Ecography

Appendix ECOG-04503
Vollering, J., Halvorsen, R., Auestad, I. and Rydgren, K. 2019. Bunching up the background betters bias in species distribution models. -- Ecography doi: 10.1111/ecog.04503

Appendix 1

Contents

A Detailed methodology: generating the virtual species and sampling bias

A virtual species distribution was created across an area spanning the northern Pacific coast of North America, similar to that used in the models of Sitka spruce. The virtual species' true probability of presence was defined as a logistic function of the linear combination of four BIOCLIM variables from the WorldClim data product (v. 2.0; Fick and Hijmans 2017): bio01 (mean annual temperature), bio02 (mean diurnal range), bio03 (isothermality), and bio12 (annual precipitation). These four variables were chosen because they tend to capture different dimensions of climatic variation (Appendix B, fig. A.3). The function for probability of presence defined a unimodal relationship to mean annual temperature, positive linear relationships to isothermality and annual precipitation, and a negative linear relationship to mean diurnal range. Parameter values were chosen to create a spatial distribution similar to the spatial distribution of Sitka spruce presences (Fig. A.1, top panel). To transform the true probability of presence into a particular realization of binary presences and absences, i.e. the true occurrence distribution, we sampled the Bernoulli distribution with the logistic values as probabilities (Fig. A.1, bottom panel).

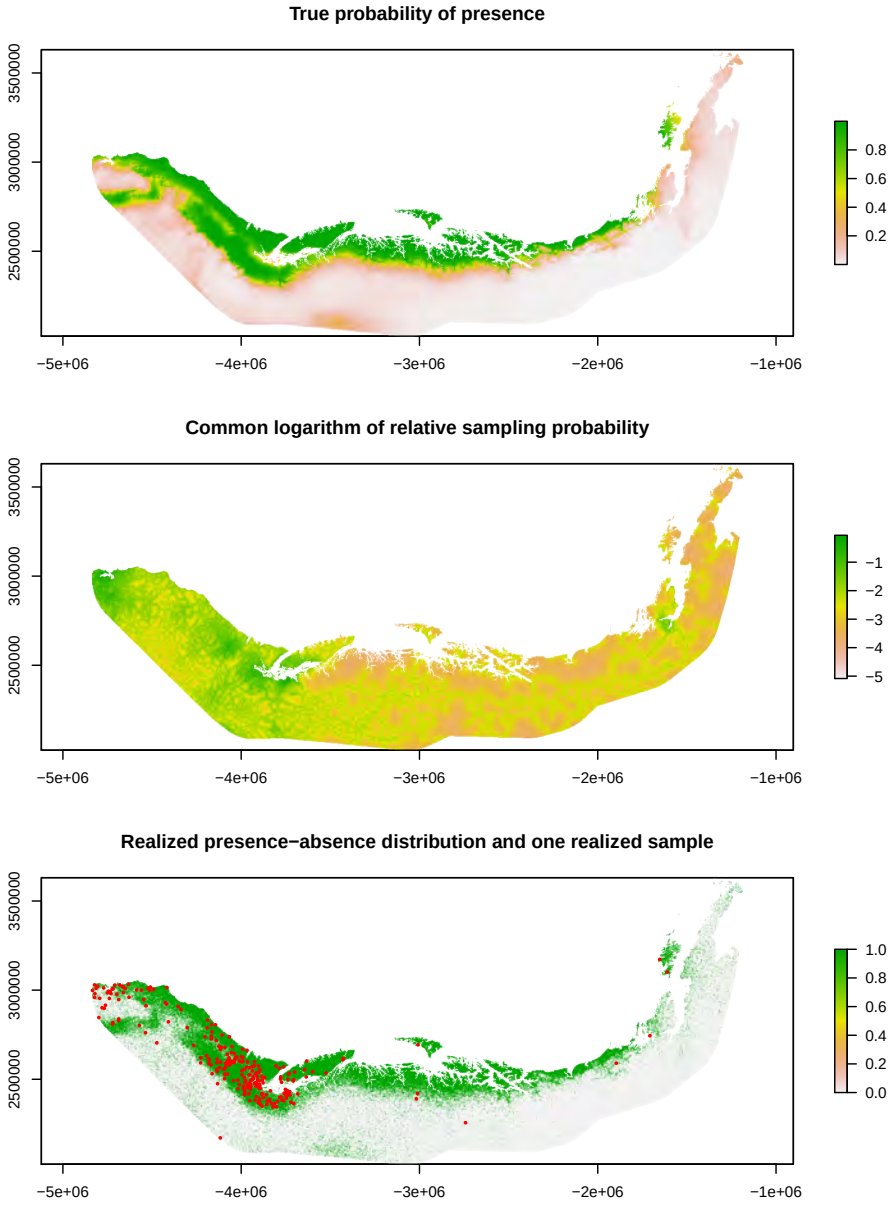

Figure A.1: Maps illustrating the virtual species simulation. The true probability of presence (top), common logarithm of relative sampling probability (middle), and realized presence-absence distribution, showing one sample of 250 presence locations as red points (bottom). Note that the maps are in the Lambert Azimuthal Equal Area projection (ESRI:102017), such that the north-south axis runs approximately right to left.

Next, a virtual sampling bias distribution was created based on real data on human population density and proximity to roads. The population density surface was derived from census data in the United States and in Canada. For the USA, 2010 Census data were downloaded for Alaska, Washington, Oregon, and California, and population density (persons per sq. mile) for each Census Area or Borough (AK), or County (WA, OR, CA) was joined to the appropriate spatial polygon. For Canada, 2011 Census data were downloaded and population densities for British Columbia and Yukon census divisions were joined to the appropriate spatial polygons. To create a smoothed surface of population density, which is likely to represent sampling probability more realistically, population density was extracted from the polygons on a point grid and point values were spatially interpolated. The grid spacing was set to 10 km to limit the number of points but maintain sufficient coverage for each census polygon. Spatial interpolation was performed using kernel smoothing in the Geostatistical Analyst in ArcGIS (v. 10.5), with an exponential kernel, a bandwidth of 50 km, and otherwise default parameters. Lastly, the smoothed population density surface was converted to raster format with the same grain and extent as the WorldClim climate data.

Proximity to roads was quantified by downloading national-level spatial data (line features) for major roads in the USA and Canada from `www.diva-gis.org`. The Euclidean Distance tool in ArcGIS was used to create a raster with the same grain and extent as the WorldClim climate data, with values representing distance to road. The multiplicative inverse of distance to road was taken to obtain proximity to road. The final sampling probability distribution used in the simulation was created as the product of linearly rescaled population density and proximity to road, raised to the power of 0.75 to lessen the dominance of urban areas (Fig. A.1, middle panel).

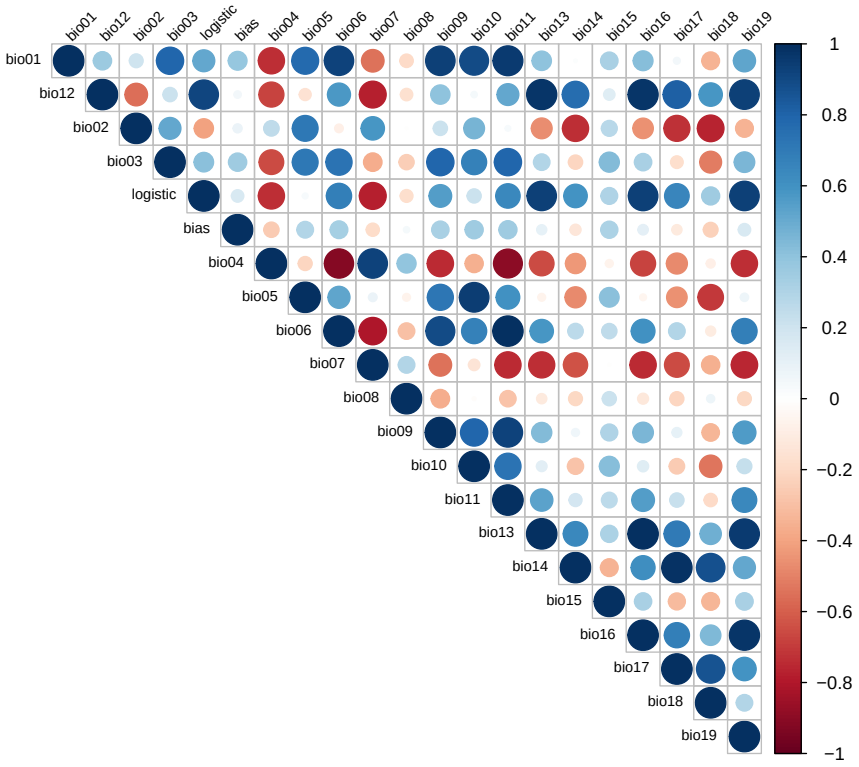

Figure A.2: Correlations matrix including the four BIOCLIM variables used to generate the virtual species, the true distribution (logistic), the sampling probability distribution (bias), and the 15 BIOCLIM variables used as explanatory variables in the models.

B Detailed methodology: building models of Sitka spruce

B.1 Presence data

Sitka spruce occurrence data obtained from the Global Biodiversity Information Facility were reviewed manually. We revised coordinates with precision less than 0.1 decimal degree and coordinates not in accordance with their named locality, if the named locality could be identified with sub-kilometer precision using Google Earth (v. 7.1) and gazetteers. Otherwise these records were discarded. Occurrences identified as planted, for example from botanical gardens, were also discarded. Coordinates falling outside the coverage of the gridded climate data — due to proximity to the coast — were shifted to the nearest data cell using ArcGIS (v. 10.5). Of the 455 presence records obtained from the query for North American Sitka spruce records, 426 remained after manual review, which resulted in 243 rasterized presence observations.

