Received: 21 October 2020 | Revised: 3 June 2021 | Accepted: 27 August 2021

DOI: 10.1111/jbi.14268

Journal of Biogeography WILEY

METHODS

Solving sampling bias problems in presence–absence or presence-only species data using zero-inflated models

Victoria Nolan | Francis Gilbert | Tom Reader

Life Sciences Building, University of Nottingham, Nottingham, UK

Correspondence
Victoria Nolan, Life Sciences Building, University of Nottingham, Nottingham NG7 2TQ, UK.
Email: victorianolan888@gmail.com

Funding information
University of Nottingham; Woodland Trust

Handling Editor: Enrique Martínez-Meyer

Abstract

Aim: Large databases of species records such as those generated through citizen science projects, archives or museum collections are being used with increasing frequency in species distribution modelling (SDM) for conservation and land management. Despite the broad spatial and temporal coverage of the data, its application is often limited by the issue of sampling bias and consequently, zero inflation; there are more zeros (which are potentially 'false absences') in the data than expected. Here, we demonstrate how pooling species presence data into a 'pseudo-abundance' count can allow identification and removal of sampling bias through the use of zero-inflated (ZI) models, and thus solve a common SDM problem.

Location: All locations

Taxon: All taxa

Methods: We present the results of a series of simulations based on hypothetical ecological scenarios of data collection using random and non-random sampling strategies. Our simulations assume that the locations of occurrence records are known at a high spatial resolution, but that the absence of occurrence records may reflect undersampling. To simulate pooling of presence–absence or presence-only data, we count occurrence records at intermediate and coarse spatial resolutions, and use ZI models to predict the counts (species abundance per grid cell) from environmental layers.

Results: ZI models can successfully identify predictors of bias in species data and produce abundance prediction maps that are free from that bias. This phenomenon holds across multiple spatial scales, thereby presenting an advantage over presence-only SDM methods such as binomial GLMs or MaxEnt, where information about species density is lost, and model performance declines at coarser scales.

Main Conclusions: Our results highlight the value of converting presence–absence or presence-only species data to 'pseudo-abundance' and using ZI models to address the problem of sampling bias. This method has huge potential for ecological researchers when using large species datasets for research and conservation.

KEYWORDS
conservation, presence–absence, presence-only, sampling bias, species abundance, species distribution model, species occurrence, zero inflation

1 | INTRODUCTION

Species distribution modelling (SDM) is widely used to address important ecological questions about species distributions and the environment (Elith et al., 2011; Dormann et al., 2007; Phillips et al., 2009). Species occurrence or abundance data from large, observational datasets such as citizen science projects, museum or herbarium collections and record lists are increasingly being used in SDM (Pearce & Boyce, 2006; Schmeller et al., 2009; Tiago, Pereira, et al., 2017). The extensive spatial and temporal coverage of the data, as well as the growing ease of online access provide numerous benefits over often costly and labour-intensive sampling methods employed in more focused scientific studies of distribution (Dickinson et al., 2010; Dwyer et al., 2016; Gouraguine et al., 2019). Nevertheless, although some collections of species records can be generated using hypothesis-led, systematic sampling protocols (Pocock & Evans, 2014; Schmeller et al., 2009), much of these data comprise presence-only occurrence records, where there is often little information about the source or survey effort accompanying the records (Boakes et al., 2010; Rocchini et al., 2011). As a result, sampling bias (also called sample selection or survey bias) is often present—certain temporal periods, geographical areas or taxa are sampled more intensively or frequently than others (Bird et al., 2014; Dickinson et al., 2010; Phillips et al., 2009).

Sampling bias in SDM can lead to over- or underestimation of important species–environment relationships (Syfert et al., 2013), and predicted distribution maps may partly represent survey effort rather than species niche requirements (Mair & Ruete, 2016; Phillips et al., 2009). Proposed methods to correct for sampling bias generally rely on either spatial filtering of occurrence records, or the manipulation of background data ('pseudo-absences') (Boria et al., 2014; Fourcade et al., 2014; Kramer-Schadt et al., 2013; Phillips et al., 2009). Both of these techniques have limitations: the former results in a dataset of reduced sample size and statistical power (Wisz et al., 2008), whereas the latter usually requires some prior knowledge of the source of the bias (Dudík et al., 2005; Phillips, 2008). A third option is the use of statistical models that can account for some of the causes of sampling bias (Bird et al., 2014; Isaac et al., 2014), for example geographically weighted regression (GWR) (Brunsdon et al., 1998), or maximum entropy (MaxEnt) with a bias layer, although again, most of these require prior knowledge of the source of the bias.

One specific problem relating to sampling bias that is particularly noticeable in species abundance databases is zero inflation: the presence of more recorded zeros or locations where data are absent than expected under standard distributions (binomial, Poisson, negative binomial etc.) (Martin et al., 2005). These excess zeros can arise from multiple processes. Some are considered to be 'true zeros', which result from either ecological processes that render a site unsuitable for occupancy, or stochastic processes, such as a sudden random extinction event in an otherwise suitable location (Cunningham & Lindenmayer, 2005; Martin et al., 2005). In contrast, 'false zeros' are locations where a species occurs but was not recorded because of errors or omissions in the sampling method (Dénes et al., 2015). These errors are either systematic and occur repeatedly throughout the survey process (e.g. through a lack of detection or poor survey design), or are owing to sampling bias, because some geographical areas have not been sampled at all (Bird et al., 2014).

Generalised Linear Models (GLMs) are a common method for analysing relationships between species occurrences or abundance and environmental variables, but excess zeros are problematic for GLMs, and if unaccounted for, can result in biased parameter estimates and poor predictive power (Lambert, 1992). As a possible solution to this problem, zero-inflated (ZI) models and their components (extensions of GLMs) have been widely discussed in the literature (Lambert, 1992; Welsh et al., 1996; Zuur et al., 2009). ZI models consist of two parts, namely a logistic component that models the probability of an observation being an excess zero (hereafter called the 'zero component') and a 'count component' that models a count (e.g. species abundance) under an assumed distribution (Lambert, 1992). Both components of ZI models are capable of producing zeros, and a key feature is the ability to include different predictor combinations in each component. In other words, they can model the different sources of zeros independently (Wenger & Freeman, 2008; Zuur et al., 2009).

ZI models, which require counts of occurrences (i.e. abundance), are rarely considered in SDM, because most large datasets record species presences, not abundance. SDM methods that can use presence-only data, such as MaxEnt, are therefore most commonly applied (Fitzpatrick et al., 2013; Fourcade et al., 2014; Phillips & Dudík, 2008). However, the ability of ZI models to separate the two processes underlying the generation of zeros in a species dataset could provide an alternative method to model and account for sampling bias. ZI models can be used with any species database that records abundance directly, or by aggregating presence-only or presence–absence data into counts of occurrence. In this study, we therefore propose ZI models as a new, alternative method to address problems of sampling bias in SDM. We present here the results of a series of simulations, based on hypothetical ecological scenarios representing the large-scale collection of species occurrence data, that aim to address three particular research questions.

Our first research question is to test our main theory of whether sampling bias (resulting in excess 'false' zeros) can be modelled and accounted for using ZI models, in order to improve species distribution predictions. ZI models have been used effectively to model true and false zeros in ecological count data, such as when modelling the abundance of rare species (Cunningham & Lindenmayer, 2005; Martin et al., 2005; Welsh et al., 1996). They are also particularly prevalent in the field of occupancy-abundance modelling (Sileshi et al., 2009; Smith et al., 2012), especially when there are false zeros in the data owing to systematic sampling errors from imperfect detection (Sólymos et al., 2012; Wenger & Freeman, 2008; Williams et al., 2016). However, research into zero inflation caused by spatio-temporal sampling bias in species occurrence data is scarce. A few studies have used ZI models to identify and quantify sources of bias in species data (Dwyer et al., 2016; Tiago, Ceia-Hasse, et al., 2017;

Williams et al., 2016), yet none has tested the ability of the models to produce accurate predictions of species distributions from biased data. We outline through our simulations how accurate distribution maps can be produced using ZI models in this way, and we describe the required criteria during model fitting and prediction for this to occur. In particular, our simulations also address our second research question: under what levels of zero inflation is our ZI model method most appropriate?

Our final research question considers the issue of scale, and the benefits of pooling fine-scale occurrence data to model occurrence density across coarser spatial scales. Species presence is normally modelled at the smallest spatial scale (grid cell size) possible, given the resolution of the records and environmental layers used to build the model. Counting or aggregating presences across grid cells at a larger spatial scale to generate 'abundance' data intuitively seems to be a bad idea, because it throws away information about the precise location of the records. However, this may be inevitable if predictor layers have lower spatial resolution than occurrence location data, and we propose here that it may actually present considerable advantages. Aggregated counts of occurrences are commonly not a direct measure of true abundance (the total number of individuals of the target species), since each raw occurrence often represents a locality which is home to several or many individuals. Regardless, modelling 'abundance', and any zero inflation therein, may give important clues to sources of bias in the data which are not obvious in the raw occurrences, and the benefits of being able to identify and eliminate bias could outweigh the costs of any loss of spatial resolution caused by aggregation. Therefore, counting occurrence records at larger spatial scales in order to model 'occurrence density' may be a better alternative to traditional presence-only SDM methods. Indeed, abundance models have been shown to perform better than presence–absence models fitted using the same data across multiple spatial scales (Howard et al., 2014; Johnston et al., 2015).

Other methods do exist that propose aggregating occurrences into counts of 'abundance' that may also provide advantages when using spatially biased species data, including Poisson point models (Komori et al., 2020; Renner et al., 2015). These models can incorporate bias predictors when modelling intensity rather than occurrences across the study area. Nevertheless, they still require a priori knowledge about potential bias predictors, whereas we show here that ZI models are able to provide an indication of potential sources of sampling bias in the data when the exact sources are unknown.

We do not attempt to provide a detailed statistical summary of ZI models and theory (there is much associated literature already available), but aim to draw attention to the main modelling methods and usefulness of ZI models for ecological researchers and species distribution modellers dealing with large, biased databases. We argue that ZI models can provide insight into, and correction methods for, the bias in large species databases, and that they can be powerful and effective SDM tools.

2 | MATERIALS AND METHODS

Our general approach was to use ZI models to predict the observed number of species occurrences per grid cell for a series of simulated species using predictors of either the biology of the species and/or sampling bias in the data. We envisaged a large species for which it is theoretically possible to survey all individuals in a landscape (e.g. trees, large animals). The true distribution of all individuals was simulated for each species, and this distribution was then sampled incompletely, with or without spatial sampling bias. Before sampling, the true abundance of the species could be calculated by summing occurrences per grid square. But with incomplete sampling, the observed or 'sampling abundance' per grid cell is an underestimate. An alternative way to view our simulations, which is more realistic for species which are small or hard to enumerate (e.g. smaller plants, most insects), is to consider each occurrence in the raw data to represent a recorded encounter with the species at a local site which may contain many individuals. In such cases, the models do not strictly predict abundance, but instead they predict what we might call 'occurrence density'.

As a result of the two-part nature of ZI models, two types of abundance predictions can be produced. Assuming that all excess zeros arise from incomplete sampling, the first type of prediction is of true, biological abundance (or occurrence density) across the study area, created only from the count component of the model, which we call here the 'count abundance prediction'. This is likely to be the desired modelling outcome, especially for conservation and land management planning. The second type of prediction, which we here call the 'sampling abundance prediction', comes from the whole model (combining both the count and zero components) and therefore represents the predicted abundance (or occurrence density) that would be recorded if sampling was carried out in the same way as when collecting the data that were used to fit the model. Bias in sampling will be reflected in this second prediction. However, if some excess zeros arise also from biological zero inflation, for example if a species is clustered, the zero component will reflect some of the underlying biological processes as well as the sampling bias. In this case, the count abundance prediction will only partially reflect the true species abundance. The best type of prediction to use will therefore depend on the estimated strength of biological zero inflation versus the bias in the data.

