Appendix S1: Additional figures and tables

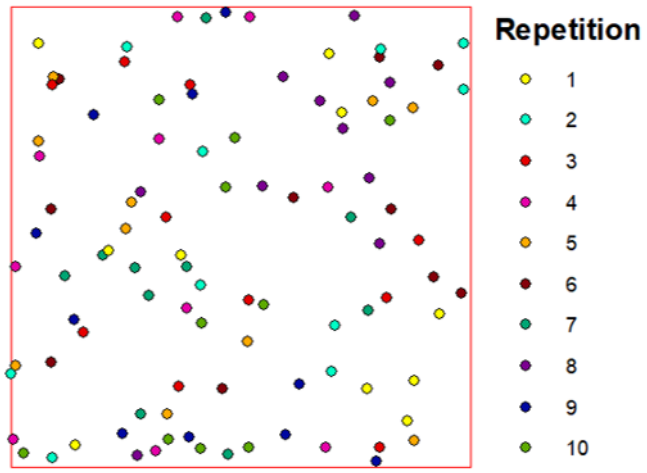

Figure S1.1. Spatial positions of the 10 randomly placed hypothetical 'town centres' across the simulation study area for each of the 10 simulation repetitions.

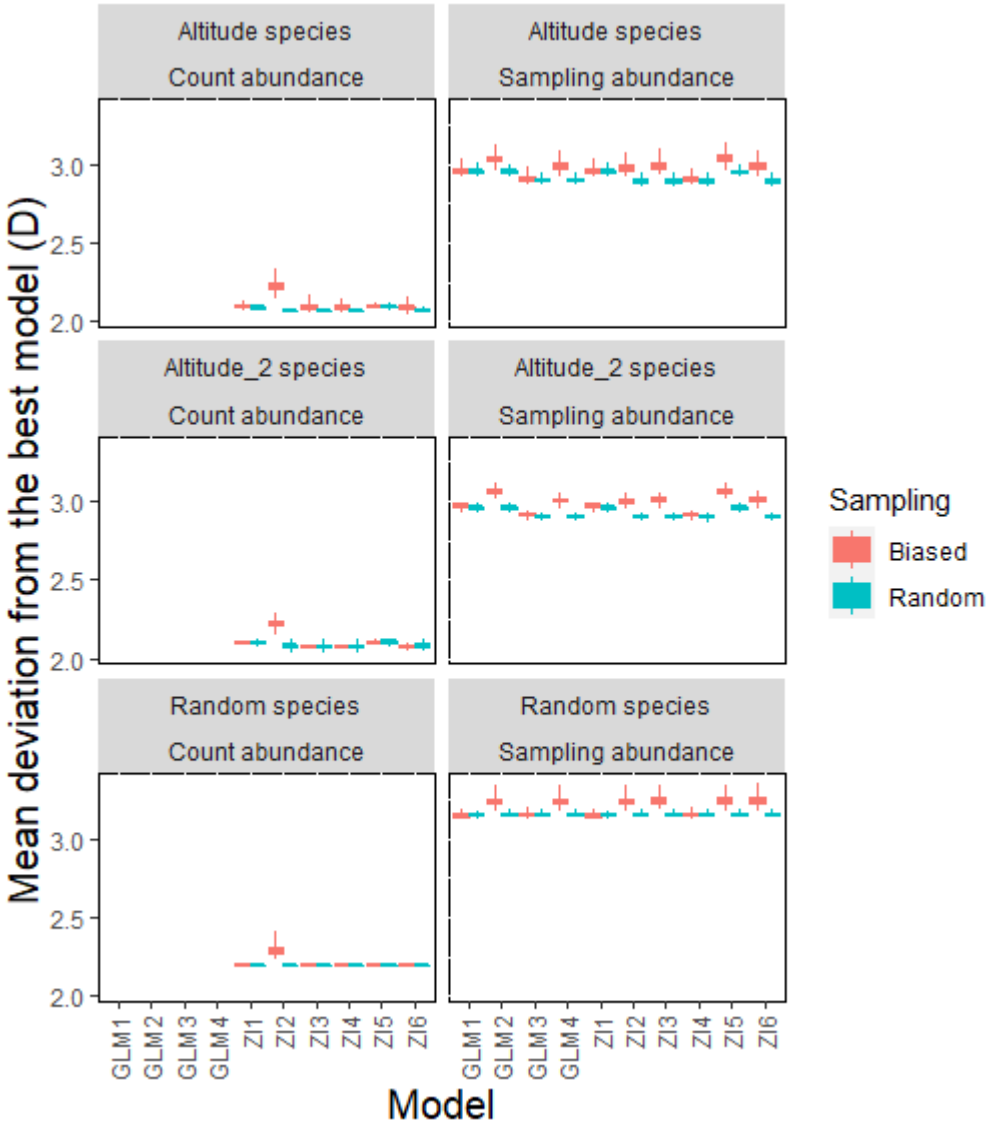

Figure S1.2. Evaluation of model predictions of abundance (based on D = 'deviation from the best model') for three hypothetical organisms (one with randomly simulated occurrences = random species, and two with occurrences simulated based on biological preferences = altitude species or altitude_randomised species (here termed altitude_2 species)). Mean D (± SE and data range) is shown for each sampling strategy (random or biased) across 10 different sets of hypothetical 'town centres' for each model. There are four non-zero-inflated generalised linear models, and six zero-inflated (ZI) models. For explanations of the structure of each model, see Tab. 3. Two types of prediction were evaluated: the count abundance predictions from the count component of the ZI models and the sampling abundance predictions from the whole of the ZI models or from the GLMs. Note the different scales on the vertical axes for the two types of predictions.