B.2 Climate data

The 19 BIOCLIM variables were obtained as raster data with global land coverage at 30 arc-second resolution from two data products: WorldClim (v. 2.0; Fick and Hijmans 2017) and CHELSA (v. 1.2; Karger et al. 2017). Climate data products are most uncertain around significant topographic features, near coastlines, and in polar regions (Daly 2006; Morales-Barbero and Vega-Álvarez 2018), all of which are prevalent in the study area. The WorldClim product is the result of thin-spline interpolation of weather station data using topographic and satellite-derived predictors, for the period 1970–2000. The CHELSA product is the result of a statistical downscaling of a general circulation model reanalysis, with orographic predictors and a bias correction used in the precipitation data, for the period 1979–2013. The largest differences in variable estimates from these different methodologies are expected to occur for precipitation in mountainous areas with low weather station density (Bobrowski and Schickhoff 2017; Karger et al. 2017). A small practical difference between the two is that the CHELSA product tends to have fewer no-data cells than the WorldClim product near coastlines. All climate data were resampled to an equal-area raster with 1 km grain size (Budic et al. 2016), using bilinear interpolation in ArcGIS (v. 10.5).

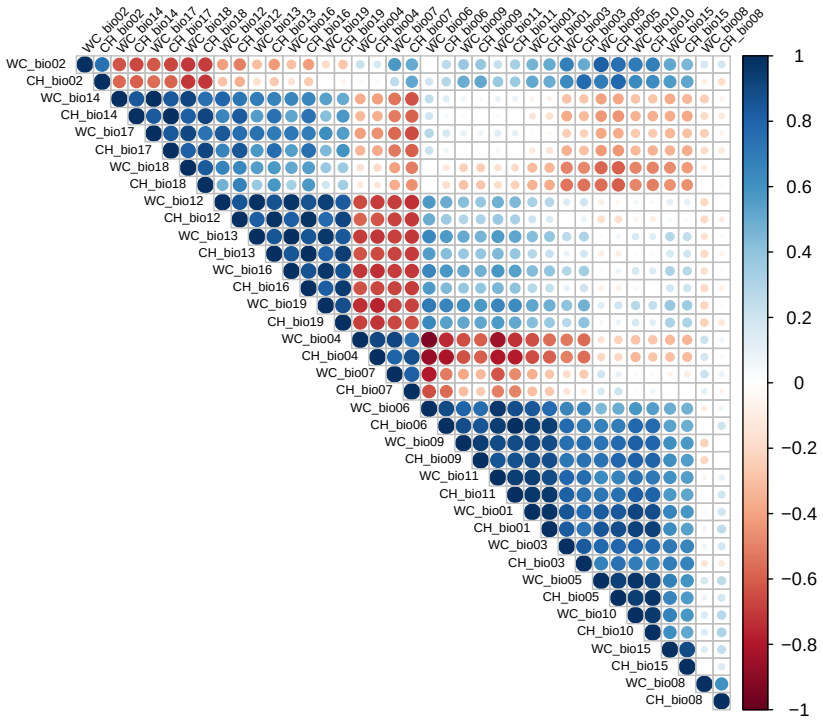

Figure A.3: Matrix showing the correlation structure among WorldClim and CHELSA's BIOCLIM variables in the study area.

B.3 Ecological model

We used an ecological model (*sensu* Austin 2002) grounded in concepts from gradient analysis of vegetation (Whittaker 1967) to underpin distribution models of Sitka spruce. This theoretical foundation recognizes that most environmentally controlled variation in species composition — or variation in performance, in the case of a single species — can be explained by just a few complex-gradients (Halvorsen 2012). Thus, it suggests that a limited number of explanatory variables should be included in the distribution model, especially when the full set of explanatory variables subject to selection consists of more or less collinear variables.

Furthermore, the gradient analysis literature shows (1) that species responses to complex-gradients are generally unimodal, (2) that the way gradients are

scaled will affect the symmetry of the unimodal response, and (3) that the observed portion of the unimodal response is often truncated by sampling (Austin 2007; Halvorsen 2012). These observations prescribe the types of relationships that the model should be able to capture and demarcates what should be considered a realistic model. Specifically, any shape obtainable through a particular truncation or scaling of a unimodal response may be considered sensible. This expectation was operationalized by using transformations of explanatory variables ("derived variables") as predictors, and rejecting univariate responses that showed local minima (Halvorsen 2013; Halvorsen et al. 2015).

B.4 Model building

Statistically, models were fitted based on the principle of maximum entropy, implying that their predictions give "the least biased estimate possible on the given information" (Jaynes 1957a,b). In the context of distribution modeling, this is interpreted most easily as giving predictions which are as close as possible to geographically uniform, subject to particular constraints (Merow et al. 2013). Specifically, for all predictors entering into the model, the mean of the predicted probability distribution (i.e. expected value) is constrained to match the empirical mean among presences (Phillips et al. 2006). Under this condition, the maximum entropy model takes the form of an exponential-family function called a Gibbs distribution (Phillips et al. 2006). This model may alternatively be interpreted as an inhomogeneous Poisson process model, or as an infinitely weighted logistic regression model (Fithian and Hastie 2013). In practice, models of Sitka spruce were fitted by infinitely weighted logistic regression, by regressing presence-background data on the derived variables (Vollering et al. 2018).

Up until recently, variable selection in maximum entropy distribution models has been performed nearly exclusively using shrinkage methods — specifically lasso regularization — whereby the coefficients of predictors with lower explanatory power are reduced, potentially to zero (Phillips et al. 2006; Halvorsen 2013). An alternative is to perform subset variable selection, whereby a discrete set of predictors that each carries explanatory power above a certain threshold is retained. In simulations, this alternative has been shown to outperform shrinkage methods when a minority of candidate predictors truly affect the distribution and the sample size is sufficiently large (Reineking and Schröder 2006). When modeling real distributions, subset variable selection seems to produce more parsimonious models than lasso regularization (Halvorsen et al. 2016).

For the above reasons, variable selection for Sitka spruce models was carried out by subset selection in a two-stage procedure outlined by Halvorsen et al. (2015) and implemented in the "MIAmaxent" R package (v. 1.0.0; Vollering et al. 2018). First, a parsimonious group of derived variables was selected from those created for each explanatory variable, to serve as an integral set representing the explanatory variable; thereafter, these integral sets of derived variables functioned as inseparable units and a second round of variable selection picked among them. For each combination of climate product and sampling bias cor-

rection approach, an identical variable selection procedure was applied to arrive at the final model, as follows:

1. Derived variable selection and subsequent explanatory variable selection was performed by nested model comparison using the chi-squared test under a level of significance (α) of 0.01. Interaction terms were not allowed.

2. If one or more of the selected explanatory variables showed an unrealistic single-effect response curve — defined as a response curve with one or more local minima — then:

   (a) Derived variable selection was performed under $\alpha = 1\text{x}10^{-3}$, to reduce the complexity of the explanatory variable responses, followed by explanatory variable selection under $\alpha = 0.01$.

   (b) If an explanatory variable still showed an unrealistic response, then this explanatory variable was dropped from the pool of candidate explanatory variables, and the selection procedure was repeated from step 1 until all response curves conformed with expectations. If multiple explanatory variables showed unrealistic responses, the explanatory variable with the highest peaks bounding the local minimum was dropped.

It should be noted that exploratory analyses showed that an alpha value of 0.01 resulted in models that explained more deviance per derived variable than values of $\alpha = 1\text{x}10^{-3}$ or $\alpha = 1\text{x}10^{-6}$, such that this appeared to be an suitable threshold level.

C Virtual species simulation: thickening radius

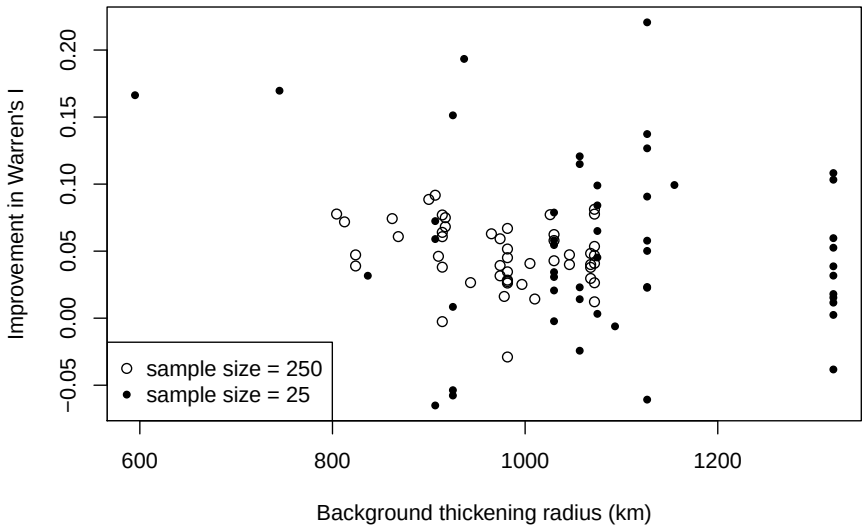

Figure A.4: Variation in the efficacy of background thickening with the length of the background thickening radius. The efficacy of background thickening is measured by the improvement in Warrens I compared to no correction. Each point represents one of the 100 presence samples used in the simulation.

D Sitka spruce case study: individual model details

D.1 No correction, WorldClim

Training data map

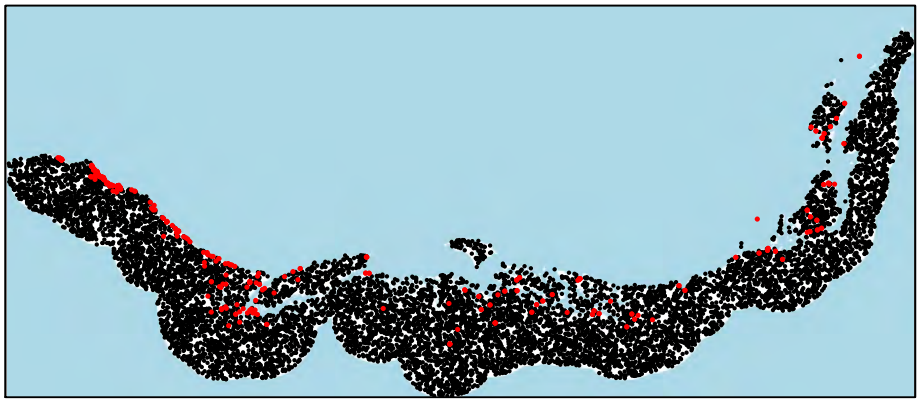

Figure A.5: Presence locations (red) and uninformed background locations (black) plotted across the training area. The map projection is as used during modeling: the Lambert Azimuthal Equal Area projection (ESRI:102017).