2.1 | Simulation study area and predictor variables

We simulated the occurrence of a hypothetical species in a study area that consisted of a 100 × 100 cell grid at 1-km^2 resolution placed randomly within the boundary of England (Figure 1a). The total area covered by the grid is therefore 10,000 km^2 and there are 10,000 individual grid cells. Two predictor variables were selected across this area. The first was a 'biological predictor' that we chose to be 'altitude', which we used to define the relationship between the simulated species occurrences and environment (Meynard et al.,

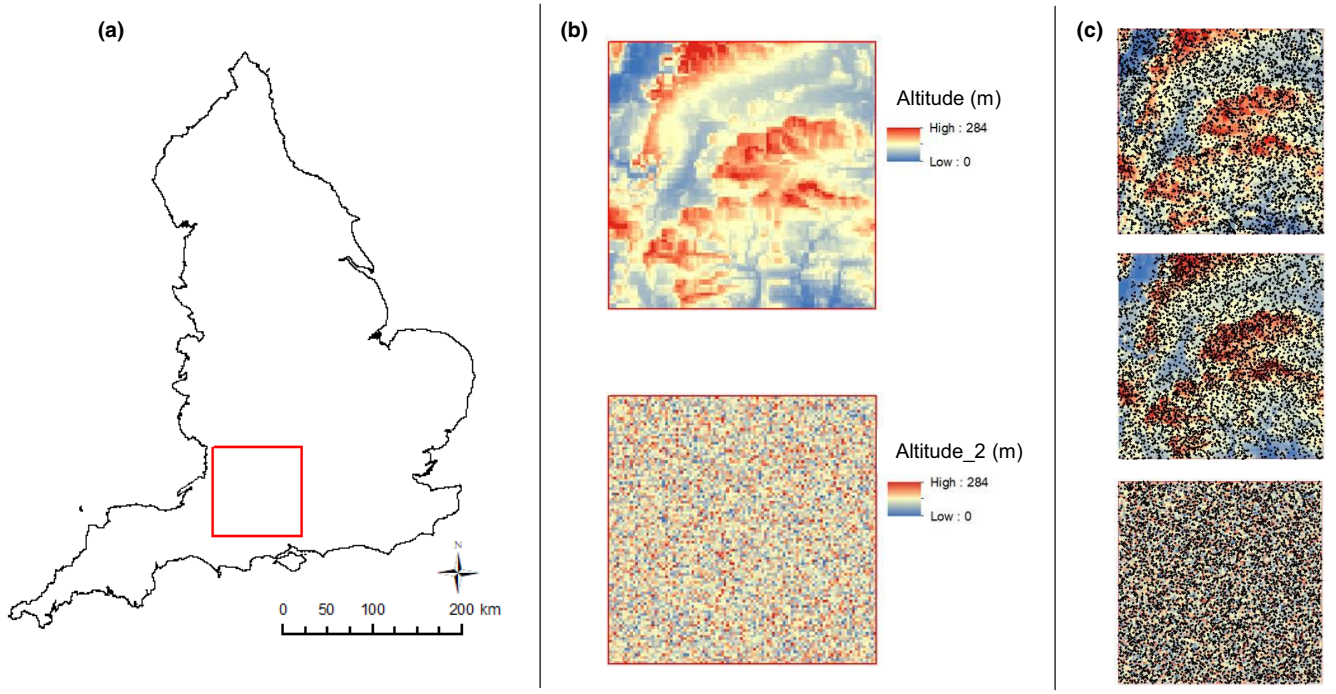

FIGURE 1 (a) Simulation study area consisting of a group of 100 × 100 grid squares of 1-km^2 size randomly placed within England covering a total area of 10,000 km^2 (outlined in red) (left). (b) The biological predictors used to fit the models: altitude (m) (top) and altitude_randomised (m) (randomised altitude layer with no spatial autocorrelation, labelled here as 'altitude_2') (bottom) shown for the study area. (c) A simulated species with 5000 occurrence points showing no preference for altitude (random species) (top), a preference for high altitudes based on a logarithmic scaler of altitude (altitude species) (middle) and a preference for high altitudes based on a logarithmic scaler of altitude_randomised (altitude_randomised species) (bottom). Although the occurrence points in this bottom map appear randomly distributed across the study area (similar to that of the 'random species' in the top map), it is actually only the altitude values that are randomly distributed (in the 'altitude randomised' layer): species occurrences are still placed based on preferring high-altitude values, so are not actually random with respect to the environmental predictor

2019). Real values for altitude (m) across the study area were obtained from WorldClim DEM (accessed 10/05/18) at a 1-km^2 resolution and ranged from 0 to 284 m above sea level (Figure 1b). The choice of biological predictor for a simulation study of this sort is necessarily somewhat arbitrary, but we chose altitude because it is both a plausible predictor of occurrence for a range of organisms, and it is quite strongly spatially autocorrelated, an important possible source of biological zero inflation in the abundance data formed when occurrences are counted across grid cells at intermediate spatial scales. The actual biological mechanism underlying the relationship between altitude and species occurrences is not important for this study, but altitude is a good proxy for a suite of environmental variables such as temperature or precipitation commonly used in SDM which have direct effects on species distributions.

Because altitude is spatially autocorrelated, and so is the sampling bias we wanted to investigate (see below), there was a risk that biological and sampling bias predictors in our simulations could correlate: depending on the positions of the simulated towns on our map, there could be a strong correlation between real altitude and sampling effort. Thus, in order to allow us to investigate the impact of sampling bias completely independently of the biological predictor, we also generated an alternative 'biological predictor' with no autocorrelation: a spatially random control variable. This control variable (henceforth labelled 'altitude_randomised') was created by randomising the real altitude values across the study area at a 1-km^2 resolution (Figure 1b), and hence removed any correlation between altitude and distance from town.

The second predictor of observed species occurrence was a 'bias predictor' ('distance from nearest town') which affected the virtual sampling of the simulated species. We assumed that the greater the distance from a town, the lower the feasibility and likelihood of sampling occurring, as has previously been seen in ecological studies (Kadmon et al., 2004; Parnell et al., 2003; Reddy & Dávalos, 2003). Unlike with altitude, we chose to simulate a hypothetical bias layer rather than use values based on the locations of real towns, in order to ensure the lowest possible correlation between the two predictors, although some correlation between them was likely because of spatial autocorrelation in both. Within the study area, 10 points representing 'town centres' were randomly placed, and the distance from the nearest town (m) was calculated for each grid cell, creating a continuous predictor layer at 1-km^2 resolution across the study area. To reduce the influence of collinearity between predictors, the process of generating the 'town centres' was repeated 10 times, creating 10 sets of randomly placed 'town centres' (Appendix S1, Figure S1.1). As a result, mean Pearson's correlation coefficients across the 10 repetitions show weak correlations between the bias predictor

TABLE 1 Sources of zero inflation in the simulated species occurrence data

	Source of zero inflation		
	True abundance (before sampling)	Random sampling	Biased sampling
Species			
Random	No zero inflation	Sampling	Sampling
Altitude	Biological	Biological and sampling	Biological and sampling
Altitude 2	Biological	Biological and sampling	Biological and sampling

'distance from nearest town' and the biological predictor 'altitude' (r = −0.0499, SD ±0.228), and even weaker correlations with the biological predictor 'altitude_randomised' (−0.0044, SD ±0.012).

To summarise, we had three variables in total across the simulation study area: two biological predictors ('altitude' and 'altitude_randomised') and one bias predictor ('distance from nearest town'). All predictors were centred (the mean of each predictor was subtracted from each value of the predictor) and scaled (the centred values were divided by the standard deviation of the predictor values) so that the differences in units of the predictors were removed.

2.2 | Simulating the virtual species

To obtain counts of 'abundance' to use in ZI models, we first simulated species occurrences across the study area and then aggregated them into counts of 'abundance' (alternatively interpreted as occurrence density—see above), because we assumed that the simulated distribution of occurrences was the complete true distribution, all other locations are assumed to be 'true absences'. Therefore, when aggregating the raw occurrence points into 'abundance' counts, a value of 0 represented a true absence and any value greater than 0 represented a true presence.

The recommended first step in a simulation study is to define the relationship between the environment and occurrence points (Meynard et al., 2019). We modelled the distributions of three simulated species each with 5000 occurrence points (Figure 1c). The occurrence points of the first species ('random species') were simulated randomly across the study area, and show no preference for any environmental condition. The second and third species were simulated based on the two biological predictors ('altitude' and 'altitude_randomised') and were assumed to favour high altitudes; these species were named 'altitude species' and 'altitude_randomised species' respectively. We chose these three scenarios in order to create datasets in which different kinds of zero inflation occur. For the random species, zero inflation can only occur as a result of sampling (where sites which are not sampled might be incorrectly recorded as zeros), while for the altitude species and altitude_randomised species, zero inflation can result both from sampling and from the fact that grid cells are potentially not suitable for the species because of environmental conditions.

We then simulated the effect of the relationship between our biological predictors and species occurrences by creating layers of the probability of occurrence which varied according to altitude or altitude_randomised (see Meynard & Kaplan, 2013; Meynard et al., 2019). Initially we tried using a linear relationship between the altitude predictor layers and probability of occurrence, but this introduced relatively little zero inflation in the data. For the purposes of investigating sampling bias and zero inflation, we therefore chose to use a logarithmic relationship, whereby the probability of occurrence rapidly increases initially with small increases in altitude, but gradually tapers off at higher altitudes. This heavily disfavours low-altitude values, and the majority of these will be assigned low probability values close to zero. Hence, biological aggregation of the occurrence points was effectively increased, yielding greater zero inflation. Each biological predictor was resampled to a 100 × 100 m resolution across the study area, and were then rescaled using the 'rescale by function' tool in ARCGIS version 10.3.1 (ESRI, 2013), such that the new probability of occurrence layers (ranging between 0 and 1) was logarithmically related to the biological predictors.

Five thousand occurrence points were placed across the study area (using the ArcGIS tool: 'Create Spatially Balanced Points') based on these altitude and altitude_randomised occurrence probability layers. Due to computation limitations of the 'Create Spatially Balanced Points' tool, only one occurrence point can be placed within a single raster cell. Therefore, a resolution of 100 × 100 m was chosen for the probability layers so that up to 100 species occurrences could be placed in each 1-km^2 grid cell. Although visually the altitude_randomised species appears to be randomly distributed across the study area, it is actually the underlying altitude grid square values that are randomised: occurrences of the altitude_randomised species still occur at higher densities in grid squares with higher altitude values. As we used a logarithmic species response to the altitude_randomised layer, significant (biological) zero inflation still occurs in the raw data: occurrences are unlikely in low-altitude grid cells, generating lots of true zeros when occurrences were counted per grid cell (Table 1). Only the random species distribution is completely random across the study area.

Finally, true (raw) species abundance (total number of occurrence points) was calculated for each 1-km^2 grid cell. We felt the chosen grid scale was appropriate because, although the maximum abundance per grid cell is strictly 100, no grid cells reached this value (the maximum was six occurrences per 1-km grid cell), and we therefore assumed that it was unlikely that the shape of the distribution of abundances would be significantly affected by the upper bound (i.e. unbounded distributions such as Poisson or negative binomial

were likely to be appropriate). In addition, using this grid scale sets up a situation where location data are available at a higher resolution than the environmental predictors. Hence, we are simulating a situation in which modellers must make a decision about how to aggregate high-resolution data across grid cells to create models which predict species distributions based on lower resolution environmental predictors.

2.3 | Simulating the sampling strategies

We considered two sampling strategies across the study area to represent alternative scenarios of ecological data collection. The first is random sampling, where every 1-km grid cell has an equal chance of being visited and sampled. If visited, we assume all species occurrences in the cell are recorded (i.e. there is no detection error) and the result is the true (raw) abundance (count of all occurrences) for each visited grid cell. The second sampling strategy is affected by spatial sampling bias and relates to the 'bias predictor', where the probability of a grid cell being sampled decreases as distance from the nearest 'town centre' increases. The grid cells selected for this strategy were chosen based on a probability layer created using a logarithmic scaler of the 'distance from nearest town' predictor, again using the 'rescale by function' ArcGIS tool. This time high probability values close to 1 were assigned to cells with small numerical values, that is cells closer to towns and more likely to be sampled, whereas low probability values close to 0 were assigned to cells with large 'distance from nearest town' values. For each strategy, 2000 grid cells (20% of the total) were sampled and species abundance was noted for each one. All other (unsampled) squares were assigned an observed abundance of zero, creating a ZI dataset. All sources of zero inflation in the simulated species abundance data before and after sampling are shown in Table 1.

