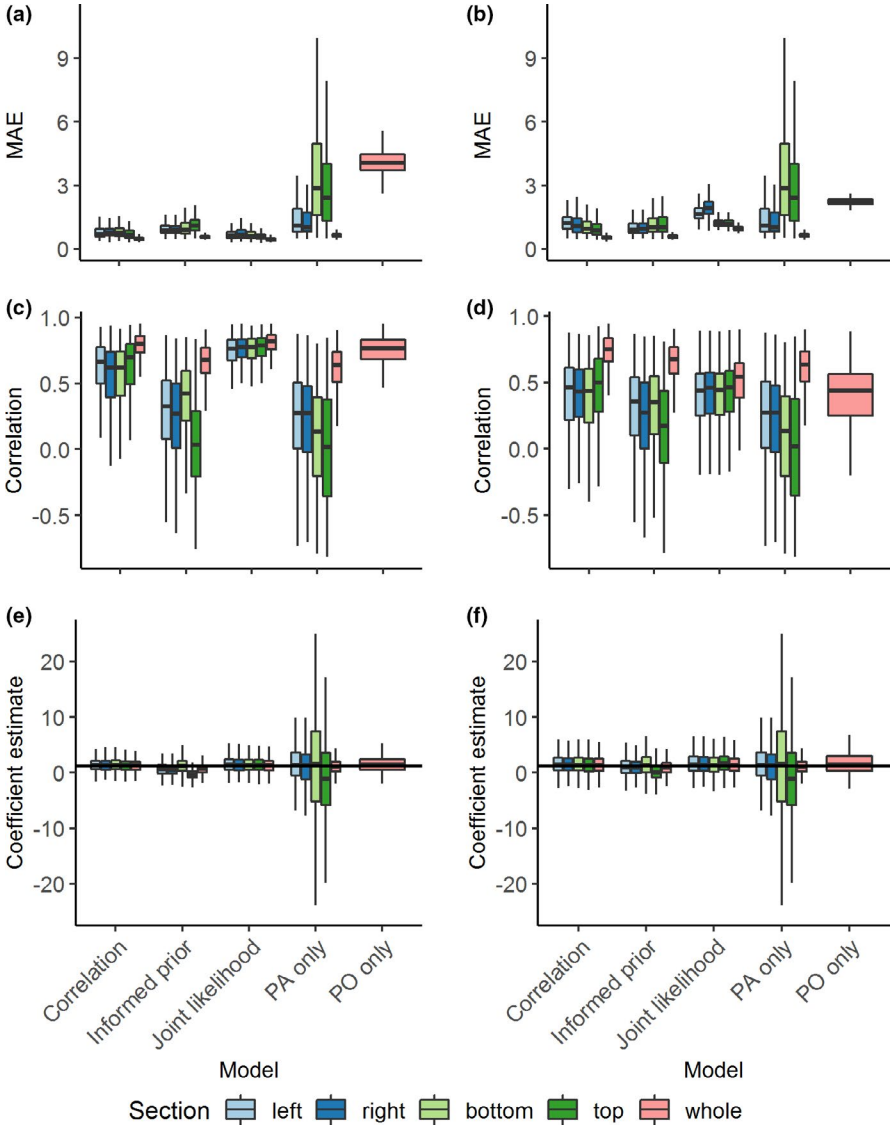

FIGURE 3 Performance of each integrated model and single dataset models when the structured PA data has either partial or full coverage of the total area of interest. Panels (a) and (b) show mean absolute error (MAE), panels (c) and (d) show correlation with true intensity and panels (e) and (f) show estimation of the environmental coefficient. Panels (a), (c) and (e) show results when the bias covariate x_2 was available; panels (b), (d) and (f) show results when the bias covariate x_2 was missing. The true value of the environmental coefficient is shown by a solid horizontal line in panels (e) and (f). A visualization of the PA data placement is available in Appendix S1 in Supporting Information

if either the PA data did not cover the environmental gradient (in the "top" position) or if x_2 was available.