Stepwise forward selection of explanatory variables

round	variables	m	Dsq	Chisq	df	P
1	bio04	1	0.158	719.674	1	1.58e-158
1	bio06	1	0.146	661.296	1	7.81e-146
1	bio07	2	0.141	642.493	2	3.05e-140
1	bio11	1	0.099	449.521	1	9.17e-100
1	bio09	2	0.094	426.168	2	2.88e-93
1	bio01	2	0.09	408.713	2	1.77e-89
1	bio03	2	0.09	407.179	2	3.82e-89
1	bio10	2	0.087	394.158	2	2.57e-86
1	bio05	3	0.062	283.888	3	3.05e-61
1	bio08	3	0.056	255.636	3	3.95e-55

round	variables	m	Dsq	Chisq	df	P
1	bio15	2	0.051	232.026	2	4.13e-51
1	bio19	1	0.049	222.963	1	2.04e-50
1	bio16	1	0.046	210.979	1	8.4e-48
1	bio13	1	0.045	205.093	1	1.62e-46
1	bio12	1	0.039	179.153	1	7.42e-41
1	bio18	3	0.023	103.118	3	3.32e-22
1	bio02	1	0.02	92.236	1	7.69e-22
1	bio14	1	0.002	9.383	1	0.00219
2	bio04 + bio15	3	0.17	52.91	2	3.24e-12
2	bio04 + bio02	2	0.166	35.19	1	2.99e-09
2	bio04 + bio10	3	0.165	31.366	2	1.55e-07
2	bio04 + bio05	4	0.165	29.911	3	1.44e-06
2	bio04 + bio09	3	0.162	16.793	2	0.000226
2	bio04 + bio08	4	0.163	18.728	3	0.000311
2	bio04 + bio07	3	0.161	11.874	2	0.00264
2	bio04 + bio14	2	0.16	8.03	1	0.0046
2	bio04 + bio03	3	0.16	8.044	2	0.0179
2	bio04 + bio19	2	0.16	4.895	1	0.0269
2	bio04 + bio11	2	0.159	3.377	1	0.0661
2	bio04 + bio12	2	0.159	1.554	1	0.213
2	bio04 + bio18	4	0.159	4.355	3	0.226
2	bio04 + bio16	2	0.159	1.266	1	0.26
2	bio04 + bio06	2	0.159	1.212	1	0.271
2	bio04 + bio13	2	0.159	0.882	1	0.348
2	bio04 + bio01	3	0.159	2.093	2	0.351
3	bio04 + bio15 + bio08	6	0.176	25.116	3	1.46e-05
3	bio04 + bio15 + bio10	5	0.174	19.23	2	6.67e-05
3	bio04 + bio15 + bio05	6	0.174	19.472	3	0.000218
3	bio04 + bio15 + bio02	4	0.173	12.583	1	0.000389
3	bio04 + bio15 + bio14	4	0.171	3.958	1	0.0467
3	bio04 + bio15 + bio09	5	0.171	2.222	2	0.329

round	variables	m	Dsq	Chisq	df	P
3	bio04 + bio15 + bio07	5	0.171	2.075	2	0.354
4	*bio04 + bio15 + bio08 + bio10*	*8*	*0.182*	*29.558*	*2*	*3.82e-07*[1]
4	bio04 + bio15 + bio08 + bio05	9	0.181	26.225	3	8.56e-06
4	bio04 + bio15 + bio08 + bio02	7	0.179	15.448	1	8.48e-05
5	bio04 + bio15 + bio08 + bio10 + bio02	9	0.183	5.41	1	0.02
5	bio04 + bio15 + bio08 + bio10 + bio05	11	0.182	0.45	3	0.93

Model parameters

A vector of model predictions in Probability Ratio Output, denoted $\dot{q}$, is calculated as follows:

$$\dot{q}_i = N \cdot e^{\alpha + \sum_{k=1}^{m} \beta_k x_{ik}}$$

where N is the number of background locations (presences plus uninformed background locations), α is a normalizing constant, β is a vector of coefficients, and x is a matrix of derived variables.

The names of the coefficients in the table below are made up of the name of the explanatory variable followed by an extension that denotes the type of transformation applied to obtain the derived variable. The extensions follow defaults in Vollering et al. 2018.

parameter	value
N	10243
α	-13.09555709
bio04_M	-7.623956034
bio15_D05	2.116578274
bio15_HF17	-3942.554924
bio08_D1	4.322301555
bio08_HF16	-52.19756051
bio08_L	11.6014134
bio10_D1	-5.653978695

[1] selected model

parameter	value
bio10_M	-1.790029681

Response curves

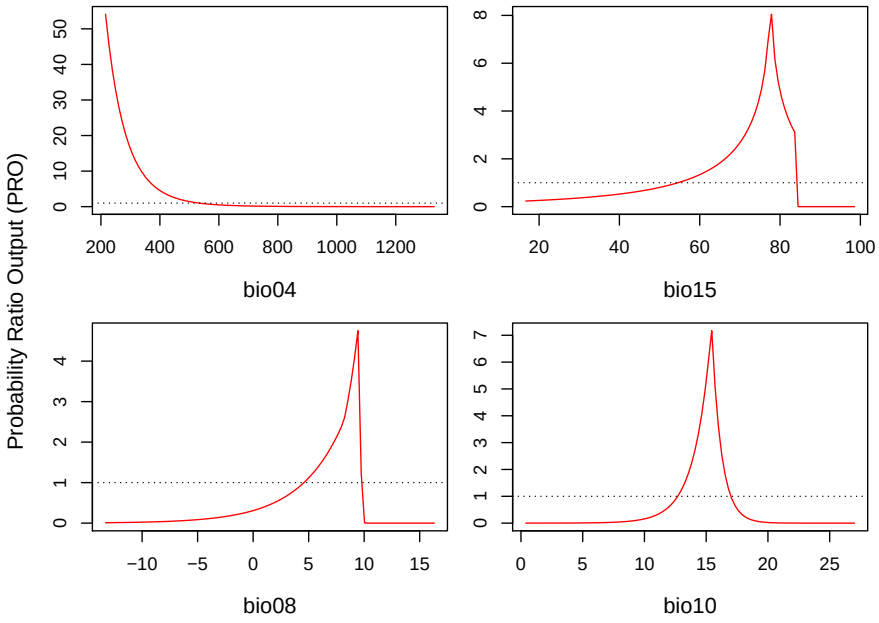

Figure A.6: Single-effect response curves of the model

D.2 No correction, CHELSA

Training data map

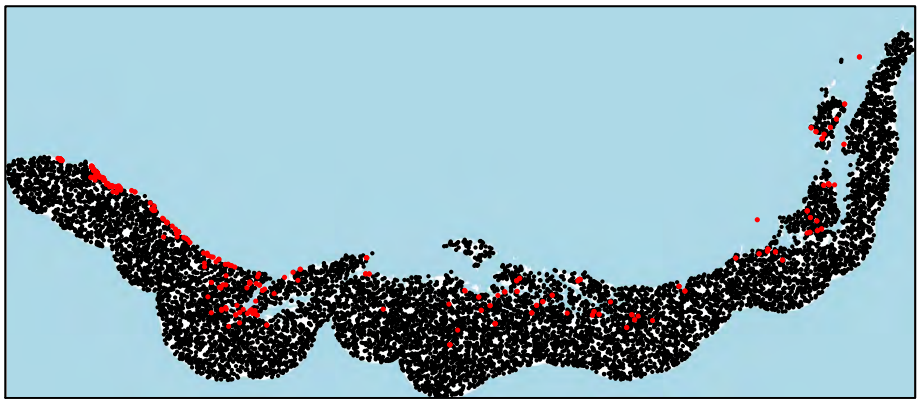

Figure A.7: Presence locations (red) and uninformed background locations (black) plotted across the training area. The map projection is as used during modeling: the Lambert Azimuthal Equal Area projection (ESRI:102017).

Stepwise forward selection of explanatory variables

round	variables	m	Dsq	Chisq	df	P
1	bio06	1	0.142	642.902	1	7.81e-142
1	bio04	2	0.131	594.287	2	8.96e-130
1	bio11	1	0.114	518.345	1	9.69e-115
1	bio07	1	0.107	485.679	1	1.24e-107
1	bio01	3	0.098	444.179	3	5.95e-96
1	bio10	3	0.081	367.764	3	2.12e-79
1	bio09	1	0.073	332.913	1	2.23e-74
1	bio05	3	0.057	258.373	3	1.01e-55
1	bio19	1	0.049	223.054	1	1.95e-50
1	bio13	2	0.048	220.125	2	1.59e-48
1	bio16	1	0.047	214.045	1	1.8e-48
1	bio12	2	0.046	207.952	2	6.98e-46
1	bio02	3	0.028	126.211	3	3.54e-27

round	variables	m	Dsq	Chisq	df	P
1	bio18	2	0.019	86.834	2	1.39e-19
1	bio14	2	0.006	25.379	2	3.08e-06
2	bio06 + bio02	4	0.174	148.428	3	5.75e-32
2	bio06 + bio05	4	0.163	95.867	3	1.2e-20
2	bio06 + bio11	2	0.159	79.585	1	4.62e-19
2	bio06 + bio10	4	0.16	84.413	3	3.47e-18
2	bio06 + bio01	4	0.16	82.214	3	1.03e-17
2	bio06 + bio07	2	0.156	66.702	1	3.16e-16
2	bio06 + bio18	3	0.154	57.272	2	3.66e-13
2	bio06 + bio14	3	0.154	56.809	2	4.61e-13
2	bio06 + bio12	3	0.151	42.869	2	4.91e-10
2	bio06 + bio04	3	0.148	29.59	2	3.75e-07
2	bio06 + bio16	2	0.147	22.863	1	1.74e-06
2	bio06 + bio13	3	0.146	18.961	2	7.63e-05
2	bio06 + bio19	2	0.144	10.581	1	0.00114
2	bio06 + bio09	2	0.142	2.089	1	0.148
3	bio06 + bio02 + bio10	7	0.18	27.114	3	5.57e-06
3	bio06 + bio02 + bio04	6	0.18	24.046	2	6e-06
3	bio06 + bio02 + bio12	6	0.179	21.062	2	2.67e-05
3	bio06 + bio02 + bio01	7	0.179	22.55	3	5.01e-05
3	bio06 + bio02 + bio13	6	0.177	13.747	2	0.00103
3	bio06 + bio02 + bio16	5	0.176	9.081	1	0.00258
3	bio06 + bio02 + bio05	7	0.177	12.651	3	0.00546
3	bio06 + bio02 + bio14	6	0.175	5.449	2	0.0656
3	bio06 + bio02 + bio07	5	0.175	3.074	1	0.0796
3	bio06 + bio02 + bio19	5	0.175	3.052	1	0.0807
3	bio06 + bio02 + bio11	5	0.175	1.768	1	0.184
3	bio06 + bio02 + bio18	6	0.175	1.484	2	0.476
4	bio06 + bio02 + bio10 + bio12	9	0.184	15.935	2	0.000347
4	bio06 + bio02 + bio10 + bio01	10	0.183	14.208	3	0.00263

round	variables	m	Dsq	Chisq	df	P
4	bio06 + bio02 + bio10 + bio05	10	0.183	14.045	3	0.00284
4	bio06 + bio02 + bio10 + bio16	8	0.182	7.903	1	0.00494
4	bio06 + bio02 + bio10 + bio13	9	0.182	7.884	2	0.0194
4	bio06 + bio02 + bio10 + bio04	9	0.182	6.22	2	0.0446
5	*bio06 + bio02 + bio10 + bio12 + bio05*	*12*	*0.187*	*16.853*	*3*	*0.000758*[2]
5	bio06 + bio02 + bio10 + bio12 + bio01	12	0.186	10.183	3	0.0171
5	bio06 + bio02 + bio10 + bio12 + bio16	10	0.184	2.295	1	0.13