2.3.1 | Simulation 1: investigating the accuracy of species distribution maps from ZI models

To address our first question regarding the accuracy of ZI model predictions of abundance, we focused initially on the performance of ZI Poisson models, and how this compared with equivalent conventional Poisson GLMs. We include comparisons between (a) ZI and GLM models, (b) count and sampling abundance predictions from ZI models and (c) alternative ZI models fitted using different combinations of biological and bias predictors.

We chose to fit four GLMs and six ZI models for each of the three sets of species abundances per 1-km^2 (random, altitude and altitude_randomised), all fitted with a Poisson distribution but with different combinations of the biological or bias predictors (Table 2). These included combinations where different predictors were tested in the count and zero components of the ZI models. Where the biological predictor was included, models for the 'altitude species' were fitted using altitude as a predictor, and models for the altitude_randomised species were fitted using altitude_randomised. Model fitting was repeated 10 times, each time using a different set of simulated 'town centres' (Appendix S1, Figure S1.1). Thus, there are three species (random, altitude and altitude_randomised), two sampling strategies (random and biased) and 10 repeats, resulting in 60 total simulation runs. All ZI and GLM models were fitted in R version 3.6.3 (R Core Team, 2019) using packages 'stats' (R Core Team, 2019) and 'pscl' (Zeileis et al., 2008).

TABLE 2 Ten predictor combinations were considered when modelling the simulated species distributions

Model	Predictors (GLM/ ZI count component)	Predictors (ZI zero component)
GLM1	Null (No predictors)	N/A
GLM2	Biased	N/A
GLM3	Biological	N/A
GLM4	Biological + bias	N/A
ZI1	Null (No predictors)	Null
ZI2	Biological + bias	Biological
ZI3	Biological	Biological + bias
ZI4	Biological	Biological
ZI5	Bias	Bias
ZI6	Biological + bias	Biological + bias

Four Generalised Linear Model (GLM) and six zero-inflated (ZI) model structures were considered using combinations of the biological predictors (either altitude or altitude_randomised) and the bias predictor (distance from nearest town), including different combinations in the count and zero components of the ZI models.

Abundance predictions from each model were created using 10-fold cross-validation, where the data were split into 10 subsets and each subset was used iteratively as the test data for which predictions were created and the other nine subsets as training data. For the ZI models, both count abundance and sampling abundance predictions were evaluated. Model predictions were evaluated using a novel metric based on the probability of obtaining the model predictions, that we named 'deviation from the best model' (*D*) (See Appendix S3 for more information). We used this metric, rather than conventional measures of performance (e.g. root mean square) typically employed in presence-only or presence–absence modelling, because it produces a measure of fit for count or abundance predictions which is independent of the mean. *D* ranges from a minimum of 1 for a perfect model where model predictions are equal to the true raw abundance data, and increases without limit as model predictive performance decreases. Spearman's rank correlation coefficients (r_s) were also used to compare model abundance predictions to the original model covariates.

To check that our results were not overly sensitive to the choice of predictor, simulations using average temperature (°C) (WorldClim, accessed 10/05/18) at a 1-km^2 resolution, as an alternative biological predictor, were also carried following the same methodology (see Appendix S2)—the results were parallel to those of altitude, and so were omitted from the main results and discussion.

2.3.2 | Simulation 2: examining the impact of the extent of zero inflation in the data

To address our second question, about the effect of varying the extent of zero inflation in the data (both as a result of biological processes and sampling bias) on the effectiveness of the ZI models, we carried out a second simulation. In our first simulation, we assumed 20% of grid cells were sampled, but in Simulation 2 zero inflation resulting from sampling bias was adjusted by varying the number of cells sampled from the grid, ranging from 1000 (10%) to 10,000 (100%) at 10% increments. Therefore, the highest level of zero inflation occurred when 1000 cells were sampled, and thus 9000 cells were assigned an abundance of zero simply because they were not sampled, and the lowest level of zero inflation occurred when 10,000 cells were sampled and none was assigned an abundance of zero for this reason. At the same time, zero inflation resulting from biological processes was adjusted by adding a threshold below which the altitude species can no longer survive, but keeping constant the number of true occurrence points generated each time. With higher altitude thresholds, the species occurrences were increasingly aggregated, and more cells were classified as true zeros. Altitude across the study area ranged from 0 to 284 m, so we tested threshold values of 0, 50, 100, 125, 150, 175 and 200 m (see Appendix S1, Table S1.1 for number of cells above each threshold). Above these thresholds, species occurrences were placed in a similar way based on weighted probability calculated from a logarithmic scaler of the original altitude predictor as described previously. Both the random species and altitude species were examined in scenarios with varying sample sizes, but obviously only the latter was tested using the altitude threshold method.

Based on the results of Simulation 1, we selected three predictor combinations to fit the models and create predictions. These included the GLM with both the bias and biological predictor (GLM4) and two of the ZI models which differ only in the inclusion (ZI6) or exclusion (ZI2) of the bias predictor from the zero component (Table 2). Although theoretically a ZI model that has only the biological predictor in the count component, but both the biological and bias predictors in the zero component (as with ZI3), would be the most obvious choice, in the real world the bias predictor may also have some biological influence on the species distribution, and the researcher may not be sure whether it is a better predictor of bias or biology. We therefore chose to use ZI6 rather than ZI3, to simulate better a real-world modelling scenario in which the causes of bias are unknown.

Model performance (*D*) was calculated for each simulation run with a particular combination of sample size and altitude threshold. Finally, in order to evaluate the improvement in model performance created by adding predictors of zero inflation, the difference in '*D*' was calculated between each model (GLM4 and ZI2, GLM4 and ZI6, and ZI2 and ZI6). This was repeated using both count abundance and sampling abundance predictions for the ZI models. Again, model fitting was repeated 10 times each with two sampling strategies (random and biased). Therefore, there were 200 simulation runs for the random species (10 repeats, two sampling strategies and 10 levels of sampling zero inflation) and 1400 simulation runs for the altitude species (10 repeats, two sampling strategies, 10 levels of sampling zero inflation and seven altitude thresholds (levels of biological zero inflation)).

2.3.3 | Simulation 3: comparing abundance versus presence–absence when aggregating spatial data

Often when fitting distribution models the only data available are presence-only, and multiple species occurrences within a grid cell are usually classified as a single presence. Often the predictors are only available at a coarser spatial scale than the species occurrence data, forcing the modeller to aggregate occurrences into coarser scale presence-only or presence–absence estimates. The coarser the resolution at which the distribution is modelled, the more information is lost about both the precise location of species occurrences, and species abundance (or occurrence density). However, if occurrences are instead aggregated into count data, information about abundance or occurrence density is retained at all scales, which may be more beneficial to conservation purposes. Therefore, even if only presence-only data are available, ZI models fitted at a larger spatial scale using the summed counts of occurrence may provide a better modelling method than traditional presence-only SDM that aggregate multiple occurrences into presence–absence data. This effect is likely to be more pronounced when the species data are biased, because ZI models attempt to model the excess zeros from sampling bias, whereas other methods, unless they explicitly incorporate bias correction, make no attempt to model or remove the bias.

Our final simulation study addressed this question by comparing the performance of Poisson GLM and ZI models predicting the abundance of the altitude species (as was carried out in Simulation 1) with two commonly used modelling methods that predict presence–absence: presence–absence binomial GLMs and presence-only MaxEnt models. This represents a scenario where the raw species occurrences (simulated at a 100-m resolution) are available at a greater resolution than the predictors (at a 1-km resolution), so the modeller is required to make a decision on how to aggregate the data.

To fit the binomial GLM presence–absence models, the source data for which need to be in the form of presence–absence rather than abundance, simulated 1-km cells that received an abundance count of zero based on either the random or biased sampling strategy for the ZI models in Simulation 1 (i.e. 80% of cells that were not considered to have been sampled) were classified automatically as an absence, and any cell with species occurrences that was sampled was classified as a presence. All binomial GLMs were fitted using the package 'stats' in R. As with Simulation 1, two GLMs were fitted, one with only the biological predictor ('Binomia-GLM1' equivalent to GLM3) and one with the biological and bias predictors ('Binomial-GLM2' equivalent to GLM4). Binomial occurrence predictions (i.e.

predicted probability of presence) were estimated across the study area from each model using 10-fold cross-validation.

Two MaxEnt presence-only models were also fitted to the altitude species occurrence data, one with altitude as the only predictor ('Maxent1') and one with both altitude and distance from nearest town as predictors ('Maxent2'). To produce presence-only data collected under a random or biased sampling strategy, only occurrence points at a 100-m resolution that fell within a 1-km cell that had been sampled for the ZI models in Simulation 1 were retained; only these cells would be classified by MaxEnt as a presence. Each model was fitted using the 'dismo' package (Hijmans et al., 2017) in R, at a 1-km resolution with 10,000 randomly selected background 'pseudo-absences' and 10 repetitions across each set of town centres.

Comparing the performance of count/abundance models (Poisson GLM and ZI models) and presence/presence–absence models (MaxEnt and binomial GLMs) required evaluation metrics which could work with both types of model. As it is less feasible to convert presence–absence predictions to abundance to use 'D', two other evaluation metrics were selected: Area under the curve (AUC) and the Spearman's rank correlation coefficient (r_s) between the model predictors ('altitude' and/ or 'distance from town') and each of the model predictions of count/abundance (GLM/ ZI) or habitat suitability (MaxEnt/ binomial GLM). In order to calculate AUC for the ZI and GLM models, abundance predictions were converted to binary presence–absence predictions, using an abundance threshold above which the species was considered to be predicted to be present. Because some models produced predicted abundances that all fell below 1, the threshold for conversion was chosen to be the mean abundance prediction across all grid cells for each individual model, that is the threshold varied across each GLM or ZI model. Mean AUC was calculated across the 10 repetitions for each model based on the presence–absence predictions for all models compared to the true presence–absence based on all occurrence locations across the study area. It should be noted that neither of these metrics offer a perfect measure of model performance. AUC causes a loss of information from the Poisson GLMs and ZI models, which are designed to predict abundance, while Spearman's rank retains more of the information in the predictions of both types of model, but is necessarily relatively crude.

Finally, in order to assess the impact of the scale of data aggregation on the performance of abundance and presence–absence models, additional models were fitted and compared across two other scales of increasing coarseness: 2 and 5-km. The larger the grid cell, the larger the mean count of occurrences per cell, and hence the more data potentially lost by converting to presence–absence. ZI count abundance predictions at a 2 and 5-km scale were obtained following the methodology of Simulation 1 using the ZI6 model structure and again converted to presence–absence predictions. MaxEnt and binomial GLM presence–absence predictions at a 2 and 5-km scale were obtained following the methodology outlined previously in Simulation 3. Model predictors (altitude and distance from town) were converted to coarser scales by calculating the mean values of each predictor at a 1-km resolution for each 2 or 5-km cell. As before, all predictions were evaluated using AUC and Spearman's rank correlation coefficient (r_s).

3 | RESULTS

3.1 | Simulation 1: investigating the accuracy of species distribution maps from ZI models

The results from Simulation 1 confirm that count abundance predictions from the ZI models provide the most accurate estimates (according to the metric *D*) of true species abundance (Figure 2 and Appendix S1, Figure S1.2). Estimating true abundance based purely on the biology of the species rather than sampling processes is usually the aim of ecological research, and these results suggest the count abundance predictions are most likely able to fulfil these aims. In contrast, all GLMs are poor at predicting true abundance because they do not separately model the excess (false) zeros generated by grid cells that have not been sampled. The problem is exaggerated when sampling is not just incomplete, but is also biased; if the GLM includes a predictor which is correlated with sampling effort (distance from nearest town), the model performs even less well (compare pink and blue bars for GLM3 (without bias predictor) and GLM4 (with bias predictor) in Figure 2) because it detects a spurious negative association between this predictor and abundance (top panels, Appendix S1, Figure S1.3). Similarly, ZI sampling abundance predictions (predictions from the whole model that potentially include the influence of sampling bias) perform poorly; rather than estimating true abundance, reflecting the species niche, they predict abundance as it would appear to observers employing each sampling strategy (Figure 2 and Appendix S1, Figure S1.2). Again, these predictions are particularly poor when sampling is biased (compare pink and blue bars for ZI2 and ZI6 in Figure 2). These findings hold true for all three species (altitude, altitude_randomised and random) (Appendix S1, Figures S1.2 and S1.3).