4 | DISCUSSION

This study set out to investigate the performance of three integrated species distribution models under a range of scenarios. In real life, structured or PA sampling is often limited, expensive and often covers a smaller portion of area than opportunistic PO data. The number of samples of PA data, positioning and size of coverage area of PA samples were therefore manipulated to create the limitations in the simulation study. Another variable manipulated was the degree of bias in PO data that often varies in real life caused by different elements like terrain, accessibility and observer density. Models were evaluated by looking at the parameter estimates of the environmental covariate and at prediction performance in terms of the mean absolute error (MAE) and correlation between predicted values and the true values.

The simulation study demonstrated that the joint model did not always perform better than the single PA model unlike the study by Pacifici et al. (2017) whose integrated models always produced better predictions than its single model when the underlying assumption that the two data sources were related was met. Pacifici et al looked at situations where either detection was constant across space or where spatial variation in effort was very well known. Here we considered that detection in PO data almost always varies in space and there may be no information to provide a suitable covariate for spatial variation in detection or effort. We demonstrate that the joint likelihood model performs poorly when the PO data source is biased and that bias cannot be accounted for by a covariate, also shown by Simmonds et al. (2020). The informed prior model was robust to unknown spatial bias in PO data but provided little benefit over the PA-only models in most scenarios. The correlation model was less sensitive to unknown spatial bias in PO data than the joint likelihood model and performed only slightly worse than the joint likelihood model in the unbiased scenario, suggesting this may be a good choice of model for ecologists faced with data of unknown quality.

Both the informed prior and correlation IDMs were less affected by spatially biased PO data, even when no covariate was available to explain the bias. In the informed prior model, the PO data contribute via the prior while in the correlation model the spatial fields are allowed to be correlated but not completely shared. Therefore in both these alternative IDMs, the information contributed by the PO data is lower, reducing the sensitivity of these models to spatial bias in this data source. The informed prior model in particular was unaffected by bias in the PO data; however, this model also provided little benefit beyond modelling PA data only unless the PA data were limited in coverage. This indicates that the prior obtained from the PO data was sufficiently vague to mean that the posterior was largely informed by the PA data. Only when the spatial coverage of the PA data was limited did the benefit of this IDM become apparent. Fletcher et al. (2019) argued that incorporating PO data via informed priors would be similar to joint likelihood modelling. However, our results indicated that the informed prior model provided little improvement over analysing the PA data alone when the full domain was covered. The informed prior model also consistently underestimated the environmental coefficient despite both single data source models producing unbiased estimates.

In agreement with the research done by Koshkina et al. (2017), the performance of integrated models with PA data restricted to a small part of the total area of interest can be higher than using PO data only, suggesting that spatially restricted PA data can still be valuable in estimating species distributions. However, joint likelihood models were influenced by the location of the PA data in relation to gradients of bias. If the PA data did not cover the gradient of bias of the PO data, then joint likelihood models produced poorer results than if the PA data alone had been used unless a covariate was available to explain bias in the PO data. Therefore for PA data to be useful in separating bias from the true spatial distribution, it must cover both areas with high sampling effort or detectability and areas with low effort or detectability. It is likely to be impossible for a researcher to be able to estimate whether PA data cover this gradient unless good covariates are available, in which case these can be included in the model anyway, reducing the utility of the joint likelihood model in this scenario. Surprisingly, the correlation model had comparable or lower error to the joint likelihood model, even though there was a restricted spatial area over which to correlate spatial fields. This suggests that even when PO data are spatially biased, the cause of this bias is unknown and unbiased PA data are only available for a small subset of the domain, data integration via the correlation model can still provide better estimates of species distributions than considering either dataset separately.