Model parameters

A vector of model predictions in Probability Ratio Output, denoted $\dot{q}$, is calculated as follows:

$$\dot{q}_i = N \cdot e^{\alpha + \sum_{k=1}^{m} \beta_k x_{ik}}$$

where N is the number of background locations (presences plus uninformed background locations), α is a normalizing constant, β is a vector of coefficients, and x is a matrix of derived variables.

The names of the coefficients in the table below are made up of the name of the explanatory variable followed by an extension that denotes the type of transformation applied to obtain the derived variable. The extensions follow defaults in Vollering et al. 2018.

parameter	value
N	10243
α	-10.41059894
bio06_D05	-4.124739923
bio02_HF11	-4.320127444
bio02_HR3	0.655044182
bio02_HF14	-21.10690146
bio10_D05	-5.581766522

[2]selected model

parameter	value
bio10_M	-685.4582259
bio10_L	713.8110925
bio12_D2	-7.158134136
bio12_HR4	0.698618265
bio05_D05	1.565778989
bio05_L	49.06017132
bio05_M	-72.07752552

Response curves

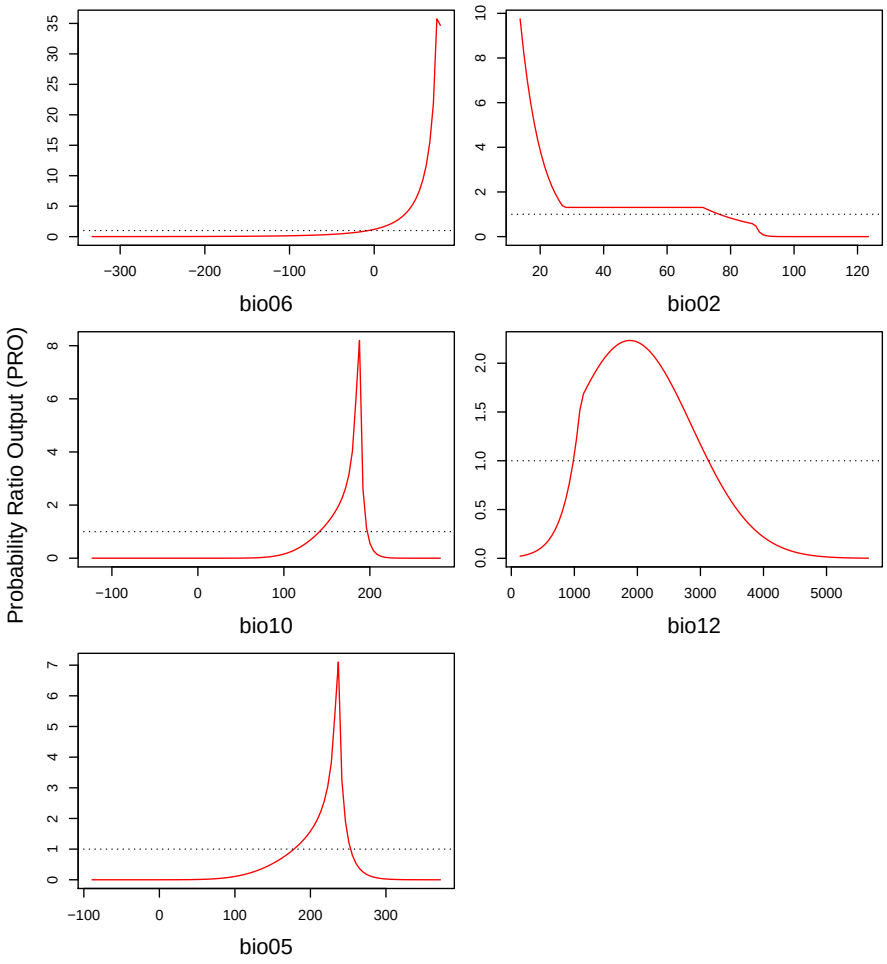

Figure A.8: Single-effect response curves of the model

D.3 Target group background selection, WorldClim

Training data map

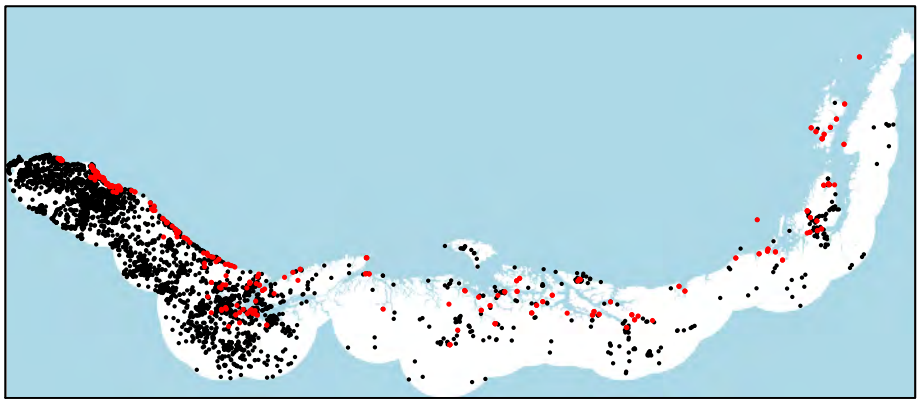

Figure A.9: Presence locations (red) and uninformed background locations (black) plotted across the training area. The map projection is as used during modeling: the Lambert Azimuthal Equal Area projection (ESRI:102017).

Stepwise forward selection of explanatory variables

round	variables	m	Dsq	Chisq	df	P
1	bio07	1	0.068	277.839	1	2.22e-62
1	bio02	1	0.052	213.434	1	2.45e-48
1	bio04	1	0.043	173.674	1	1.17e-39
1	bio06	1	0.04	163.349	1	2.1e-37
1	bio05	1	0.038	155.701	1	9.84e-36
1	bio10	2	0.037	151.298	2	1.4e-33
1	bio09	2	0.037	150.167	2	2.46e-33
1	bio08	1	0.028	113.606	1	1.59e-26
1	bio15	2	0.029	117.196	2	3.56e-26
1	bio18	1	0.023	92.229	1	7.72e-22
1	bio01	2	0.021	85.203	2	3.15e-19
1	bio17	1	0.018	72.795	1	1.44e-17
1	bio14	1	0.017	70.688	1	4.18e-17

round	variables	m	Dsq	Chisq	df	P
1	bio12	1	0.015	60.433	1	7.61e-15
1	bio19	1	0.012	49.634	1	1.85e-12
1	bio11	1	0.012	49.608	1	1.88e-12
1	bio13	1	0.01	41.025	1	1.5e-10
1	bio16	1	0.01	39.278	1	3.68e-10
2	bio07 + bio15	3	0.101	134.364	2	6.65e-30
2	bio07 + bio18	2	0.08	50.216	1	1.38e-12
2	bio07 + bio02	2	0.077	37.373	1	9.76e-10
2	bio07 + bio14	2	0.077	34.871	1	3.52e-09
2	bio07 + bio05	2	0.076	31.085	1	2.47e-08
2	bio07 + bio17	2	0.075	30.312	1	3.68e-08
2	bio07 + bio10	3	0.076	31.89	2	1.19e-07
2	bio07 + bio09	3	0.074	24.617	2	4.51e-06
2	bio07 + bio01	3	0.074	22.633	2	1.22e-05
2	bio07 + bio08	2	0.07	8.722	1	0.00314
2	bio07 + bio04	2	0.069	4.232	1	0.0397
2	bio07 + bio19	2	0.069	3.378	1	0.0661
2	bio07 + bio12	2	0.069	2.582	1	0.108
2	bio07 + bio06	2	0.068	1.181	1	0.277
2	bio07 + bio13	2	0.068	0.768	1	0.381
2	bio07 + bio16	2	0.068	0.365	1	0.546
2	bio07 + bio11	2	0.068	0.024	1	0.876
3	bio07 + bio15 + bio08	4	0.114	51.92	1	5.78e-13
3	bio07 + bio15 + bio01	5	0.106	21.865	2	1.79e-05
3	bio07 + bio15 + bio10	5	0.105	17.146	2	0.000189
3	bio07 + bio15 + bio09	5	0.105	15.579	2	0.000414
3	bio07 + bio15 + bio18	4	0.102	6.024	1	0.0141
3	bio07 + bio15 + bio17	4	0.101	1.071	1	0.301
3	bio07 + bio15 + bio02	4	0.101	0.751	1	0.386
3	bio07 + bio15 + bio05	4	0.101	0.541	1	0.462
3	bio07 + bio15 + bio14	4	0.101	0.208	1	0.649

round	variables	m	Dsq	Chisq	df	P
4	*bio07 + bio15 + bio08 + bio10*	*6*	*0.118*	*18.349*	*2*	*0.000104*[3]
4	bio07 + bio15 + bio08 + bio09	6	0.117	13.009	2	0.0015
4	bio07 + bio15 + bio08 + bio01	6	0.114	3.222	2	0.2
5	bio07 + bio15 + bio08 + bio10 + bio09	8	0.118	0.303	2	0.859

Model parameters

A vector of model predictions in Probability Ratio Output, denoted $\dot{q}$, is calculated as follows:

$$\dot{q}_i = N \cdot e^{\alpha + \sum_{k=1}^{m} \beta_k x_{ik}}$$

where N is the number of background locations (presences plus uninformed background locations), α is a normalizing constant, β is a vector of coefficients, and x is a matrix of derived variables.