The ability to model excess zeros separately led to dramatically improved predictive power of true abundance for all ZI models (see count abundance predictions in Figure 2 and Appendix S1, Figure S1.2), although one (ZI2) performed relatively less well than the others when sampling was biased (Figure 2 and Appendix S1, Figure S1.2). In ZI2, the bias predictor was included in the count component but not the zero component, meaning that like the GLMs it detected a spurious negative association between abundance and distance from the nearest town (middle panels, Appendix S1, Figure S1.3); if they included the bias predictor, the other ZI models (e.g. ZI3 or ZI6) correctly detected that it was positively associated with the probability of an excess zero being recorded (lower panels, Appendix S1, Figure S1.3).

Predicted distribution maps based on both the count abundance predictions and sampling abundance predictions also support these findings (Figure 3 and Appendix S1, Figure S1.4). Maps produced using ZI count abundance predictions that account for bias where necessary (i.e. including predictors of bias in the zero

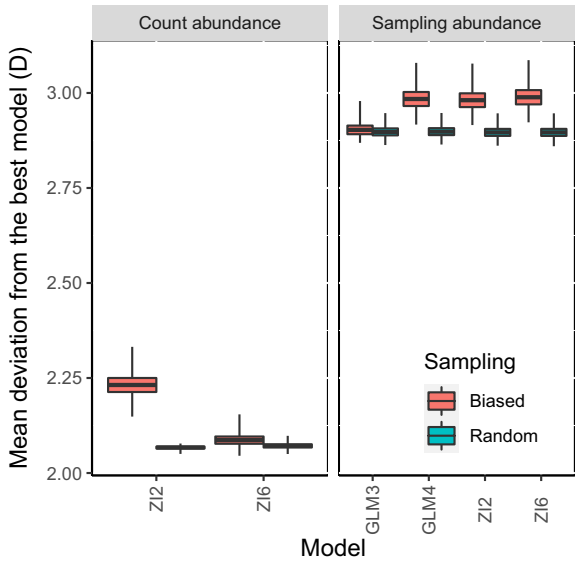

FIGURE 2 Evaluation of abundance predictions (based on D = 'deviation from the best model') for a hypothetical organism with occurrences simulated based on a preference for high altitudes (altitude species). Mean D values (±SE and data range) are shown for each sampling strategy (random or biased) across the 10 model repetitions. (a) (left) shows the evaluation of the count abundance predictions (from the zero-inflated (ZI) model count component only) from one model that accounts for sampling bias in the zero component (ZI6) and one model that does not (ZI2). (b) (right) shows the evaluation of the sampling abundance predictions (predictions from the whole model, and thus can be obtained from both ZI models and Generalised Linear Models (GLMs)) for four models: the same two ZI models as in (a), along with two GLMs: GLM3 including only the biological predictor and GLM4 including the biological and bias predictors. Only sampling abundance can be obtained from the GLMs, hence why (a) only shows results from the ZI models

component when sampling is biased) correlate strongly with the biological predictor layer (altitude) ($r_s > 0.9$) and show little influence of bias (distance from towns) (Appendix S1, Figure S1.5). When sampling is biased, neglecting to account for the bias in the zero component, or using the sampling abundance predictions both result in low-accuracy distribution maps that correlate more strongly with the bias predictor (r_s value between −0.64 and −0.71) and less strongly with the biological predictor (r_s values between 0.60 and 0.74) (Appendix S1, Figure S1.5). Distribution maps produced by the GLMs were also less accurate when sampling was biased and predictors correlating with bias were included (Figure 3 and Appendix S1, Figure S1.4). Maps from the GLMs which include the bias predictor (GLM4) show a strong influence of sampling bias similar to that seen in the ZI sampling abundance predictions. These maps show relatively weak correlations to the altitude predictor ($r_s = 0.60$) compared to their counterpart GLMs that do not include the bias predictor (GLM3) ($r_s = 0.99$) (Appendix S1, Figure S1.5). The prediction map from the GLM including both the biological and bias predictors (GLM4) with biased sampling also shows a strong correlation to the bias predictor ($r_s = -0.72$).

Additional maps that depict the probability of each grid cell being an excess zero (i.e. predictions from the zero component of a ZI model) further highlight the ability of ZI models to model separately the biological and sampling processes, as well as provide insight into the nature of bias in the species data (Figure 3 and Appendix S1, Figure S1.4). This means that in real studies in which the sources of sampling bias are unknown, inclusion of predictors that may correlate with sampling bias (e.g. distance to towns or roads, accessibility, land use etc.) in both the count and zero components of ZI models can help to both model and identify likely causes of bias. This is a unique feature of the ZI models, and is something which the GLMs are unable to reproduce; these models cannot provide insight into the bias or prediction maps that eliminate sampling effects within the data.

3.2 | Simulation 2: examining the impact of the extent of zero inflation in the data

Real species occurrence or abundance data will suffer from variable levels of zero inflation resulting from both biological and sampling processes. Therefore, the better performance of ZI models compared with GLMs described in Simulation 1 may not occur in all circumstances, so exploring this issue was our aim of Simulation 2. As anticipated, ZI count abundance predictions and GLM abundance predictions have similar accuracy when the data are not zero-inflated; when the whole study area is surveyed, all absences are 'true absences', the species is randomly distributed with no biological zero inflation and the difference in performance is zero (Figure 4, see random species (R) in left and middle panels). When considering the random species only (i.e. with no biological zero inflation), as less of the study area is surveyed, zero inflation as a result of sampling increases, and therefore the effectiveness of ZI model count abundance predictions improves in comparison to GLMs. Although this phenomenon occurs under both sampling strategies, it is most noticeable when both sampling is biased and that bias is accounted for in the model (e.g. by including the bias predictors in the ZI zero component as in ZI6).

As with the random species, when there are high levels of incomplete sampling for the altitude species (e.g. ~20% or fewer cells are sampled), ZI model count abundance predictions are consistently better than GLM predictions, regardless of biological zero inflation (Figure 4, left and middle panels). However, as more of the area is surveyed (>20%), the difference in performance decreases. At low levels of biological zero inflation, this difference tends towards zero. However, at higher levels of biological zero inflation, GLM predictions are actually more accurate than the ZI model count abundance predictions under both random and biased sampling scenarios. This can best be understood by looking at Appendix S1, Figure S1.6 showing the results based on sampling abundance predictions from the ZI model, rather than count abundance predictions; in contrast to the count abundance predictions, as biological zero inflation increases, ZI sampling abundance

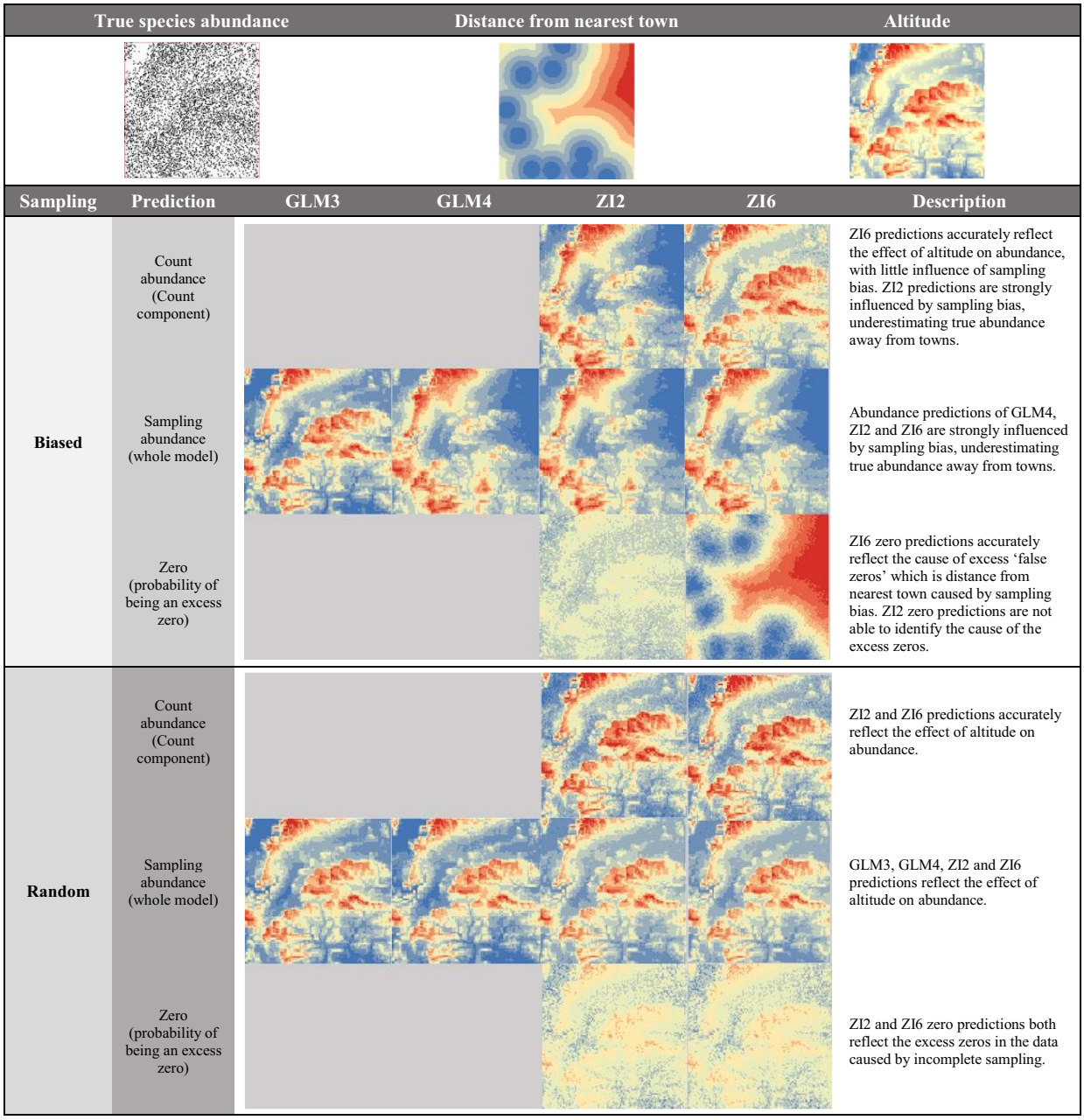

FIGURE 3 Example maps of abundance for a hypothetical species ('altitude species') whose occurrence is positively influenced by altitude, produced from two Generalised Linear Models (GLMs) and two zero-inflated (ZI) models. Models were built with either data collected by randomly sampling grid cells (random) or with sampling bias (biased). Abundance maps from GLM3 (including the biological predictor only) and GLM4 (including both the biological and bias predictors) are produced using sampling abundance predictions (i.e. from the whole model). Both count abundance and sampling abundance predictions can be produced from the ZI models along with a map of the probability a cell is an excess zeros (zero). Both ZI models include a biological predictor (altitude) of both abundance and excess zeros, and bias predictor (distance from the nearest town) of abundance. ZI6 also includes 'distance from the nearest town' as a predictor of excess zeros. Individual cells are colour-coded based on abundance for the abundance predictions or on probability of being an excess zero for the zero predictions (high = red, low = blue)

predictions increasingly outperform those of the GLM. This is because the zero component, which is combined with the count component to create the sampling abundance prediction, is able to predict the excess zeros caused by the biological driver, while the GLM cannot. Therefore, if high levels of biological zero inflation are suspected in the data, both the count and sampling abundance predictions should be considered and evaluated before choosing the best predictions of species abundance.

Reiterating our results from Simulation 1, when sampling is random there is no benefit of including the bias predictor in the zero component under any levels of sampling or biological zero inflation (Figure 4 and Appendix S1, Figure S1.6, top right panels).