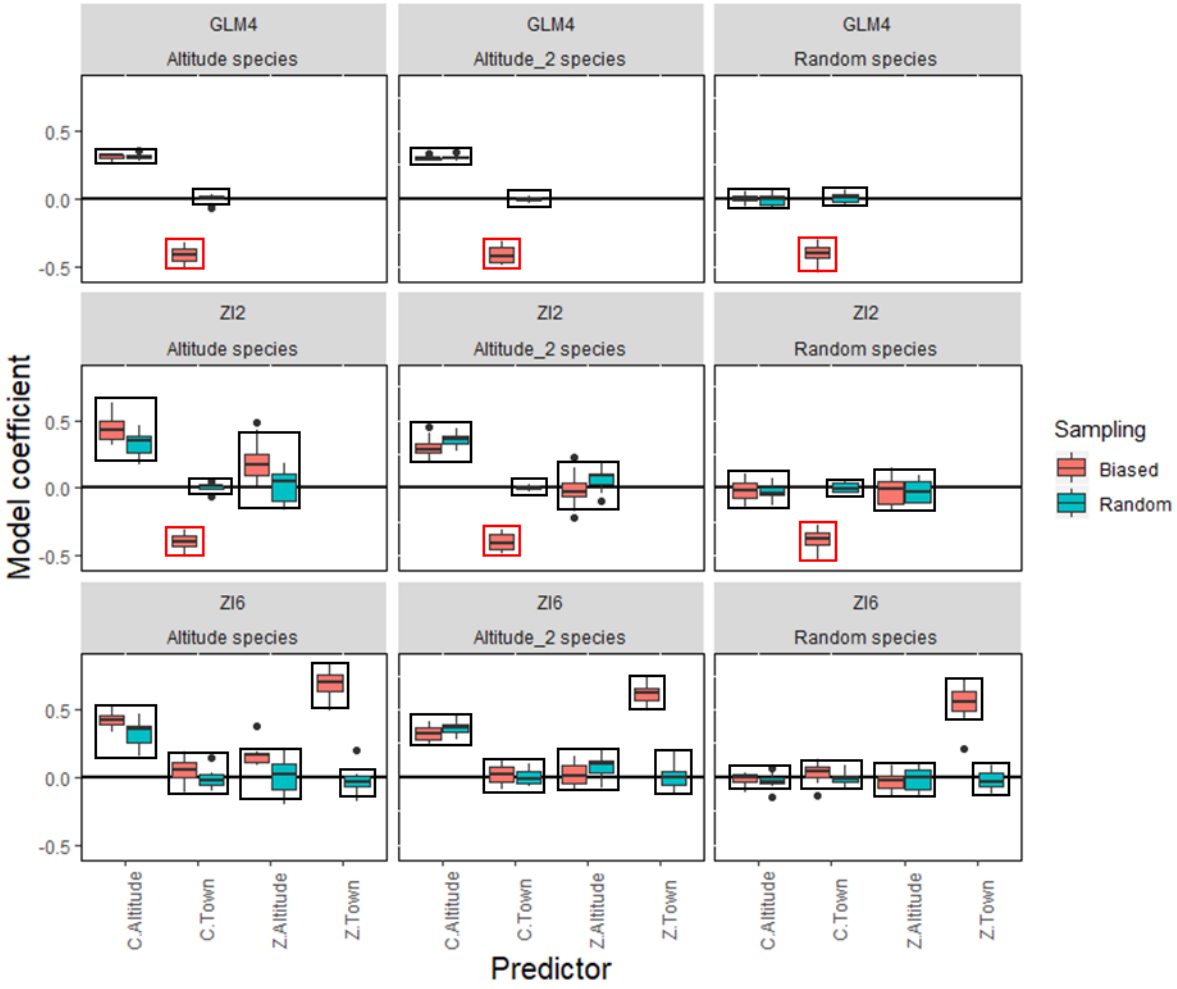

Figure S1.3. Model coefficients estimating the effects of a biological predictor (altitude or altitude_randomised (here termed altitude_2)) and a sampling bias predictor (distance to nearest hypothetical town) on the abundance of a hypothetical organism from a non-zero-inflated generalised linear model containing both the bias and biological predictor (GLM4), and two zero-inflated models which either exclude (ZI2) or include (ZI6) the bias predictor in the zero component. Zero-inflated (ZI) models include components which model both the count (C) of organisms per grid cell, and excess zeros (Z) caused by zero-inflation. For explanations of the structure of each model, see Tab. 3. Median model coefficients and range are shown for models fitted with data simulated using two different sampling strategies: random sampling and biased sampling. Results highlighted in red boxes indicate where the model is including the bias variable as a predictor of abundance where it should not. Black boxes are results that are correctly predicted.