The names of the coefficients in the table below are made up of the name of the explanatory variable followed by an extension that denotes the type of transformation applied to obtain the derived variable. The extensions follow defaults in Vollering et al. 2018.

parameter	value
N	3983
α	-6.152092426
bio07_HR9	2.059009573
bio15_HF17	-22.60395871
bio15_T6	-1.533943035
bio08_D05	-3.650384035
bio10_HF12	-7.146835695
bio10_D2	-2.742808015

[3] selected model

Response curves

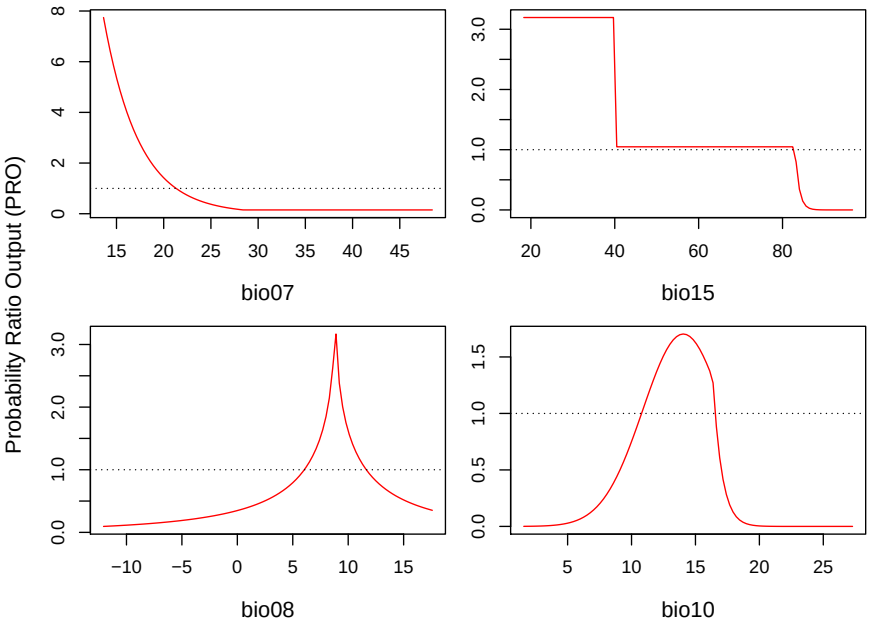

Figure A.10: Single-effect response curves of the model

D.4 Target group background selection, CHELSA

Training data map

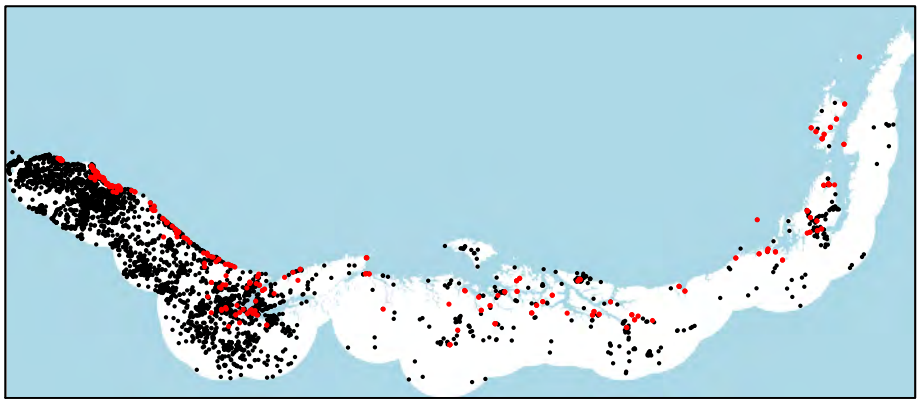

Figure A.11: Presence locations (red) and uninformed background locations (black) plotted across the training area. The map projection is as used during modeling: the Lambert Azimuthal Equal Area projection (ESRI:102017).

Stepwise forward selection of explanatory variables

round	variables	m	Dsq	Chisq	df	P
1	bio02	2	0.099	405.35	2	9.54e-89
1	bio07	1	0.069	282.462	1	2.18e-63
1	bio08	1	0.049	201.718	1	8.81e-46
1	bio04	1	0.039	158.052	1	3.02e-36
1	bio15	2	0.035	145.607	2	2.41e-32
1	bio06	1	0.033	137.057	1	1.17e-31
1	bio05	2	0.033	134.132	2	7.48e-30
1	bio03	2	0.026	104.948	2	1.62e-23
1	bio17	1	0.023	93.768	1	3.55e-22
1	bio18	1	0.022	90.96	1	1.47e-21
1	bio11	1	0.022	88.784	1	4.4e-21
1	bio12	1	0.02	80.589	1	2.78e-19
1	bio14	1	0.019	78.674	1	7.33e-19

round	variables	m	Dsq	Chisq	df	P
1	bio10	2	0.02	81.939	2	1.61e-18
1	bio01	2	0.016	65.364	2	6.4e-15
1	bio09	1	0.013	53.106	1	3.16e-13
1	bio13	1	0.012	48.959	1	2.61e-12
1	bio16	1	0.011	44.752	1	2.24e-11
1	bio19	1	0.008	33.006	1	9.19e-09
2	bio02 + bio15	4	0.111	50.812	2	9.26e-12
2	bio02 + bio03	4	0.106	29.166	2	4.64e-07
2	bio02 + bio01	4	0.104	22.059	2	1.62e-05
2	bio02 + bio13	3	0.103	17.116	1	3.52e-05
2	bio02 + bio12	3	0.102	14.388	1	0.000149
2	bio02 + bio16	3	0.102	13.347	1	0.000259
2	bio02 + bio11	3	0.102	12.918	1	0.000325
2	bio02 + bio19	3	0.102	12.686	1	0.000368
2	bio02 + bio07	3	0.102	12.156	1	0.000489
2	bio02 + bio08	3	0.101	9.422	1	0.00214
2	bio02 + bio06	3	0.101	9.332	1	0.00225
2	bio02 + bio04	3	0.1	5.966	1	0.0146
2	bio02 + bio09	3	0.1	3.419	1	0.0645
2	bio02 + bio10	4	0.1	4.107	2	0.128
2	bio02 + bio17	3	0.099	2.2	1	0.138
2	bio02 + bio05	4	0.099	0.905	2	0.636
2	bio02 + bio18	3	0.099	0.015	1	0.903
2	bio02 + bio14	3	0.099	0	1	0.986
3	bio02 + bio15 + bio06	5	0.129	71.594	1	2.64e-17
3	bio02 + bio15 + bio01	6	0.128	69.771	2	7.07e-16
3	bio02 + bio15 + bio11	5	0.127	64.705	1	8.7e-16
3	bio02 + bio15 + bio07	5	0.119	31.143	1	2.4e-08
3	bio02 + bio15 + bio08	5	0.118	26.95	1	2.09e-07
3	bio02 + bio15 + bio13	5	0.117	24.213	1	8.62e-07
3	bio02 + bio15 + bio19	5	0.116	18.836	1	1.42e-05

round	variables	m	Dsq	Chisq	df	P
3	bio02 + bio15 + bio16	5	0.114	13.189	1	0.000282
3	bio02 + bio15 + bio12	5	0.113	7.277	1	0.00698
3	bio02 + bio15 + bio03	6	0.113	8.242	2	0.0162
4	*bio02 + bio15 + bio06 + bio13*	*6*	*0.132*	*12.629*	*1*	*0.00038*[4]
4	bio02 + bio15 + bio06 + bio16	6	0.131	10.668	1	0.00109
4	bio02 + bio15 + bio06 + bio12	6	0.131	9.56	1	0.00199
4	bio02 + bio15 + bio06 + bio01	7	0.131	8.747	2	0.0126
4	bio02 + bio15 + bio06 + bio19	6	0.13	5.556	1	0.0184
4	bio02 + bio15 + bio06 + bio11	6	0.13	3.821	1	0.0506
4	bio02 + bio15 + bio06 + bio07	6	0.129	2.967	1	0.085
4	bio02 + bio15 + bio06 + bio08	6	0.129	0.618	1	0.432
5	bio02 + bio15 + bio06 + bio13 + bio16	7	0.132	0.873	1	0.35
5	bio02 + bio15 + bio06 + bio13 + bio12	7	0.132	0.362	1	0.547

Model parameters

A vector of model predictions in Probability Ratio Output, denoted $\dot{q}$, is calculated as follows:

$$\dot{q}_i = N \cdot e^{\alpha + \sum_{k=1}^{m} \beta_k x_{ik}}$$

where N is the number of background locations (presences plus uninformed background locations), α is a normalizing constant, β is a vector of coefficients, and x is a matrix of derived variables.

The names of the coefficients in the table below are made up of the name of the explanatory variable followed by an extension that denotes the type of transformation applied to obtain the derived variable. The extensions follow defaults in Vollering et al. 2018.