Random — Altitude Threshold

Number of sampled grid squares	Difference between GLM4 and ZI6								Difference between GLM4 and ZI2								Difference between ZI2 and ZI6							
	R	0	50	100	125	150	175	200	R	0	50	100	125	150	175	200	R	0	50	100	125	150	175	200
1000	-1.63	-1.47	-1.63	-1.94	-2.23	-2.52	-3.56	-7.38	-1.64	-1.06	-1.19	-1.42	-1.55	-1.63	-1.99	-3.63	0.00	0.00	0.00	0.01	0.01	0.02	0.02	0.02
2000	-0.96	-0.83	-0.91	-1.02	-1.04	-1.11	-1.12	-1.34	-0.96	-0.83	-0.91	-1.02	-1.04	-1.12	-1.12	-1.34	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00
3000	-0.59	-0.47	-0.52	-0.63	-0.50	-0.33	0.03	0.83	-0.59	-0.47	-0.53	-0.64	-0.51	-0.34	0.02	0.83	0.00	0.00	0.01	0.00	0.01	0.01	0.01	0.00
4000	-0.37	-0.24	-0.17	-0.25	-0.07	0.15	0.78	2.40	-0.37	-0.25	-0.18	-0.26	-0.08	0.15	0.78	2.39	0.00	0.00	0.00	0.01	0.01	0.01	0.01	0.00
5000	-0.22	-0.10	-0.06	0.10	0.16	0.50	1.27	3.60	-0.22	-0.09	-0.06	0.10	0.16	0.49	1.26	3.59	0.00	-0.01	0.00	0.00	0.01	0.01	0.01	0.00
6000	-0.13	-0.02	-0.03	0.10	0.44	0.78	1.63	4.43	-0.13	-0.02	-0.03	0.10	0.44	0.78	1.63	4.43	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00
7000	-0.07	-0.01	-0.01	0.15	0.43	0.96	1.91	4.71	-0.07	-0.01	-0.01	0.15	0.43	0.96	1.90	4.71	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
8000	-0.03	0.00	0.00	0.19	0.48	1.08	2.10	5.02	-0.03	0.00	0.00	0.19	0.48	1.08	2.10	5.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
9000	-0.01	0.00	0.02	0.24	0.59	1.14	2.15	4.97	-0.01	0.00	0.02	0.24	0.59	1.14	2.15	4.97	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10000	0.00	0.00	0.03	0.28	0.65	1.22	2.21	5.09	0.00	0.00	0.03	0.28	0.65	1.22	2.21	5.09	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Biased — Altitude Threshold

Number of sampled grid squares	Difference between GLM4 and ZI6								Difference between GLM4 and ZI2								Difference between ZI2 and ZI6							
	R	0	50	100	125	150	175	200	R	0	50	100	125	150	175	200	R	0	50	100	125	150	175	200
1000	-1.71	-1.58	-1.70	-2.02	-2.14	-2.39	-3.00	-4.80	-1.64	-1.44	-1.58	-2.00	-2.13	-2.48	-2.86	-4.69	-0.07	-0.13	-0.11	-0.02	0.06	0.09	0.08	0.05
2000	-1.04	-0.90	-1.00	-1.15	-1.16	-1.17	-1.44	-2.55	-0.95	-0.75	-0.86	-1.14	-1.23	-1.23	-1.49	-2.58	-0.09	-0.14	-0.13	0.00	0.00	0.07	0.05	0.02
3000	-0.68	-0.53	-0.57	-0.65	-0.55	-0.38	-0.07	0.70	-0.60	-0.40	-0.43	-0.64	-0.61	-0.45	-0.13	0.66	-0.09	-0.13	-0.14	-0.02	0.06	0.07	0.06	0.04
4000	-0.44	-0.32	-0.27	-0.30	-0.16	0.05	0.64	1.96	-0.37	-0.19	-0.14	-0.14	-0.22	-0.01	0.59	1.92	-0.07	-0.13	-0.13	-0.16	0.06	0.06	0.05	0.04
5000	-0.29	-0.16	-0.11	0.18	0.48	0.42	1.16	2.68	-0.23	-0.06	-0.06	0.04	0.05	0.35	1.11	2.65	-0.06	-0.10	-0.05	-0.06	0.04	0.06	0.05	0.03
6000	-0.18	-0.07	-0.05	0.10	0.30	0.61	1.45	3.18	-0.13	-0.02	-0.03	0.10	0.33	0.55	1.40	3.15	-0.05	-0.05	-0.01	0.00	-0.03	0.06	0.05	0.03
7000	-0.10	-0.03	-0.02	0.13	0.42	0.95	1.74	4.43	-0.07	-0.01	-0.02	0.13	0.41	0.90	1.71	4.42	-0.03	-0.03	0.00	0.00	0.01	0.04	0.03	0.01
8000	-0.05	-0.02	0.00	0.19	0.47	0.98	2.04	4.76	-0.03	0.00	0.00	0.18	0.47	0.94	2.01	4.75	-0.02	-0.01	0.00	0.00	0.00	0.04	0.03	0.01
9000	-0.02	-0.01	0.01	0.23	0.57	1.11	2.11	4.81	-0.01	0.00	0.01	0.23	0.57	1.10	2.06	4.80	-0.01	0.00	0.00	0.00	0.00	0.00	0.05	0.00
10000	0.00	0.00	0.03	0.28	0.65	1.22	2.21	5.09	0.00	0.00	0.03	0.28	0.65	1.22	2.21	5.09	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

-10	-7.5	-5.0	-2.5	0.0	2.5	5	7.5	10

-0.4	-0.3	-0.2	-0.1	0.0	0.1	0.2	0.3

FIGURE 4 Comparisons of model predictive power of true abundance between a Generalised Linear Model (GLM) and two zero-inflated (ZI) models across varying levels of biological and sampling bias zero inflation. Values represent the mean difference in *D* ('deviation from the best model') between GLM4 (containing both biological and bias predictors), ZI2 (excludes the bias predictor from the zero component) and ZI6 (includes the bias predictor in the zero component). Biological zero inflation was increased by introducing a minimum altitude threshold below which the species cannot survive and therefore reducing its environmental niche. Sampling-related zero inflation was increased by increasing the number of grid cells sampled across the study area in increments of 10%. Negative (red) values show scenarios where the ZI model performs better than the GLM (left and middle panels) or where ZI6 performs better than ZI2 (right panel), whereas positive (blue) values show scenarios where GLM4 outperforms the ZI models or ZI2 outperforms ZI6. 'R' represents the values for the random species whose occurrence is not related to altitude

Under biased sampling scenarios, models accounting for bias (e.g. by including the bias predictor in the zero component as in ZI6) are most effective when there are high levels of sampling-related zero inflation and low levels of biological zero inflation. As either the area surveyed or biological zero inflation increases, the effectiveness of these models reduces compared to models that fail to account for bias (Figure 4, bottom right panel). Nevertheless, the majority of differences seen between ZI models are relatively small compared to those between the ZI models and GLMs.

3.3 | Simulation 3: comparing abundance versus presence–absence data across multiple spatial scales

The results from Simulation 3 support our hypothesis that, when dealing with biased species data, modelling aggregated count data using ZI models is a better choice than modelling aggregated presence–absence or presence-only data, as is commonly done in traditional SDM studies, using approaches such as binomial GLMs or MaxEnt (Figure 5). The only model to perform consistently well across all spatial scales when dealing with the biased species data was the ZI model, which maintained strong correlations to the biological predictor ($r_s > 0.9$) and low correlations to the bias predictor ($-0.12 < r_s < 0.07$) across all scales (Figure 5). Predicted maps of the altitude species distribution also show that the ZI model count abundance predictions provide the most accurate reflection of the true species distribution as the scale of data aggregation increases (Appendix S1, Figure S1.7). Binomial-GLM2 and MaxEnt2 models, which incorporate the bias predictor, produced predictions that are heavily influenced by sampling bias at a 1-km scale, with strong correlations to the bias predictor ($r_s < -0.75$) (Figure 5 and Appendix S1, Figure S1.7). These increase in strength as scale increases to 2 and 5 km, so that both model predictions produce correlations to the bias predictor close to 1 ($r_s < -0.92$). Both MaxEnt1 and binomial-GLM1 (which do not include the bias predictor) were able to produce accurate predictions with the biased data at a 1-km resolution, although performance declined as the scale became coarser. Even when the species data were collected using a random sampling strategy, the performance of the presence–absence models declined as the scale became coarser and more information was lost with data aggregation (Figure 5); this phenomenon was not seen in the ZI models and performance remained high as scale increased.

Model evaluation using mean AUC based on the presence–absence predictions also supports these findings (Figure 6 and Appendix S1, Figure S1.8). Across all three scales, the ZI model was best suited to model the biased species data compared to the MaxEnt and binomial GLM models that were fitted using the bias predictor (Figure 6). The presence–absence models have a much larger variance in performance than the ZI abundance models, especially at coarser scales, with some repetitions producing AUC values below 0.5 and above 0.9 (Figure 6 and Appendix S1, Figure S1.8). The ZI

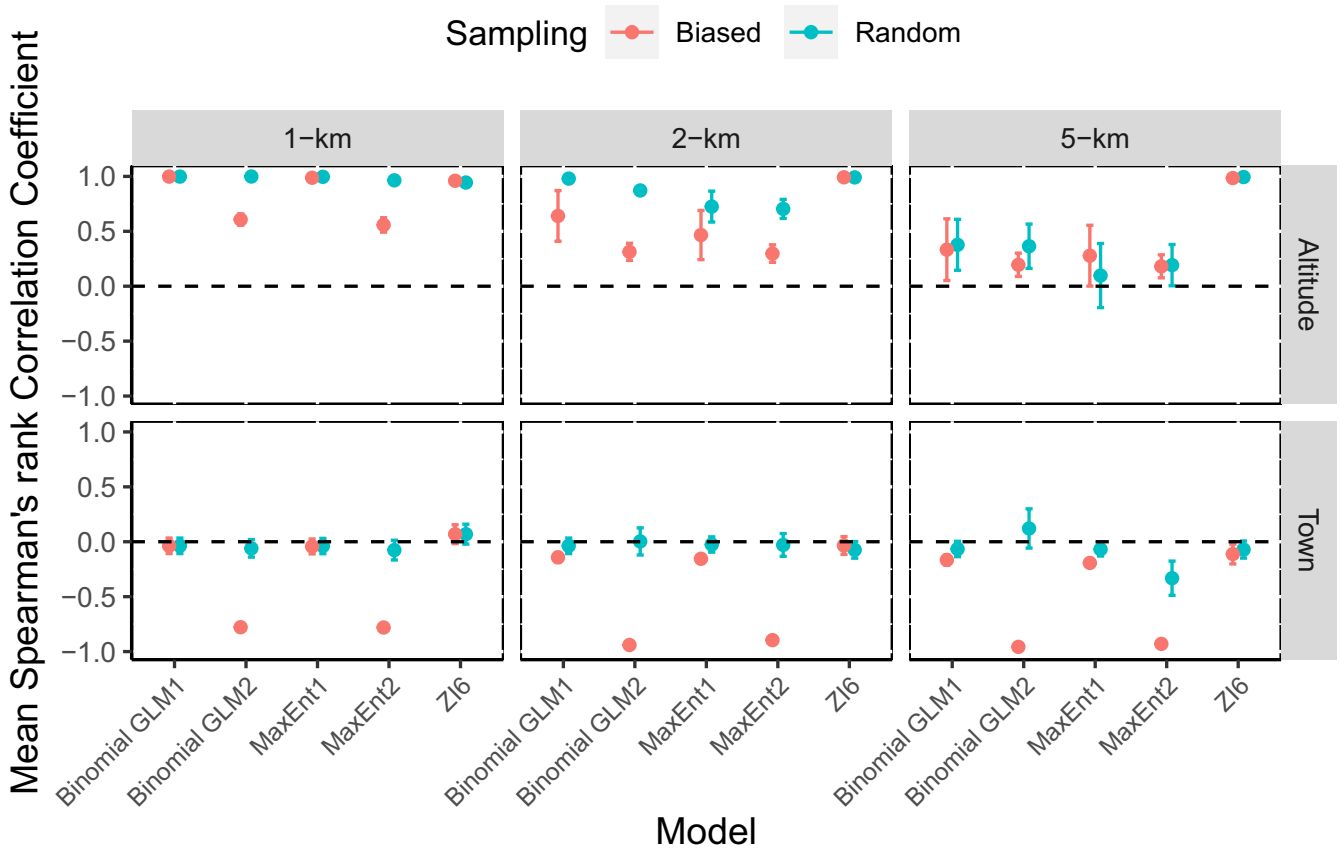

FIGURE 5 Mean spearman's rank correlation coefficients (r_s) (±SE) between the model predictors (altitude and distance from nearest town) and model predictions for altitude species across three modelling scales: 1, 2 and 5 km and two sampling strategies (random and biased). Three types of model are compared: (1) binomial Generalised Linear Models (GLMs) that predict the probability of occurrence, (2) maximum entropy (MaxEnt) models that predict the probability of occurrence and (3) zero-inflated (ZI) models that predict the true (count) abundance of the species. Binomial-GLM1 and MaxEnt1 include only the biological predictor in the model, whereas Binomial-GLM2 and MaxEnt2 include both the biological and bias predictors. ZI6 model includes the bias and biological predictor in both the count and zero components

model also outperformed several of the MaxEnt and binomial GLMs fitted without the bias predictor, including the MaxEnt1 model at a 2-km scale and the binomial-GLM1 at a 5-km scale (Appendix S1, Figure S1.8), although it produced slightly lower mean AUC values than some of the presence–absence models when the bias predictor was excluded. Nevertheless, if the sampling bias source is unknown, it might be difficult to exclude completely predictors correlating with the bias, so choosing a ZI model is still likely to be the safest option to produce the best, most robust predictions least affected by sampling bias.