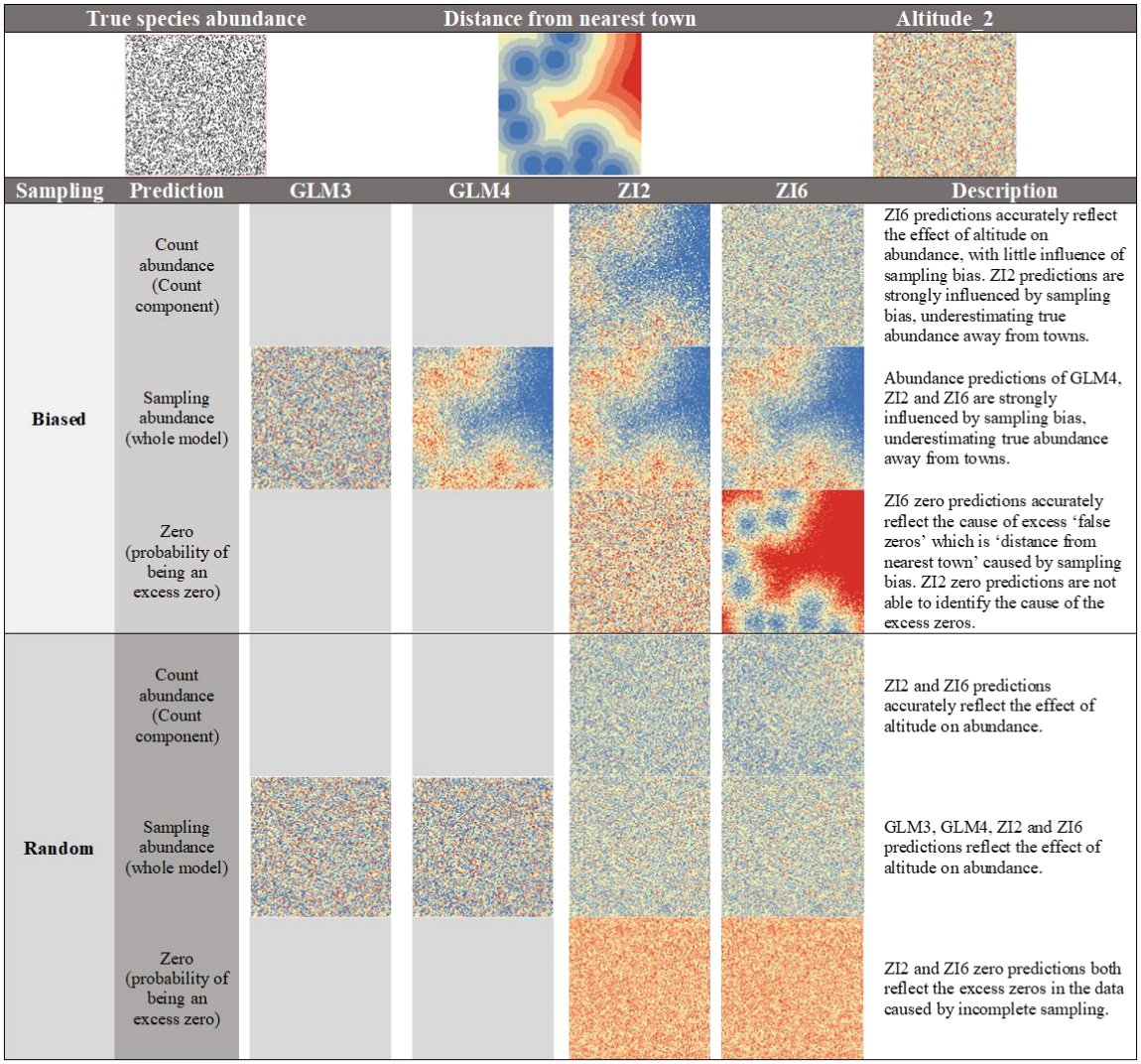

Figure S1.4. Example maps of abundance for a hypothetical species ('altitude_randomised species') whose occurrence is positively influenced by a randomised altitude layer, produced from two generalised linear models (GLMs) and two Zero-Inflated (ZI) models. Models were built with either data collected 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 predictor) 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).

Altitude species

Predictor	Model	Sampling: Random						Sampling: Biased					
		True abundance		Sampling abundance/ occurrence		Zero predictions		True abundance		Sampling abundance/ occurrence		Zero predictions	
		r	se	r	se	r	se	r	se	r	se	r	se
Altitude	GLM3			0.999	0.000					0.999	0.000		
	GLM4			0.992	0.003					0.603	0.052		
	ZI2	0.979	0.007	0.993	0.003	0.183	0.254	0.738	0.029	0.623	0.050	0.822	0.082
	ZI6	0.944	0.013	0.993	0.003	0.071	0.198	0.960	0.012	0.607	0.053	0.187	0.072
Distance from town	GLM3			-0.037	0.071					-0.037	0.070		
	GLM4			-0.046	0.082					-0.781	0.019		
	ZI2	-0.054	0.081	-0.042	0.080	0.016	0.060	-0.648	0.039	-0.765	0.020	0.002	0.054
	ZI6	0.069	0.091	-0.044	0.081	0.262	0.152	0.069	0.086	-0.769	0.021	0.958	0.009