[4] selected model

parameter	value
N	4155
α	-8.955329309
bio02_D2	-2.582887546
bio02_T14	-2.465615119
bio15_HF16	-20.59776809
bio15_HR10	2.903867844
bio06_HF15	2.511405467
bio13_T6	0.66094999

Response curves

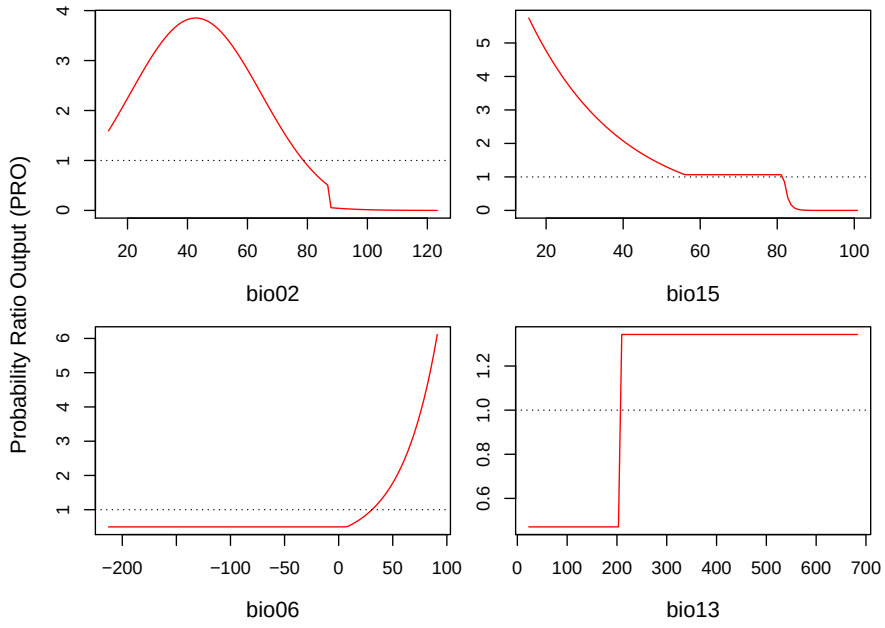

Figure A.12: Single-effect response curves of the model

D.5 Presence thinning, WorldClim

Training data map

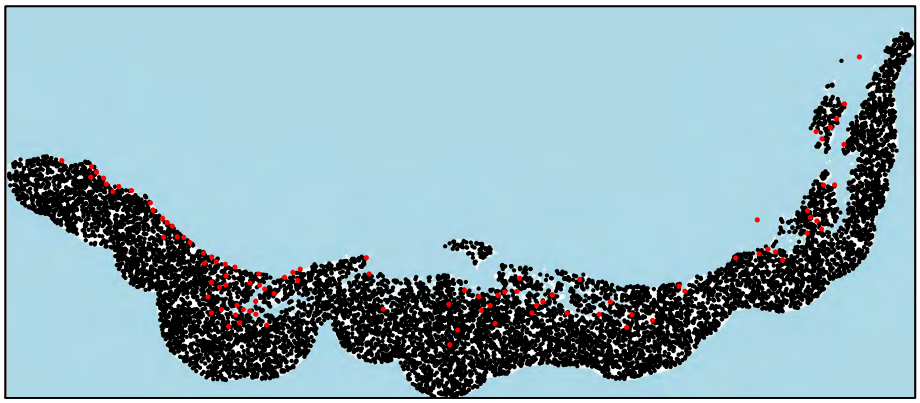

Figure A.13: Presence locations (red) and uninformed background locations (black) plotted across the training area. The map projection is as used during modeling: the Lambert Azimuthal Equal Area projection (ESRI:102017).

Stepwise forward selection of explanatory variables

round	variables	m	Dsq	Chisq	df	P
1	bio07	2	0.082	163.626	2	2.94e-36
1	bio04	1	0.075	148.151	1	4.4e-34
1	bio06	1	0.069	137.057	1	1.17e-31
1	bio09	2	0.053	104.696	2	1.84e-23
1	bio11	1	0.047	94.023	1	3.12e-22
1	bio01	2	0.049	96.85	2	9.32e-22
1	bio10	1	0.045	90.263	1	2.09e-21
1	bio12	1	0.04	79.591	1	4.61e-19
1	bio16	1	0.039	77.205	1	1.54e-18
1	bio13	1	0.039	76.862	1	1.83e-18
1	bio19	2	0.041	80.962	2	2.63e-18
1	bio08	1	0.032	64.276	1	1.08e-15
1	bio05	2	0.03	59.275	2	1.34e-13

round	variables	m	Dsq	Chisq	df	P
1	bio02	1	0.027	52.877	1	3.55e-13
1	bio03	1	0.016	32.446	1	1.23e-08
1	bio17	1	0.015	30.199	1	3.9e-08
1	bio14	1	0.011	22.735	1	1.86e-06
1	bio15	2	0.01	20.21	2	4.09e-05
1	bio18	1	0.004	7.048	1	0.00793
2	bio07 + bio10	3	0.097	28.507	1	9.34e-08
2	bio07 + bio05	4	0.097	29.202	2	4.56e-07
2	bio07 + bio01	4	0.095	24.693	2	4.34e-06
2	bio07 + bio06	3	0.09	14.681	1	0.000127
2	bio07 + bio09	4	0.091	17.411	2	0.000166
2	bio07 + bio19	4	0.089	13.927	2	0.000946
2	bio07 + bio11	3	0.087	10.21	1	0.0014
2	bio07 + bio03	3	0.087	9.719	1	0.00182
2	bio07 + bio15	4	0.088	11.333	2	0.00346
2	bio07 + bio04	3	0.086	6.653	1	0.0099
2	bio07 + bio02	3	0.086	6.509	1	0.0107
2	bio07 + bio08	3	0.085	6.054	1	0.0139
2	bio07 + bio18	3	0.084	2.652	1	0.103
2	bio07 + bio12	3	0.083	2.113	1	0.146
2	bio07 + bio16	3	0.083	0.999	1	0.318
2	bio07 + bio13	3	0.083	0.807	1	0.369
2	bio07 + bio17	3	0.083	0.543	1	0.461
2	bio07 + bio14	3	0.083	0.529	1	0.467
3	*bio07 + bio10 + bio19*	*5*	*0.107*	*20.042*	*2*	*4.45e-05*[5]
3	bio07 + bio10 + bio15	5	0.101	7.746	2	0.0208
3	bio07 + bio10 + bio11	4	0.098	2.699	1	0.1
3	bio07 + bio10 + bio05	5	0.098	3.524	2	0.172
3	bio07 + bio10 + bio04	4	0.098	1.795	1	0.18
3	bio07 + bio10 + bio09	5	0.098	2.302	2	0.316

[5] selected model

round	variables	m	Dsq	Chisq	df	P
3	bio07 + bio10 + bio03	4	0.097	0.785	1	0.376
3	bio07 + bio10 + bio01	5	0.097	0.747	2	0.688
3	bio07 + bio10 + bio06	4	0.097	0.009	1	0.924

Model parameters

A vector of model predictions in Probability Ratio Output, denoted $\dot{q}$, is calculated as follows:

$$\dot{q}_i = N \cdot e^{\alpha + \sum_{k=1}^{m} \beta_k x_{ik}}$$

where N is the number of background locations (presences plus uninformed background locations), α is a normalizing constant, β is a vector of coefficients, and x is a matrix of derived variables.

The names of the coefficients in the table below are made up of the name of the explanatory variable followed by an extension that denotes the type of transformation applied to obtain the derived variable. The extensions follow defaults in Vollering et al. 2018.

parameter	value
N	10097
α	-5.329641179
bio07_D05	-5.689978932
bio07_T7	-1.347230325
bio10_D1	-5.443519377
bio19_D2	2.587513628
bio19_HR3	-4.702145218

Response curves

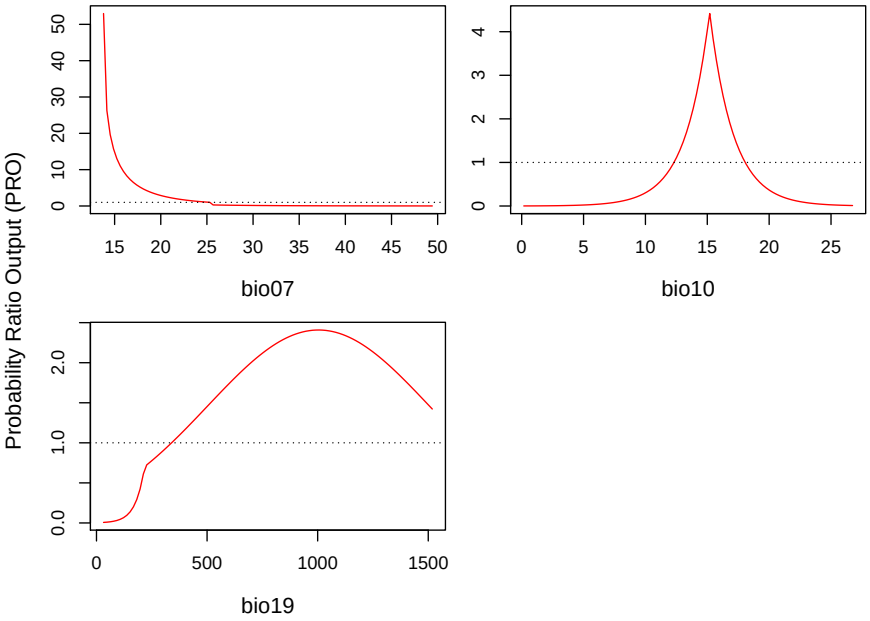

Figure A.14: Single-effect response curves of the model

D.6 Presence thinning, CHELSA

Training data map

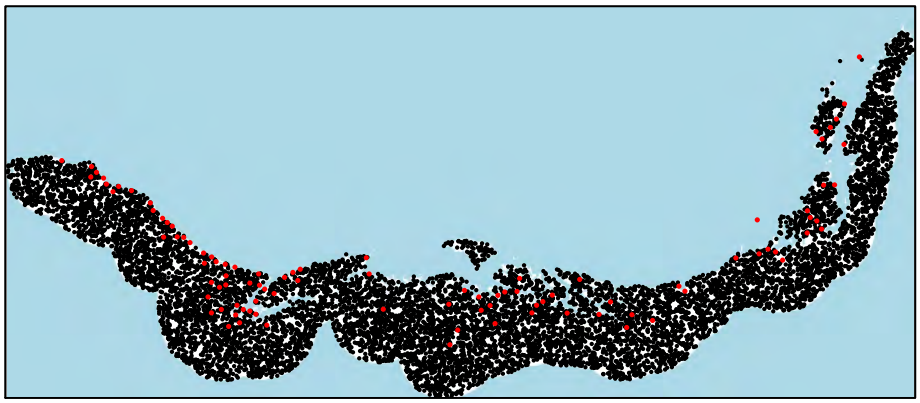

Figure A.15: Presence locations (red) and uninformed background locations (black) plotted across the training area. The map projection is as used during modeling: the Lambert Azimuthal Equal Area projection (ESRI:102017).