One important assumption of our models is that detection in PA data is perfect whereas Pacifici et al. allowed for both datasets to have imperfect detection. Therefore, our results may be more applicable to plants or other relatively immobile taxa where the assumption of perfect detection is more likely to be reasonable (e.g. Fithian et al., 2015). However, as long as imperfect detection in PA data is not spatially biased, we would hypothesize that the relative performance of the different IDMs would be similar. We also assumed that PA data locations were spatially unbiased, which corresponds to the PA designs we most often use (e.g. Norton et al., 2012) but may not be the case for other PA datasets.

Another limitation is the simplistic way in which the environmental covariate and spatial bias were assumed to have perpendicular gradients, so that their effects could be easily separated. If these variables are correlated, then integrated models may perform poorly (Simmonds et al., 2020), although the relative performance of IDM types under scenarios of correlation between environmental suitability and sampling bias has not yet been investigated. Simmonds et al. suggested that models containing two spatial fields, one capturing the spatial bias, could be useful where there is correlation between drivers of occurrence and sampling bias. The robustness of each type of IDM to other potential patterns in sampling such as preferential sampling in areas of high occupancy is also useful topics for future work.

Overall, the study confirms that joint likelihood models provide the best estimates when data are unbiased, or bias is well accounted for, but are very sensitive to unexplained spatial biases. The two alternative IDMs were robust to unknown spatial bias; however, the informed prior model showed little improvement over modelling the PA data alone unless the PA data were spatially constrained. The correlation model performed well under conditions of unexplained bias and provided an improvement over modelling single data sources. We suggest that the correlation model may be the best choice in many applications where researchers are faced with spatially biased PO data. The cost of using this model when data are unbiased, or when effort can be explained by a covariate, is low, with only a small reduction in performance compared to the joint likelihood model. The correlation model also performs well when PA data are spatially restricted. Understanding the complex spatial biases in PO data is a real challenge for researchers so providing integrated modelling approaches that are more robust to unknown biases is important to allow researchers to apply integrated approaches to a wider range of datasets.

ACKNOWLEDGEMENTS

The work presented here formed part of a Master's thesis by SSAS as part of an MSc Data Science at Lancaster University. We thank the UK Centre for Ecology & Hydrology for hosting SSAS during this project and providing facilities and support. We also thank Pete Henrys and Emily Simmonds for statistical advice. Additional funding for SJ was provided under the Natural Environment Research Council UK-SCaPE programme delivering National Capability award NE/R016429/1, and for GSB under his EPSRC Senior Fellowship grant (EP/P002285/1).

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/ddi.13255.

DATA AVAILABILITY STATEMENT

All R code required to conduct the simulation study and create all figures and tables is publically available on Github at https://github.com/NERC-CEH/IDM_comparisons.

ORCID

Susan G. Jarvis https://orcid.org/0000-0001-5382-5135

REFERENCES

Blangiardo, M., Cameletti, M., Baio, G., & Rue, H. (2013). Spatial and spatio-temporal models with R-INLA. *Spatial and Spatio-Temporal Epidemiology*, *4*, 33–49. https://doi.org/10.1016/j.sste.2012.12.001

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*, 7766. https://doi.org/10.1038/s41598-019-44075-9

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*, 424–438. https://doi.org/10.1111/2041-210X.12242

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*, e02710. https://doi.org/10.1002/ecy.2710

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 and Evolution*, *35*, 56–67. https://doi.org/10.1016/j.tree.2019.08.006

Isaac, N. J. B., & Pocock, M. J. O. (2015). Bias and information in biological records. *Biological Journal of the Linnean Society*, *115*, 522–531. https://doi.org/10.1111/bij.12532

Isaac, N. J. B., van Strien, A. J., August, T. A., de Zeeuw, M. P., & Roy, D. B. (2014). Statistics for citizen science: Extracting signals of change from noisy ecological data. *Methods in Ecology and Evolution*, *5*, 1052–1060. https://doi.org/10.1111/2041-210X.12254

Johnston, A., Moran, N., Musgrove, A., Fink, D., & Baillie, S. R. (2020). Estimating species distributions from spatially biased citizen science data. *Ecological Modelling*, *422*, 108927. https://doi.org/10.1016/j.ecolmodel.2019.108927

Kery, M., & Royle, J. A. (2016). *Applied hierarchical modeling in ecology: Analysis of distribution, abundance and species richness in R and BUGS*. Academic Press.