Altitude_2 species

Predictor	Model	Sampling: Random						Sampling: Biased					
		True abundance		Sampling abundance/ occurrence		Zero predictions		True abundance		Sampling abundance/ occurrence		Zero predictions	
		r	se	r	se	r	se	r	se	r	se	r	se
Altitude	GLM3			0.999	0.000					0.999	0.000		
	GLM4			0.985	0.007					0.576	0.020		
	ZI2	0.980	0.005	0.986	0.007	0.497	0.194	0.565	0.042	0.593	0.021	-0.289	0.238
	ZI6	0.944	0.011	0.986	0.007	0.414	0.159	0.957	0.013	0.584	0.022	0.045	0.046
Distance from town	GLM3			-0.002	0.003					-0.003	0.003		
	GLM4			-0.057	0.043					-0.787	0.017		
	ZI2	-0.044	0.035	-0.053	0.043	0.002	0.004	-0.781	0.032	-0.773	0.018	0.002	0.004
	ZI6	-0.025	0.085	-0.052	0.042	0.012	0.188	0.056	0.072	-0.774	0.020	0.980	0.006

Figure S1.5. Spearman's Rank correlation coefficients (r_s) between the model predictors (altitude/ altitude_randomised (here termed altitude_2) and distance from nearest town) and model predictions under two sampling strategies (random and biased). The top panel represents results for altitude species, whereas the bottom panel represents results for altitude_randomised species. These predictions are either abundance predictions from the whole model (shown for the generalised linear models (GLMs), sampling abundance predictions from the zero-inflated (ZI) models, count abundance predictions of true abundance (shown for the ZI models) and predictions of the probability an observation is an excess zero (shown for the ZI models). GLM3 and the zero component of ZI2 do not include the bias predictor, whereas GLM4 and the zero component of ZI6 do contain the bias predictor. Values represent the mean coefficients (including standard error (se)) across the 10 simulated sets of 'town centres' Coefficients are colour-coded based on strength: the darker the colour, the stronger the correlation. Red values represent positive correlations, whereas blue represent negative correlations.