Stepwise forward selection of explanatory variables

round	variables	m	Dsq	Chisq	df	P
1	bio07	1	0.064	127.891	1	1.19e-29
1	bio06	1	0.06	119.115	1	9.88e-28
1	bio04	1	0.056	111.786	1	3.98e-26
1	bio11	1	0.048	95.248	1	1.68e-22
1	bio09	2	0.05	99.157	2	2.94e-22
1	bio08	1	0.045	90.344	1	2e-21
1	bio01	1	0.043	84.536	1	3.77e-20
1	bio10	2	0.038	75.892	2	3.31e-17
1	bio12	1	0.034	67.831	1	1.78e-16
1	bio16	1	0.033	65.241	1	6.63e-16
1	bio13	1	0.033	65.04	1	7.34e-16
1	bio19	1	0.03	59.75	1	1.08e-14
1	bio02	2	0.027	53.863	2	2.01e-12

round	variables	m	Dsq	Chisq	df	P
1	bio05	2	0.022	44.111	2	2.64e-10
1	bio03	2	0.021	41.321	2	1.06e-09
1	bio17	1	0.012	23.849	1	1.04e-06
1	bio14	1	0.01	18.914	1	1.37e-05
1	bio18	1	0.005	9.795	1	0.00175
1	bio15	1	0.004	7.004	1	0.00813
2	bio07 + bio10	3	0.087	45.205	2	1.53e-10
2	bio07 + bio01	2	0.081	32.621	1	1.12e-08
2	bio07 + bio05	3	0.081	34.121	2	3.9e-08
2	bio07 + bio06	2	0.078	27.462	1	1.6e-07
2	bio07 + bio08	2	0.077	24.694	1	6.72e-07
2	bio07 + bio09	3	0.079	28.378	2	6.88e-07
2	bio07 + bio11	2	0.075	21.735	1	3.13e-06
2	bio07 + bio03	3	0.074	19.628	2	5.47e-05
2	bio07 + bio02	3	0.071	13.931	2	0.000944
2	bio07 + bio15	2	0.067	6.12	1	0.0134
2	bio07 + bio04	2	0.066	2.914	1	0.0878
2	bio07 + bio13	2	0.066	2.5	1	0.114
2	bio07 + bio19	2	0.065	2.184	1	0.139
2	bio07 + bio16	2	0.065	2.012	1	0.156
2	bio07 + bio12	2	0.065	1.824	1	0.177
2	bio07 + bio18	2	0.065	1.749	1	0.186
2	bio07 + bio14	2	0.065	1.541	1	0.215
2	bio07 + bio17	2	0.064	0.098	1	0.754
3	*bio07 + bio10 + bio02*	*5*	*0.107*	*40.424*	*2*	*1.67e-09*[6]
3	bio07 + bio10 + bio11	4	0.099	22.856	1	1.75e-06
3	bio07 + bio10 + bio05	5	0.095	15.18	2	0.000505
3	bio07 + bio10 + bio08	4	0.092	9.753	1	0.00179
3	bio07 + bio10 + bio03	5	0.091	8.063	2	0.0177
3	bio07 + bio10 + bio01	4	0.089	4.589	1	0.0322

[6] selected model

round	variables	m	Dsq	Chisq	df	P
3	bio07 + bio10 + bio06	4	0.088	1.234	1	0.267
3	bio07 + bio10 + bio09	5	0.088	2.188	2	0.335
4	bio07 + bio10 + bio02 + bio11	6	0.109	3.322	1	0.0683
4	bio07 + bio10 + bio02 + bio08	6	0.108	1.83	1	0.176
4	bio07 + bio10 + bio02 + bio05	7	0.108	1.753	2	0.416

Model parameters

A vector of model predictions in Probability Ratio Output, denoted $\dot{q}$, is calculated as follows:

$$\dot{q}_i = N \cdot e^{\alpha + \sum_{k=1}^{m} \beta_k x_{ik}}$$

where N is the number of background locations (presences plus uninformed background locations), α is a normalizing constant, β is a vector of coefficients, and x is a matrix of derived variables.

The names of the coefficients in the table below are made up of the name of the explanatory variable followed by an extension that denotes the type of transformation applied to obtain the derived variable. The extensions follow defaults in Vollering et al. 2018.

parameter	value
N	10097
α	-10.14357188
bio07_D2	-2.143516457
bio10_D1	-10.04742459
bio10_M	8.443393505
bio02_D05	-4.498393715
bio02_HF14	-24.0349391

Response curves

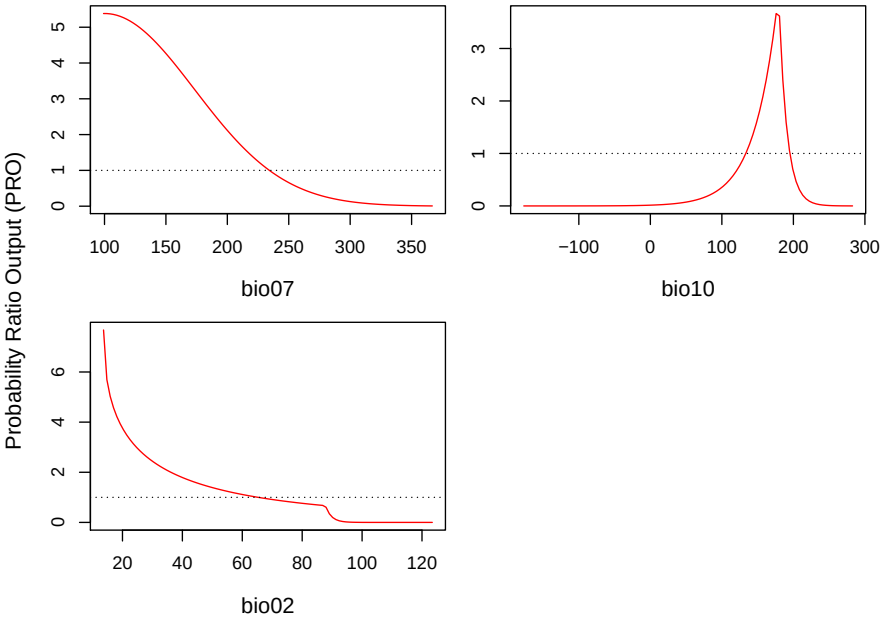

Figure A.16: Single-effect response curves of the model

D.7 Background thickening, WorldClim

Training data map

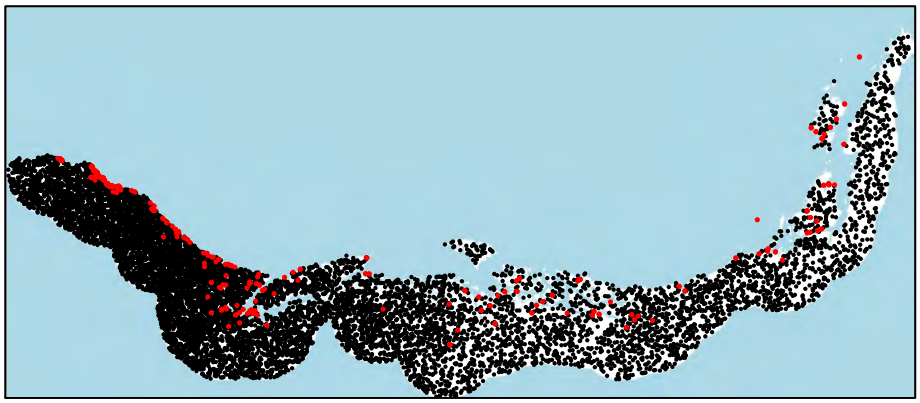

Figure A.17: Presence locations (red) and uninformed background locations (black) plotted across the training area. The map projection is as used during modeling: the Lambert Azimuthal Equal Area projection (ESRI:102017).

Stepwise forward selection of explanatory variables

round	variables	m	Dsq	Chisq	df	P
1	bio07	1	0.133	604.328	1	1.92e-133
1	bio04	1	0.128	580.693	1	2.65e-128
1	bio06	2	0.124	561.156	2	1.4e-122
1	bio03	1	0.067	302.789	1	8.13e-68
1	bio08	2	0.065	295.999	2	5.3e-65
1	bio11	2	0.063	287.433	2	3.84e-63
1	bio10	2	0.059	266.49	2	1.36e-58
1	bio09	3	0.054	247.284	3	2.53e-53
1	bio05	2	0.049	220.766	2	1.15e-48
1	bio01	2	0.046	208.766	2	4.65e-46
1	bio02	2	0.035	158.519	2	3.79e-35
1	bio16	1	0.028	127.841	1	1.22e-29
1	bio19	1	0.027	122.521	1	1.78e-28

round	variables	m	Dsq	Chisq	df	P
1	bio13	1	0.027	122.11	1	2.18e-28
1	bio12	1	0.027	120.53	1	4.84e-28
1	bio15	2	0.024	106.873	2	6.21e-24
1	bio17	1	0.003	13.429	1	0.000248
1	bio14	1	0.002	11.255	1	0.000794
2	bio07 + bio08	3	0.147	62.4	2	2.82e-14
2	bio07 + bio06	3	0.142	42.117	2	7.15e-10
2	bio07 + bio03	2	0.141	37.324	1	1e-09
2	bio07 + bio11	3	0.14	30.007	2	3.05e-07
2	bio07 + bio01	3	0.139	28.237	2	7.39e-07
2	bio07 + bio04	2	0.138	23.762	1	1.09e-06
2	bio07 + bio15	3	0.138	21.135	2	2.57e-05
2	bio07 + bio05	3	0.137	18.313	2	0.000106
2	bio07 + bio10	3	0.137	16.154	2	0.000311
2	bio07 + bio19	2	0.135	8.452	1	0.00365
2	bio07 + bio13	2	0.134	6.319	1	0.0119
2	bio07 + bio12	2	0.134	6.163	1	0.013
2	bio07 + bio16	2	0.134	5.098	1	0.0239
2	bio07 + bio02	3	0.135	6.923	2	0.0314
2	bio07 + bio17	2	0.133	1.876	1	0.171
2	bio07 + bio14	2	0.133	1.294	1	0.255
2	bio07 + bio09	4	0.134	2.503	3	0.475
3	bio07 + bio08 + bio15	5	0.155	36.277	2	1.33e-08
3	bio07 + bio08 + bio10	5	0.152	22.093	2	1.59e-05
3	bio07 + bio08 + bio05	5	0.151	19.206	2	6.75e-05
3	bio07 + bio08 + bio19	4	0.148	6.228	1	0.0126
3	bio07 + bio08 + bio01	5	0.148	6.678	2	0.0355
3	bio07 + bio08 + bio06	5	0.148	6.542	2	0.038
3	bio07 + bio08 + bio03	4	0.148	3.314	1	0.0687
3	bio07 + bio08 + bio11	5	0.147	1.509	2	0.47
3	bio07 + bio08 + bio04	4	0.147	0.045	1	0.832

round	variables	m	Dsq	Chisq	df	P
4	*bio07 + bio08 + bio15 + bio10*	*7*	*0.159*	*17.354*	*2*	*0.00017*[7]
4	bio07 + bio08 + bio15 + bio05	7	0.158	13.952	2	0.000934
5	bio07 + bio08 + bio15 + bio10 + bio05	9	0.159	0.546	2	0.761

Model parameters

A vector of model predictions in Probability Ratio Output, denoted $\dot{q}$, is calculated as follows:

$$\dot{q}_i = N \cdot e^{\alpha + \sum_{k=1}^{m} \beta_k x_{ik}}$$

where N is the number of background locations (presences plus uninformed background locations), α is a normalizing constant, β is a vector of coefficients, and x is a matrix of derived variables.