4 | DISCUSSION

Sampling bias in species data is problematic for SDM, and many researchers call for greater awareness and development of correction methods to deal with this issue (Araújo & Guisan, 2006; Bystriakova et al., 2012; Kramer-Schadt et al., 2013). Our simulations using ZI models highlight a novel approach for dealing with sampling bias and zero inflation in SDM, which we believe can be applied to a wide variety of ecological and conservation research questions that use large databases of species records. Our results reveal that ZI models have the potential both to reduce the impact of bias on predictions which are used for biological inference, and to provide insights into previously unknown causes or correlates of sampling bias. This method can be used with both raw abundance data, and with abundance data created by summing occurrences from presence-only data across a larger spatial scale, and therefore offers an alternative to traditional presence-only SDM methods. As spatial occurrence data are often present at a finer scale than the environmental predictors, decisions about data aggregation have to be made when fitting distribution models. We found that even though information about the precise location of species occurrences is sacrificed, aggregating species occurrences into counts of abundance and fitting ZI models produces better estimates of a species distribution, especially when the species data are biased by sampling methods, than aggregating occurrences into presence–absence form at a coarser spatial scale, as is common with traditional SDM methods such as binomial GLMs or MaxEnt.

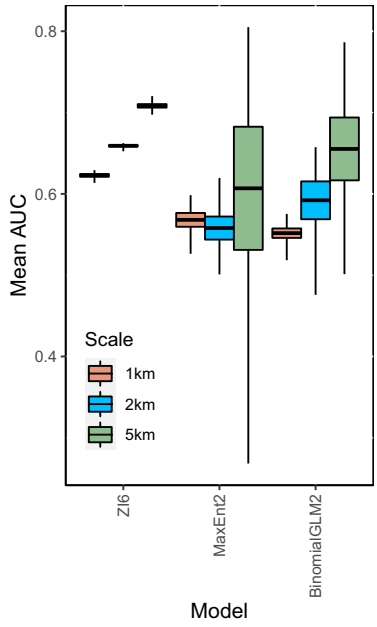

FIGURE 6 Evaluation of MaxEnt, Generalised Linear Model (GLM) and zero-inflated (ZI) model predictions of altitude species presence–absence sampled using a biased strategy across the study area. Mean area under the curve (AUC) (±SE and data range) across the 10 model repetitions is used to evaluate predictions across three scales of data aggregation: 1, 2 and 5 km. Three models are compared: (1) a ZI model able to account for the bias in the zero component (ZI6) (see Methods for more information on the conversion of ZI abundance predictions to presence–absence), (2) a MaxEnt model (MaxEnt2) that includes altitude and distance from town as predictors and (3) a binomial GLM (Binomial-GLM2), also including altitude and distance from town as predictors

Species distribution maps are an important resource for conservation planners (Rodríguez et al., 2007), yet there is often little consideration of inaccuracies or uncertainty in these maps or associated models (Elith et al., 2002; Zuquim et al., 2014). Our results show how the biological information value of maps based on GLM, MaxEnt and ZI sampling abundance predictions can be reduced by sampling bias. In contrast, the distribution maps produced from the predictions from the count component of ZI models are accurate reflections of the species niche and true abundance, even when species data are spatially biased, providing that the bias influence is accounted for in the model by included all predictors suspected of capturing or correlating with the bias in both ZI count and zero components. If in doubt about whether a predictor is likely to be a source of bias, inclusion in both parts will not only alleviate the problem of bias, but will also provide insight into whether it actually is introducing a large number of excess ('false') zeros. Additionally, ZI model coefficients allowed examination of potential causes of bias; in ZI6 (the model including both the bias and biological predictor in the zero component) from Simulation 1, 'distance from nearest town' was influential only in the zero component, and was not spuriously identified as influencing true abundance. Currently, there are few statistical models that allow post-modelling identification of bias sources. Many SDM techniques rely on prior understanding and some form of quantification of the bias in order to remove it (Phillips, 2008), so ZI models provide an advantage over these traditional bias correction methods in their ability to shed light on potential causes of bias.

If all excess zeros are false zeros, count abundance predictions from ZI models should always reflect the true species niche, and the zero component will be modelling only excess zeros from non-biological, sampling processes. However, this scenario is unlikely in ecological systems. In reality, as in our simulations with the altitude and altitude_randomised species, the excess zeros will result from a combination of biological zero inflation and sampling zero inflation. Therefore, the count abundance prediction may not always be predicting true abundance, and the zero component may actually be dominated by biological processes, as we suggest is the case for the results from Simulation 2. In this case, the sampling abundance prediction will actually be a more accurate reflection of true species abundance. Nevertheless, by examining the significance and influence of predictors in both components, their plausibility as causes of bias can be inspected—biological predictors of abundance are likely to be significant in both parts of the ZI model, whereas sampling predictors are unlikely to appear influential in the count component.

After identifying potential bias predictors, modellers can make more informed choices about whether to eliminate these predictors from either ZI component, whether the zero component is more heavily dominated by biological or sampling processes and if the count abundance or sampling abundance is more likely to reflect true species abundance. A good understanding of the biology of the species being modelled is therefore key. Additionally, despite the post-model fitting ability of ZI models to distinguish bias, beginning any analysis of a ZI dataset, it is important also to try and identify the source of excess zeros as either from biology or sampling processes (Martin et al., 2005). Consequently, although one benefit of ZI models is the ability to use different sets of covariates in the count and zero components (Lambert, 1992; Zuur et al., 2009), it is important only to include appropriate, relevant predictors in each part where possible.

The collection of species data varies widely in its scale and standardisation, from single museum specimens collected by natural history experts to more local, standardised recording schemes (Pocock & Evans, 2014) and to international, opportunistic recording schemes such as eBird (Sullivan et al., 2009). The more standardised and directed the protocols, the lower the likelihood of sampling bias and 'false zeros' in the data. In these cases, a simple Poisson or negative binomial GLM may suffice rather than a ZI model; at very low levels of zero inflation the performance of the GLMs was shown to be equal to that of the ZI models in Simulation 2. Nevertheless, our findings from Simulation 2 suggest that, regardless of biological zero inflation, when sampling is suspected to be very incomplete (estimated coverage of total study area <~20%), ZI models will always the optimum choice. At low levels of biological zero inflation, we found ZI models to be more effective than GLMs even when sampling coverage approached levels as high as 90%, as might be the case for species with broad ranges that have been extensively documented, such

as important or conspicuous species in countries with long histories of species record keeping.

In addition to the Poisson distribution, the negative binomial distribution is also often used for count data, which can also be applied within a ZI modelling framework (Minami et al., 2007; Ridout et al., 2001; Zuur et al., 2009). The negative binomial distribution is able to model an extra proportion of the excess zeros compared to the Poisson distribution through the use of an extra model parameter (Θ) (Fisher, 1941) and can therefore account for biological aggregation and overdispersion in ecological data (Lindén & Mäntyniemi, 2011). We chose not to investigate a ZI negative binomial model in these simulations to remove confusion when communicating our main message, although we acknowledge that under high levels of biological zero inflation (as in Simulation 2), such models may well be more effective than the ZI Poisson models. Therefore, when analysing presence-only species data suffering from high levels of sampling bias, a ZI Poisson model will usually be effective, but it is valuable to know that there are different ZI model types that can be used to address ecological or statistical issues that may arise in species data.

The majority of SDM research to date has focused on producing presence–absence or presence-only distribution maps of species or communities (Brotons et al., 2004; Lyashevska et al., 2016; Phillips et al., 2006). Species abundance maps are produced more infrequently, often due to the practical difficulty of measuring absolute abundance (Lyashevska et al., 2016). However, their ability to display extra information about density means they are often more informative and preferred (Barry & Welsh, 2002; Johnston et al., 2015; Pearce & Ferrier, 2001).

Although count data are known commonly to suffer from zero inflation, ZI models have been used to produce accurate species abundance maps from systematically collected species data in very few studies (Bouyer et al., 2015; Lyashaveska et al., 2016), and none have acknowledged or explored bias in their data. It is also not recommended to use SDM to predict species abundance from presence-only or presence–absence data (Jiménez-Valverde et al., 2021), so ZI models that fit abundance by default should be able to cover this methodology gap in the field of SDM. Additionally, scale is hugely important in SDM. Species distributions are often modelled at coarse resolutions across national or international scales due to the availability of predictors, even though occurrences relate more to localised environmental factors (Guisan et al., 2007; Kuemmerlen et al., 2014). The coarser the grain size used in presence–absence or presence-only SDM, the more the raw occurrences are aggregated into a binary variable and density information is lost. Therefore, it is likely that at coarse resolutions, using abundance rather than occurrence data, preserve more information and will produce more accurate maps of habitat suitability.

Our findings from Simulation 3 suggest that when having to decide how to aggregate data to match the coarser resolution of the environmental predictors, the best method is to aggregate species occurrences into counts of abundance and fit using a ZI model, rather than aggregate into presence–absence data and fit using a traditional SDM method such as MaxEnt. This provides two main benefits over presence–absence methods in that (a) ZI models are able to identify and account for bias without prior knowledge of the bias sources and (b) extra information about species abundance is retained and modelled. We found that as scale became increasingly coarser, only the ZI models retained a high level of predictive power and were an accurate reflection of species niche compared to MaxEnt or binomial GLMs, especially when the data suffered from sampling bias. We believe that ZI models have an advantage over other statistical methods in that they can be used with either presence–absence data or abundance data collected from citizen science projects—presence–absence data can just be aggregated into a count at a particular resolution. Furthermore, scale was shown to have little influence on the predictive power of ZI models providing bias was accounted for. Nevertheless, this was only simulated across relatively small resolutions (up to 5 km) due to the limitations of the study area and requirement for ZI data, whereas many studies map distributions at larger scales (>10 km) (Luoto et al., 2007; Thuiller et al., 2006). It is therefore uncertain whether this pattern holds true across more coarse scales of analysis.

In this paper, we have investigated the performance of ZI models under a relatively restricted set of scenarios. We acknowledge that our findings may therefore be case specific and we are addressing this with ongoing research (Nolan et al., unpubl.). For example, we chose to use a simple scenario in which only two predictors, a biological predictor and a bias predictor, generate patterns in the species distribution. The altitude species was assigned a simple preference for high altitudes, when in fact, there are likely several different environmental influences on the species niche. Furthermore, some of these biological predictors of species presence will also predict sampling bias. Therefore, it is important that prior consideration is given to the possible influences of any predictor included in the model on both ecological processes and sampling behaviour before it is decided whether to include it in either part of the ZI model.

GLMs, and by extension ZI models, have been criticised for their inability to capture the complex, nonlinear relationships which may often characterise species responses to the environment, in contrast to more modern methods such as MaxEnt or other machine learning techniques which are more flexible (Austin, 2002). Nevertheless, GLMs and ZIs also have some clear benefits, such as the ease with which they can be applied, and the transparency of their design. Here, we have shown an additional benefit of ZI models not yet available with any other modelling approach—the ability to simultaneously account for bias and to make inferences about it, when predicting distributions from incomplete sampling. We believe that our approach using ZI models has broad applicability to a variety of scenarios when bias is present, and there are suspected predictors of bias available. ZI models should be especially valuable when species abundance is of interest to the modeller, such as when modelling distributions of individual large animals or trees. Although we acknowledge that GLMs and ZI models have limitations, there is a range of options for more complex versions of these models, such as those incorporating polynomial terms, interactions and LASSO variable selection (Hastie et al.,

2009; Vollering et al., 2019), which might allow such models to capture nonlinear/complex responses to the environment at the same time as modelling the causes of excess zeros.