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	0.00	0.00	-0.02	-0.11	-0.21	-0.36	-0.73	-1.73	0.00	0.00	-0.02	-0.11	-0.22	-0.36	-0.75	-1.88	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.15
2000	0.00	0.00	-0.01	-0.09	-0.17	-0.31	-0.61	-1.49	0.00	0.00	-0.01	-0.09	-0.17	-0.31	-0.61	-1.53	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04
3000	0.00	0.00	-0.02	-0.08	-0.17	-0.32	-0.60	-1.49	0.00	0.00	-0.01	-0.08	-0.17	-0.32	-0.60	-1.52	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03
4000	0.00	0.00	-0.05	-0.09	-0.19	-0.34	-0.63	-1.51	0.00	0.00	-0.05	-0.09	-0.19	-0.34	-0.63	-1.52	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02
5000	0.00	0.00	-0.06	-0.25	-0.21	-0.38	-0.68	-1.59	0.00	0.00	-0.06	-0.25	-0.21	-0.38	-0.68	-1.61	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02
6000	0.00	-0.01	-0.06	-0.26	-0.25	-0.42	-0.74	-1.74	0.00	-0.01	-0.06	-0.26	-0.25	-0.42	-0.74	-1.75	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
7000	0.00	-0.01	-0.06	-0.26	-0.42	-0.48	-0.81	-1.87	0.00	-0.01	-0.06	-0.26	-0.42	-0.48	-0.81	-1.87	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
8000	0.00	-0.01	-0.06	-0.26	-0.42	-0.61	-0.88	-2.00	0.00	-0.01	-0.06	-0.26	-0.42	-0.61	-0.88	-2.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
9000	0.00	-0.01	-0.06	-0.26	-0.42	-0.63	-0.94	-2.08	0.00	-0.01	-0.06	-0.26	-0.42	-0.63	-0.94	-2.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10000	0.00	-0.01	-0.06	-0.26	-0.42	-0.63	-1.02	-2.19	0.00	-0.01	-0.06	-0.26	-0.42	-0.63	-1.02	-2.19	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	0.00	0.00	-0.01	-0.08	-0.15	-0.26	-0.49	-1.35	0.00	0.00	-0.01	-0.09	-0.27	-0.36	-0.68	-1.77	0.00	0.00	0.00	0.02	0.06	0.10	0.16	0.36
2000	0.00	0.00	-0.01	-0.09	-0.18	-0.32	-0.62	-1.56	0.00	0.00	-0.01	-0.11	-0.23	-0.42	-0.78	-1.92	0.01	0.01	0.00	0.00	0.00	0.10	0.16	0.36
3000	0.00	0.01	-0.01	-0.09	-0.18	-0.33	-0.61	-1.46	0.00	0.00	-0.03	-0.10	-0.22	-0.41	-0.74	-1.82	0.01	0.01	0.01	0.02	0.04	0.08	0.14	0.37
4000	0.01	0.01	-0.03	-0.10	-0.20	-0.36	-0.65	-1.55	0.00	0.00	-0.05	-0.16	-0.24	-0.42	-0.77	-1.90	0.01	0.01	0.02	0.06	0.04	0.06	0.12	0.35
5000	0.01	0.01	-0.05	-0.26	-0.39	-0.39	-0.70	-1.66	0.00	-0.01	-0.06	-0.26	-0.25	-0.45	-0.80	-1.88	0.01	0.02	0.02	0.04	0.03	0.06	0.09	0.22
6000	0.01	0.01	-0.05	-0.26	-0.30	-0.43	-0.76	-1.74	0.00	-0.01	-0.06	-0.26	-0.38	-0.48	-0.86	-2.00	0.01	0.02	0.01	0.00	0.08	0.06	0.09	0.26
7000	0.01	0.01	-0.06	-0.26	-0.42	-0.48	-0.82	-1.88	0.00	-0.01	-0.06	-0.26	-0.42	-0.52	-0.86	-1.99	0.01	0.01	0.00	0.00	0.00	0.04	0.05	0.11
8000	0.01	0.00	-0.06	-0.26	-0.42	-0.59	-0.88	-2.00	0.00	-0.01	-0.06	-0.26	-0.42	-0.62	-0.92	-2.08	0.01	0.01	0.00	0.00	0.00	0.03	0.04	0.08
9000	0.01	0.00	-0.06	-0.26	-0.42	-0.63	-0.98	-2.10	0.00	-0.01	-0.06	-0.26	-0.42	-0.63	-1.00	-2.15	0.01	0.00	0.00	0.00	0.00	0.00	0.02	0.04
10000	0.00	-0.01	-0.06	-0.26	-0.42	-0.63	-1.02	-2.19	0.00	-0.01	-0.06	-0.26	-0.42	-0.63	-1.02	-2.19	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Figure S1.6. Comparisons of model predictive power of sampling abundance (from the whole model) 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 predictor), 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.