Koshkina, V., Wang, Y., Gordon, A., Dorazio, R. M., White, M., & Stone, L. (2017). Integrated species distribution models: Combining presence-background data and site-occupancy data with imperfect detection. *Methods in Ecology and Evolution*, *8*, 420–430. https://doi.org/10.1111/2041-210X.12738

Lindgren, F., Rue, H., & Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*, 423–498. https://doi.org/10.1111/j.1467-9868.2011.00777.x

Miller, D. A. W., Pacifici, K., Sanderlin, J. S., & Reich, B. J. (2019). The recent past and promising future for data integration methods to estimate species' distributions. *Methods in Ecology and Evolution*, *10*, 22–37. https://doi.org/10.1111/2041-210X.13110

Norton, L. R., Maskell, L. C., Smart, S. S., Dunbar, M. J., Emmett, B. A., Carey, P. D., Williams, P., Crowe, A., Chandler, K., Scott, W. A., & Wood, C. M. (2012). Measuring stock and change in the GB countryside for policy: Key findings and developments from the Countryside Survey 2007 field survey. *Journal of Environmental Management*, *113*, 117–127. https://doi.org/10.1016/j.jenvman.2012.07.030

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*, 840–850. https://doi.org/10.1002/ecy.1710

Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *71*, 319–392. https://doi.org/10.1111/j.1467-9868.2008.00700.x

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*, 1413–1422. https://doi.org/10.1111/ecog.05146

Simpson, D., Illian, J. B., Lindgren, F., Sørbye, S. H., & Rue, H. (2016). Going off grid: Computationally efficient inference for log-Gaussian Cox processes. *Biometrika*, *103*, 49–70. https://doi.org/10.1093/biomet/asv064

Zipkin, E. F., Inouye, B. D., & Beissinger, S. R. (2019). Innovations in data integration for modeling populations. *Ecology*, *100*, e02713. https://doi.org/10.1002/ecy.2713

Zurell, D., Berger, U., Cabral, J. S., Jeltsch, F., Meynard, C. N., Münkemüller, T., Nehrbass, N., Pagel, J., Reineking, B., Schröder, B., & Grimm, V. (2010). The virtual ecologist approach: Simulating data and observers. *Oikos*, *119*, 622–635. https://doi.org/10.1111/j.1600-0706.2009.18284.x

BIOSKETCHES

Siti Sarah Ahmad Suhaimi is a data science Masters graduate from Lancaster University with interest in data modelling. She works as a data engineer in a global manufacturing organization paving her way to be an established data scientist.

Gordon Blair is a Distinguished Professor in the School of Computing and Communications, Lancaster University. He holds a prestigious EPSRC Senior Fellowship in Digital Technology and Living with Environmental Change and is co-Director of the Centre of Excellence in Environmental Data Science, a joint initiative involving Lancaster University and UKCEH.

Susan Jarvis is a quantitative ecologist interested in combining different data types, such as citizen science and professional surveys, to understand patterns in biodiversity across space and time. See her Google Scholar page https://scholar.google.co.uk/citations?user=DuIhO1IAAAAJ&hl=en for recent publications.

SUPPORTING INFORMATION

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

How to cite this article: Ahmad Suhaimi SS, Blair GS, Jarvis SG. Integrated species distribution models: A comparison of approaches under different data quality scenarios. *Divers Distrib*. 2021;27:1066–1075. https://doi.org/10.1111/ddi.13255