In our simulations, we assume that all 'false absences' are due to sampling bias, but it is likely that in many cases, particularly for rare or cryptic species, they are also generated by detection errors (Dickinson et al., 2010; Fitzpatrick et al., 2009; Kosmala et al., 2016). The species range size and the scale of detectability of the individuals is likely to influence the interpretation of the model 'abundance' predictions. For example, underestimation of true abundance could occur when modelling small organisms which appear frequently during the survey, and will be more representative of the likelihood of successfully sampling the species. On the other hand, overestimation could occur when modelling large, mobile organisms that cover multiple sampling locations, so prediction abundance might be a proxy of the probability of encountering one of a small number of individuals. Hence, there may be three sources of excess zeros, namely true zeros from unsuitable habitat, false zeros from lack of sampling and false zeros from detection error. When detection errors are significant, ZI models will not be able to distinguish between the different types of false zeros; but by including predictors in both the count and zero components of the model that capture the processes generating all types of zeros, we believe that ZI models will still be able (mostly) to account for these excess 'false' zeros, and combined with expert knowledge can provide some information about their sources.

5 | CONCLUSION

Large collections of species data are extremely useful for SDM and conservation, and yet are limited by issues associated with the recording processes, including sampling bias and zero inflation. Our simulations show that ZI models can fit biased data and identify sources of bias. Most importantly for conservation, by using only predictions from the count component of the ZI model (i.e. the count abundance predictions), biased species data can be used to produce distribution maps comparable to those using unbiased data. We also highlight the importance of considering the use of abundance data in SDM, especially at large spatial scales, when valuable ecological information about density is lost if data in each cell are converted to presence–absence or presence-only. ZI models are advantageous compared to other commonly used SDM techniques such as MaxEnt owing to their ability to retain information about abundance and also to identify and remove bias without prior knowledge of the bias sources. We believe ZI models have been largely overlooked in ecological research, even though they have a huge potential to be useful in SDM, and could have great benefits for conservation and our environment.

ACKNOWLEDGEMENTS

This research was made possible by funding support from the University of Nottingham, UK and the Woodland Trust, Grantham, UK. The authors acknowledge Tim Newbold and Richard Field for their helpful comments on a draft version of this manuscript. No permits were needed to carry out the work in this manuscript.

CONFLICT OF INTEREST

All the authors declare there are no conflict of interest associated with any of the work in this manuscript.

DATA AVAILABILITY STATEMENT

R code and data used to produce the analyses and results reported within this manuscript can be found freely available at the following online location: https://doi.org/10.6084/m9.figshare.13118417.v2. Elevation and temperature data used in the study are publicly available for download from WorldClim DEM (https://www.worldclim.org/).

ORCID

Victoria Nolan https://orcid.org/0000-0002-6069-963X
Francis Gilbert https://orcid.org/0000-0002-2727-4103
Tom Reader https://orcid.org/0000-0001-7586-8814

REFERENCES

Araújo, M. B., & Guisan, A. (2006). Five (or so) challenges for species distribution modelling. *Journal of Biogeography*, *33*(10), 1677–1688. https://doi.org/10.1111/j.1365-2699.2006.01584.x

Austin, M. P. (2002). Spatial prediction of species distribution: An interface between ecological theory and statistical modelling. *Ecological Modelling*, *157*, 101–118. https://doi.org/10.1016/S0304-3800(02)00205-3

Barry, S. C., & Welsh, A. H. (2002). Generalized additive modelling and zero inflated count data. *Ecological Modelling*, *157*(2), 179–188. https://doi.org/10.1016/S0304-3800(02)00194-1

Bird, T. J., Bates, A. E., Lefcheck, J. S., Hill, N. A., Thomson, R. J., Edgar, G. J., Stuart-Smith, R. D., Wotherspoon, S., Krkosek, M., Stuart-Smith, J. F., Pecl, G. T., Barrett, N., & Frusher, S. (2014). Statistical solutions for error and bias in global citizen science datasets. *Biological Conservation*, *173*, 144–154. https://doi.org/10.1016/j.biocon.2013.07.037

Boakes, E. H., McGowan, P. J. K., Fuller, R. A., Chang-qing, D., Clark, N. E., O'Connor, K., & Mace, G. M. (2010). Distorted views of biodiversity: Spatial and temporal bias in species occurrence data. *PLoS Biology*, *8*(6), e1000385. https://doi.org/10.1371/journal.pbio.1000385

Boria, R. A., Olson, L. E., Goodman, S. M., & Anderson, R. P. (2014). Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. *Ecological Modelling*, *275*, 73–77. https://doi.org/10.1016/j.ecolmodel.2013.12.012

Bouyer, Y., Rigot, T., Panzacchi, M., Moorter, B. V., Poncin, P., Beudels-Jamar, R., Odden, J., & Linnell, J. D. C. (2015). Using Zero-Inflated models to predict the relative distribution and abundance of roe deer over very large spatial scales. *Annales Zoologici Fennici*, *52*(1–2), 66–76. https://doi.org/10.5735/086.052.0206

Brotons, L., Thuiller, W., Araújo, M. B., & Hirzel, A. H. (2004). Presence-absence versus presence-only modelling methods for predicting bird habitat suitability. *Ecography*, *27*(4), 437–448. https://doi.org/10.1111/j.0906-7590.2004.03764.x

Brunsdon, C., Fotheringham, S., & Charlton, M. (1998). Geographically weighted regression. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *47*(3), 431–443. https://doi.org/10.1111/1467-9884.00145

Bystriakova, N., Peregrym, M., Erkens, R., Bezsmertna, O., & Schneider, H. (2012). Sampling bias in geographic and environmental space

and its effect on the predictive power of species distribution models. *Systematics and Biodiversity*, *10*. https://doi.org/10.1080/14772000.2012.705357

Cunningham, R. B., & Lindenmayer, D. B. (2005). Modeling count data of rare species: Some statistical issues. *Ecology*, *86*(5), 1135–1142. https://doi.org/10.1890/04-0589

Dénes, F. V., Silveira, L. F., & Beissinger, S. R. (2015). Estimating abundance of unmarked animal populations: Accounting for imperfect detection and other sources of zero inflation. *Methods in Ecology and Evolution*, *6*(5), 543–556. https://doi.org/10.1111/2041-210X.12333

Dickinson, J. L., Zuckerberg, B., & Bonter, D. N. (2010). Citizen Science as an ecological research tool: Challenges and benefits. *Annual Review of Ecology, Evolution, and Systematics*, *41*(1), 149–172. https://doi.org/10.1146/annurev-ecolsys-102209-144636

Dormann, C. F., McPherson, J. M., Araújo, M. B., Bivand, R., Bolliger, J., Carl, G., Davies, R. G, Hirzel, A., Jetz, W., Kissling, W. D., Kühn, I., Ohlemüller, R., Peres-Neto, P. R., Reineking, B., Schröder, B., M. Schurr, F., & Wilson, R. (2007). Methods to account for spatial autocorrelation in the analysis of species distributional data: A review. *Ecography*, *30*(5), 609–628. https://doi.org/10.1111/j.2007.0906-7590.05171.x

Dudík, M., Schapire, R. E., & Phillips, S. (2005). Correcting sample selection bias in maximum entropy density estimation. *Advances in Neural Information Processing Systems*, *17*, 323–330.

Dwyer, R. G., Carpenter-Bundhoo, L., Franklin, C. E., & Campbell, H. A. (2016). Using citizen-collected wildlife sightings to predict traffic strike hot spots for threatened species: A case study on the southern cassowary. *Journal of Applied Ecology*, *53*(4), 973–982. https://doi.org/10.1111/1365-2664.12635

Elith, J., Burgman, M. A., & Regan, H. M. (2002). Mapping epistemic uncertainties and vague concepts in predictions of species distribution. *Ecological Modelling*, *157*(2–3), 313–329. https://doi.org/10.1016/S0304-3800(02)00202-8

Elith, J., Phillips, S. J., Hastie, T., Dudík, M., Chee, Y. E., & Yates, C. J. (2011). A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*, *17*, 43–57. https://doi.org/10.1111/j.1472-4642.2010.00725.x

ESRI (2013). *ArcGIS Desktop: Release 10*. Redlands, CA: Environmental Systems Research Institute.

Fisher, R. A. (1941). The negative binomial distribution. *Annals of Eugenics*, *11*(1), 182–187. https://doi.org/10.1111/j.1469-1809.1941.tb02284.x

Fitzpatrick, M. C., Gotelli, N. J., & Ellison, A. M. (2013). MaxEnt versus MaxLike: Empirical comparisons with ant species distributions. *Ecosphere*, *4*(5), 1–15. https://doi.org/10.1890/ES13-00066.1

Fitzpatrick, M. C., Preisser, E. L., Ellison, A. M., & Elkinton, J. S. (2009). Observer bias and the detection of low-density populations. *Ecological Applications*, *19*(7), 1673–1679. https://doi.org/10.1890/09-0265.1

Fourcade, Y., Engler, J. O., Rödder, D., & Secondi, J. (2014). Mapping species distributions with MAXENT using a geographically biased sample of presence data: A performance assessment of methods for correcting sampling bias. *PLoS One*, *9*(5), e97122. https://doi.org/10.1371/journal.pone.0097122

Gouraguine, A., Moranta, J., Ruiz-Frau, A., Hinz, H., Reñones, O., Ferse, S. C. A., Jompa, J., & Smith, D. J. (2019). Citizen science in data and resource-limited areas: A tool to detect long-term ecosystem changes. *PLoS One*, *14*(1), e0210007. https://doi.org/10.1371/journal.pone.0210007

Guisan, A., Zimmermann, N. E., Elith, J., Graham, C. H., Phillips, S., & Peterson, A. T. (2007). What matters for predicting the occurrences of trees: Techniques, data, or species' characteristics? *Ecological Monographs*, *77*(4), 615–630. https://doi.org/10.1890/06-1060.1

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer Science & Business Media.

Hijmans, R., Phillips, S., Leathwick, J., & Elith, J. (2017). *dismo: Species distribution modeling*. R package version 1.1-4. https://CRAN.R-project.org/package=dismo

Howard, C., Stephens, P. A., Pearce-Higgins, J. W., Gregory, R. D., & Willis, S. G. (2014). Improving species distribution models: The value of data on abundance. *Methods in Ecology and Evolution*, *5*(6), 506–513. https://doi.org/10.1111/2041-210X.12184

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*(10), 1052–1060. https://doi.org/10.1111/2041-210X.12254

Jiménez-Valverde, A., Aragón, P., & Lobo, J. M. (2021). Deconstructing the abundance–suitability relationship in species distribution modelling. *Global Ecology and Biogeography*, *30*, 327–338. https://doi.org/10.1111/geb.13204

Johnston, A., Fink, D., Reynolds, M. D., Hochachka, W. M., Sullivan, B. L., Bruns, N. E., Hallstein, E., Merrifield, M. S., Matsumoto, S., & Kelling, S. (2015). Abundance models improve spatial and temporal prioritization of conservation resources. *Ecological Applications*, *25*(7), 1749–1756. https://doi.org/10.1890/14-1826.1

Kadmon, R., Farber, O., & Danin, A. (2004). Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. *Ecological Applications*, *14*(2), 401–413. https://doi.org/10.1890/02-5364

Komori, O., Eguchi, S., Saigusa, Y., Kusumoto, B., & Kubota, Y. (2020). Sampling bias correction in species distribution models by quasi-linear Poisson point process. *Ecological Informatics*, *55*, 101015. https://doi.org/10.1016/j.ecoinf.2019.101015

Kosmala, M., Wiggins, A., Swanson, A., & Simmons, B. (2016). Assessing data quality in citizen science. *Frontiers in Ecology and the Environment*, *14*(10), 551–560. https://doi.org/10.1002/fee.1436

Kramer-Schadt, S., Niedballa, J., Pilgrim, J. D., Schröder, B., Lindenborn, J., Reinfelder, V., Stillfried, M., Heckmann, I., Scharf, A. K., Augeri, D. M., Cheyne, S. M., Hearn, A. J., Ross, J., Macdonald, D. W., Mathai, J., Eaton, J., Marshall, A. J., Semiadi, G., Rustam, R., … Wilting, A. (2013). The importance of correcting for sampling bias in MaxEnt species distribution models. *Diversity and Distributions*, *19*(11), 1366–1379. https://doi.org/10.1111/ddi.12096

Kuemmerlen, M., Schmalz, B., Guse, B., Cai, Q., Fohrer, N., & Jähnig, S. (2014). Integrating catchment properties in small scale species distribution models of stream macroinvertebrates. *Ecological Modelling*, *277*, 77–86. https://doi.org/10.1016/j.ecolmodel.2014.01.020

Lambert, D. (1992). Zero-Inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, *34*(1), 1–14. https://doi.org/10.2307/1269547

Lindén, A., & Mäntyniemi, S. (2011). Using the negative binomial distribution to model overdispersion in ecological count data. *Ecology*, *92*(7), 1414–1421. https://doi.org/10.1890/10-1831.1

Luoto, M., Virkkala, R., & Heikkinen, R. K. (2007). The role of land cover in bioclimatic models depends on spatial resolution. *Global Ecology and Biogeography*, *16*(1), 34–42. https://doi.org/10.1111/j.1466-8238.2006.00262.x

Lyashevska, O., Brus, D. J., & van der Meer, J. (2016). Mapping species abundance by a spatial zero-inflated Poisson model: A case study in the Wadden Sea, the Netherlands. *Ecology and Evolution*, *6*(2), 532–543.