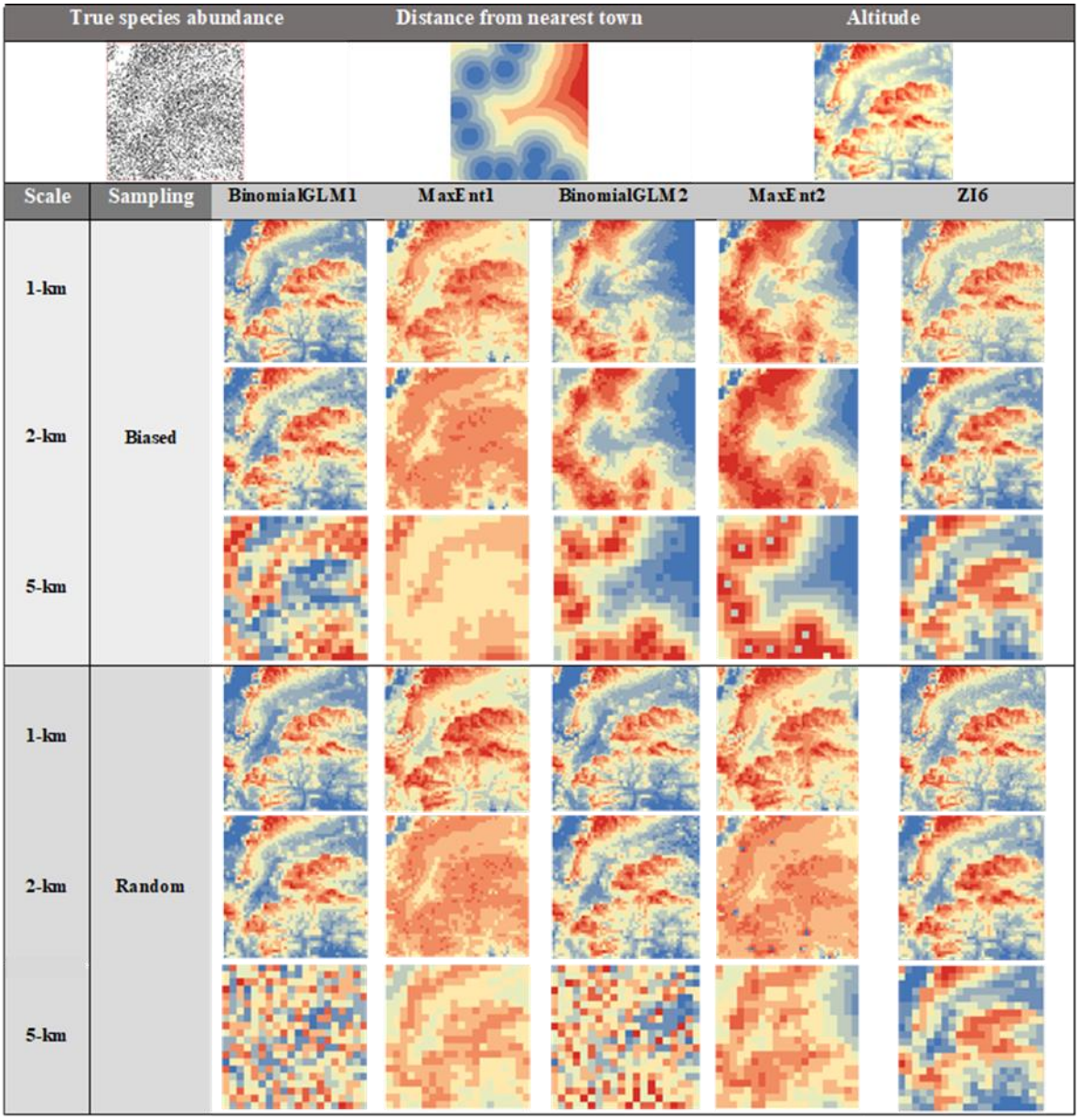

Figure S1.7. Example distribution maps for a hypothetical species whose occurrence is positively influenced by altitude (altitude species) from two binomial generalised linear models (GLMs) and two Maximum Entropy (MaxEnt) models. Maps are compared to maps of predicted abundance produced using the count abundance predictions (from the count component only) from a zero-inflated (ZI) model (ZI6) which inclues the bias in predictor in both components of the model. BinomialGLM1 and MaxEnt1 include only the biological predictor of altitude, whereas BinomialGLM2 and MaxEnt2 also include the bias predictor of distance from the nearest town. Unlike the zero-inflated (ZI) model, only one prediction can be obtained from the whole model and therefore will contain influences of sampling bias if present. Models were built with either data collected by random or biased sampling. Individual cells are colour coded based on abundance for the ZI abundance predictions or on probability of presence for the binomial GLM and MaxEnt predictions (high = red, low = blue).