Mair, L., & Ruete, A. (2016). Explaining spatial variation in the recording effort of citizen science data across multiple taxa. *PLoS One*, *11*(1), e0147796. https://doi.org/10.1371/journal.pone.0147796

Martin, T. G., Wintle, B. A., Rhodes, J. R., Kuhnert, P. M., Field, S. A., Low-Choy, S. J., Tyre, A. J., & Possingham, H. P. (2005). Zero tolerance ecology: Improving ecological inference by modelling the source of zero observations. *Ecology Letters*, *8*(11), 1235–1246. https://doi.org/10.1111/j.1461-0248.2005.00826.x

Meynard, C. N., & Kaplan, D. M. (2013). Using virtual species to study species distributions and model performance. *Journal of Biogeography*, *40*, 1–8. https://doi.org/10.1111/jbi.12006

Meynard, C. N., Leroy, B., & Kaplan, D. M. (2019). Testing methods in species distribution modelling using virtual species: What have we learnt and what are we missing? *Ecography*, *42*, 2021–2036. https://doi.org/10.1111/ecog.04385

Minami, M., Lennert-Cody, C. E., Gao, W., & Román-Verdesoto, M. (2007). Modeling shark bycatch: The zero-inflated negative binomial regression model with smoothing. *Fisheries Research*, *84*(2), 210–221. https://doi.org/10.1016/j.fishres.2006.10.019

Parnell, J. A. N., Simpson, D. A., Moat, J., Kirkup, D. W., Chantaranothai, P., Boyce, P. C., Bygrave, P., Dransfield, S., Jebb, M. H. P., Macklin, J., Meade, C., Middleton, D. J., Muasya, A. M., Prajaksood, A., Pendry, C. A., Pooma, R., Suddee, S., & Wilkin, P. (2003). Plant collecting spread and densities: Their potential impact on biogeographical studies in Thailand. *Journal of Biogeography*, *30*, 193–209. https://doi.org/10.1046/j.1365-2699.2003.00828.x

Pearce, J. L., & Boyce, M. S. (2006). Modelling distribution and abundance with presence-only data. *Journal of Applied Ecology*, *43*(3), 405–412. https://doi.org/10.1111/j.1365-2664.2005.01112.x

Pearce, J., & Ferrier, S. (2001). The practical value of modelling relative abundance of species for regional conservation planning: A case study. *Biological Conservation*, *98*(1), 33–43. https://doi.org/10.1016/S0006-3207(00)00139-7

Phillips, S. J. (2008). Transferability sample selection bias and background data in presence-only modelling: A response to Peterson et al (2007). *Ecography*, *31*(2), 272–278. https://doi.org/10.1111/j.0906-7590.2008.5378.x

Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, *190*(3), 231–259. https://doi.org/10.1016/j.ecolmodel.2005.03.026

Phillips, S. J., & Dudík, M. (2008). Modeling of species distributions with Maxent: New extensions and a comprehensive evaluation. *Ecography*, *31*(2), 161–175. https://doi.org/10.1111/j.0906-7590.2008.5203.x

Phillips, S. J., Dudik, 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. https://doi.org/10.1890/07-2153.1

Pocock, M. J. O., & Evans, D. M. (2014). The success of the Horse-Chestnut Leaf-Miner, *Cameraria ohridella*, in the UK revealed with hypothesis-led citizen science. *PLoS One*, *9*(1). https://doi.org/10.1371/journal.pone.0086226

R Core Team (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/

Reddy, S., & Dávalos, L. M. (2003). Geographical sampling bias and its implications for conservation priorities in Africa. *Journal of Biogeography*, *30*(11), 1719–1727. https://doi.org/10.1046/j.1365-2699.2003.00946.x

Renner, I. W., Elith, J., Baddeley, A., Fithian, W., Hastie, T., Phillips, S. J., Popovic, G., & Warton, D. I. (2015). Point process models for presence-only analysis. *Methods in Ecology and Evolution*, *6*, 366–379. https://doi.org/10.1111/2041-210X.12352

Ridout, M., Hinde, J., & Demétrio, C. G. B. (2001). A score test for testing a Zero-Inflated Poisson regression model Against Zero-Inflated negative binomial alternatives. *Biometrics*, *57*(1), 219–223. https://doi.org/10.1111/j.0006-341X.2001.00219.x

Rocchini, D., Hortal, J., Lengyel, S., Lobo, J. M., Jiménez-Valverde, A., Ricotta, C., Bacaro, G., & Chiarucci, A. (2011). Accounting for uncertainty when mapping species distributions: The need for maps of ignorance. *Progress in Physical Geography*, *35*(2), 211–226. https://doi.org/10.1177/0309133311399491

Rodríguez, J. P., Brotons, L., Bustamante, J., & Seoane, J. (2007). The application of predictive modelling of species distribution to biodiversity conservation. *Diversity and Distributions*, *13*(3), 243–251. https://doi.org/10.1111/j.1472-4642.2007.00356.x

Schmeller, D. S., Henry, P.-Y., Julliard, R., Gruber, B., Clobert, J., Dziock, F., Lengyel, S., Nowicki, P., Déri, E., Budrys, E., Kull, T., Tali, K., Bauch, B., Settele, J., Van swaay, C., Kobler, A., Babij, V., Papastergiadou, E., & Henle, K. (2009). Advantages of volunteer-based biodiversity monitoring in Europe. *Conservation Biology: The Journal of the Society for Conservation Biology*, *23*(2), 307–316. https://doi.org/10.1111/j.1523-1739.2008.01125.x

Sileshi, G., Hailu, G., & Nyadzi, G. I. (2009). Traditional occupancy–abundance models are inadequate for zero-inflated ecological count data. *Ecological Modelling*, *220*(15), 1764–1775. https://doi.org/10.1016/j.ecolmodel.2009.03.024

Smith, A. N. H., Anderson, M. J., & Millar, R. B. (2012). Incorporating the intraspecific occupancy–abundance relationship into zero-inflated models. *Ecology*, *93*(12), 2526–2532. https://doi.org/10.1890/12-0460.1

Sólymos, P., Lele, S., & Bayne, E. (2012). Conditional likelihood approach for analyzing single visit abundance survey data in the presence of zero inflation and detection error. *Environmetrics*, *23*(2), 197–205. https://doi.org/10.1002/env.1149

Sullivan, B. L., Wood, C. L., Iliff, M. J., Bonney, R. E., Fink, D., & Kelling, S. (2009). eBird: A citizen-based bird observation network in the biological sciences. *Biological Conservation*, *142*(10), 2282–2292. https://doi.org/10.1016/j.biocon.2009.05.006

Syfert, M. M., Smith, M. J., & Coomes, D. A. (2013). The effects of sampling bias and model complexity on the predictive performance of MaxEnt species distribution models. *PLoS One*, *8*(2), e55158. https://doi.org/10.1371/journal.pone.0055158

Thuiller, W., Lavorel, S., Sykes, M. T., & Araújo, M. B. (2006). Using niche-based modelling to assess the impact of climate change on tree functional diversity in Europe. *Diversity and Distributions*, *12*(1), 49–60. https://doi.org/10.1111/j.1366-9516.2006.00216.x

Tiago, P., Ceia-Hasse, A., Marques, T. A., Capinha, C., & Pereira, H. M. (2017). Spatial distribution of citizen science casuistic observations for different taxonomic groups. *Scientific Reports*, *7*(1), 1–9. https://doi.org/10.1038/s41598-017-13130-8

Tiago, P., Pereira, H. M., & Capinha, C. (2017). Using citizen science data to estimate climatic niches and species distributions. *Basic and Applied Ecology*, *20*, 75–85. https://doi.org/10.1016/j.baae.2017.04.001

Vollering, J., Halvorsen, R., & Mazzoni, S. (2019). The MIAmaxent R package: Variable transformation and model selection for species distribution models. *Ecology and Evolution*, *9*, 12051–12068. https://doi.org/10.1002/ece3.5654

Welsh, A. H., Cunningham, R. B., Donnelly, C. F., & Lindenmayer, D. B. (1996). Modelling the abundance of rare species: Statistical models for counts with extra zeros. *Ecological Modelling*, *88*(1), 297–308. https://doi.org/10.1016/0304-3800(95)00113-1

Wenger, S. J., & Freeman, M. C. (2008). Estimating species occurrence, abundance, and detection probability using Zero-Inflated distributions. *Ecology*, *89*(10), 2953–2959. https://doi.org/10.1890/07-1127.1

Williams, M. R., Yates, C. J., Stock, W. D., Barrett, G. W., & Finn, H. C. (2016). Citizen science monitoring reveals a significant, ongoing decline of the Endangered Carnaby's black-cockatoo *Calyptorhynchus latirostris*. *Oryx*, *50*(4), 626–635.

Wisz, M. S., Hijmans, R. J., Li, J., Peterson, A. T., Graham, C. H., Guisan, A., & NCEAS Predicting Species Distributions Working Group. (2008). Effects of sample size on the performance of species distribution models. *Diversity and Distributions*, *14*(5), 763–773. https://doi.org/10.1111/j.1472-4642.2008.00482.x

Zeileis, A., Kleiber, C., & Jackman, S. (2008). Regression models for count data in R. *Journal of Statistical Software*, *27*, 1–25.

Zuquim, G., Tuomisto, H., Jones, M. M., Prado, J., Figueiredo, F. O. G., Moulatlet, G. M., Costa, F. R. C., Quesada, C. A., & Emilio, T. (2014). Predicting environmental gradients with fern species composition

in Brazilian Amazonia. *Journal of Vegetation Science*, *25*(5), 1195–1207. https://doi.org/10.1111/jvs.12174

Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., & Smith, G. M. (2009). Zero-Truncated and Zero-Inflated models for count data. In *Mixed effects models and extensions in ecology with R* (pp. 261–293). Springer.

BIOSKETCH

Victoria Nolan is an ecologist with interests in distribution modelling and conservation at a macroecological scale, alongside an interest in large data and novel methods of statistical analysis.

Francis Gilbert is a professor of ecology with interests in distribution modelling and conservation, together with the evolution of ecological and behavioural attributes of organisms, with a specific focus on hoverflies.

Tom Reader is an associate professor with interests in ecology, behaviour and statistics, with particular focus on the evolution of animal signals such as Batesian mimicry and aposematism.

Author contributions: V.N. originally conceived of the presented idea and performed the main simulations and data analysis. T.R. and F.G. helped to refine the theory and simulations, verify the methods and provided input to the supervision of the project. The initial draft was written by V.N. with final draft input from all authors.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Nolan, V., Gilbert, F., & Reader, T. (2022). Solving sampling bias problems in presence–absence or presence-only species data using zero-inflated models. *Journal of Biogeography*, 49, 215–232. https://doi.org/10.1111/jbi.14268