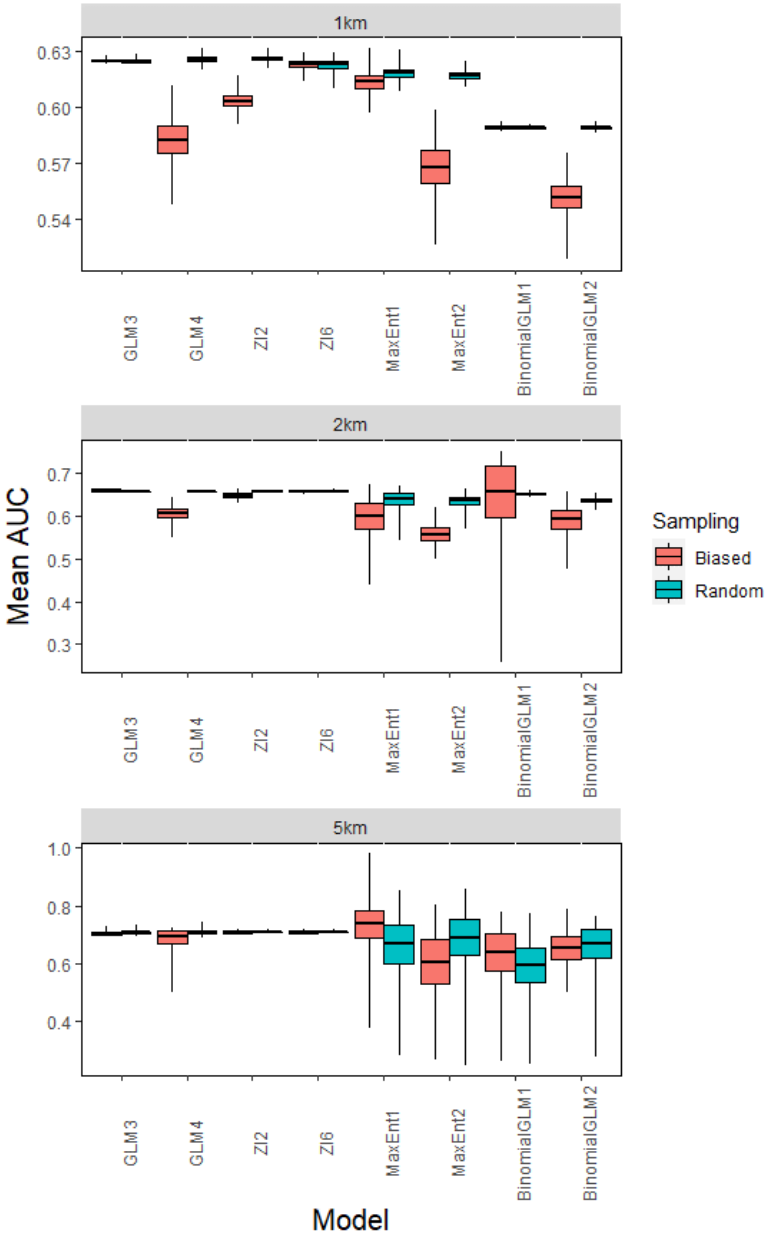

Figure S1.8. Evaluation of MaxEnt, generalised linear model (GLM) and zero-inflated (ZI) model predictions of altitude species presence-absence across the study area based on mean Area under the Curve (AUC) across three scales of data aggregation: 1-km, 2-km and 5-km. Mean AUC values (± SE and data range) for each sampling strategy (random or biased) across the 10 model repetitions are shown for two MaxEnt models and two binomial generalised linear models (GLMs): MaxEnt1 and Binomial-GLM1 which includes only altitude as a predictor, and MaxEnt2 and Binomial-GLM2 which includes altitude and distance from town as predictors. Abundance predictions were converted into binary presence-absence predictions for two non-zero-inflated generalised linear models (GLM3 including only the biological predictor and GLM4 including the biological and bias predictor) and two zero-inflated (ZI) models (ZI2 which does not account for bias in the zero component and ZI6 which does account for bias in the zero component) based on a threshold equal to the mean predicted abundance per model type (see Methods for more information).

100 *Table S1.1. Number of grid cells (at 1km^2 resolution) across the study area above each altitude threshold (m)*
101 *used for Simulation 2.*

Threshold (m)	Number of cells above threshold
0	10,000
50	9,068
100	5,329
125	3,364
150	1,993
175	1,036
200	396

102