The names of the coefficients in the table below are made up of the name of the explanatory variable followed by an extension that denotes the type of transformation applied to obtain the derived variable. The extensions follow defaults in Vollering et al. 2018.

parameter	value
N	10243
α	-5.020292945
bio07_M	-7.532369673
bio08_D05	-4.861232185
bio08_HF15	-6.705842052
bio15_D05	1.323939519
bio15_HF17	-6996.496821
bio10_D05	-1.980170475
bio10_HF12	-2.150769291

[7] selected model

Response curves

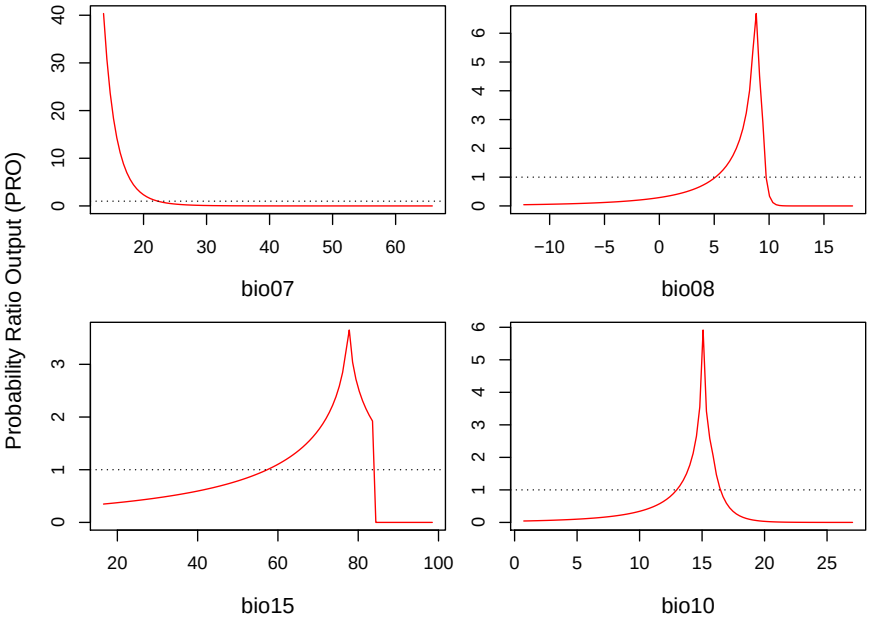

Figure A.18: Single-effect response curves of the model

D.8 Background thickening, CHELSA

Training data map

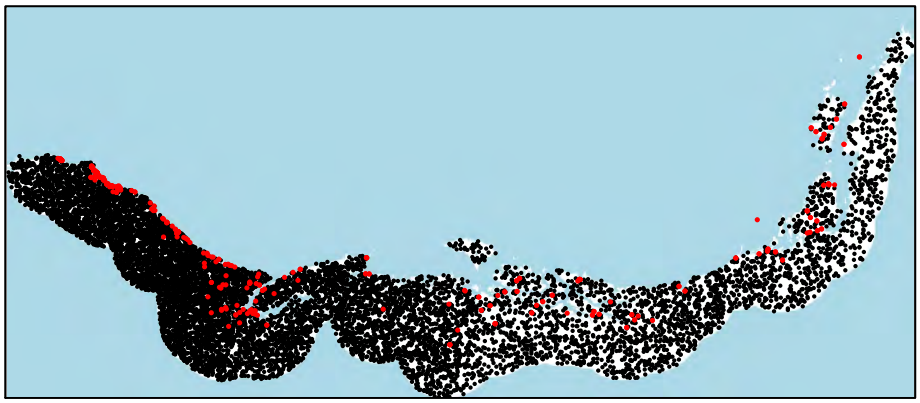

Figure A.19: Presence locations (red) and uninformed background locations (black) plotted across the training area. The map projection is as used during modeling: the Lambert Azimuthal Equal Area projection (ESRI:102017).

Stepwise forward selection of explanatory variables

round	variables	m	Dsq	Chisq	df	P
1	bio06	2	0.116	525.961	2	6.15e-115
1	bio07	2	0.108	492.624	2	1.07e-107
1	bio04	1	0.097	440.57	1	8.14e-98
1	bio08	3	0.091	412.808	3	3.72e-89
1	bio11	2	0.088	398.603	2	2.78e-87
1	bio02	2	0.06	271.927	2	8.95e-60
1	bio01	3	0.051	232.354	3	4.28e-50
1	bio10	3	0.044	199.383	3	5.73e-43
1	bio09	2	0.041	184.401	2	9.08e-41
1	bio05	2	0.034	154.445	2	2.9e-34
1	bio16	2	0.034	153.769	2	4.07e-34
1	bio12	2	0.034	153.099	2	5.69e-34
1	bio13	2	0.033	151.893	2	1.04e-33

round	variables	m	Dsq	Chisq	df	P
1	bio19	1	0.03	134.341	1	4.6e-31
1	bio18	2	0.005	23.212	2	9.11e-06
1	bio17	1	0.003	12.761	1	0.000354
1	bio14	1	0.002	9.025	1	0.00266
2	bio06 + bio02	4	0.155	178.721	2	1.55e-39
2	bio06 + bio11	4	0.145	132.41	2	1.77e-29
2	bio06 + bio05	4	0.143	121.739	2	3.67e-27
2	bio06 + bio18	4	0.137	96.183	2	1.3e-21
2	bio06 + bio07	4	0.137	94.955	2	2.4e-21
2	bio06 + bio01	5	0.134	84.256	3	3.75e-18
2	bio06 + bio10	5	0.134	84.144	3	3.96e-18
2	bio06 + bio09	4	0.13	65.342	2	6.47e-15
2	bio06 + bio17	3	0.127	53.023	1	3.3e-13
2	bio06 + bio14	3	0.126	47.829	1	4.65e-12
2	bio06 + bio08	5	0.127	48.909	3	1.36e-10
2	bio06 + bio12	4	0.124	38.953	2	3.48e-09
2	bio06 + bio16	4	0.12	21.306	2	2.36e-05
2	bio06 + bio13	4	0.12	19.061	2	7.26e-05
2	bio06 + bio19	3	0.118	9.914	1	0.00164
2	bio06 + bio04	3	0.118	9.248	1	0.00236
3	bio06 + bio02 + bio04	5	0.17	66.309	1	3.85e-16
3	bio06 + bio02 + bio10	7	0.164	42.08	3	3.86e-09
3	bio06 + bio02 + bio08	7	0.162	29.527	3	1.74e-06
3	bio06 + bio02 + bio13	6	0.16	22.81	2	1.11e-05
3	bio06 + bio02 + bio01	7	0.16	23.777	3	2.78e-05
3	bio06 + bio02 + bio16	6	0.159	18.847	2	8.08e-05
3	bio06 + bio02 + bio12	6	0.159	18.264	2	0.000108
3	bio06 + bio02 + bio07	6	0.159	15.366	2	0.000461
3	bio06 + bio02 + bio18	6	0.158	13.693	2	0.00106
3	bio06 + bio02 + bio09	6	0.158	13.048	2	0.00147
3	bio06 + bio02 + bio17	5	0.157	8.477	1	0.0036

round	variables	m	Dsq	Chisq	df	P
3	bio06 + bio02 + bio11	6	0.157	10.618	2	0.00495
3	bio06 + bio02 + bio14	5	0.156	5.256	1	0.0219
3	bio06 + bio02 + bio19	5	0.155	1.217	1	0.27
3	bio06 + bio02 + bio05	6	0.156	1.674	2	0.433
4	bio06 + bio02 + bio04 + bio08	8	0.174	19.016	3	0.000271
4	bio06 + bio02 + bio04 + bio12	7	0.173	15.311	2	0.000473
4	bio06 + bio02 + bio04 + bio13	7	0.172	11.454	2	0.00326
4	bio06 + bio02 + bio04 + bio17	6	0.171	7.541	1	0.00603
4	bio06 + bio02 + bio04 + bio16	7	0.172	9.742	2	0.00767
4	bio06 + bio02 + bio04 + bio09	7	0.172	8.84	2	0.012
4	bio06 + bio02 + bio04 + bio01	8	0.172	10.654	3	0.0138
4	bio06 + bio02 + bio04 + bio18	7	0.171	5.923	2	0.0518
4	bio06 + bio02 + bio04 + bio10	8	0.171	7.347	3	0.0616
4	bio06 + bio02 + bio04 + bio07	7	0.171	4.475	2	0.107
4	bio06 + bio02 + bio04 + bio11	7	0.17	3.375	2	0.185
5	*bio06 + bio02 + bio04 + bio08 + bio12*	*10*	*0.177*	*15.487*	*2*	*0.000434*[8]
5	bio06 + bio02 + bio04 + bio08 + bio13	10	0.176	10.518	2	0.0052
5	bio06 + bio02 + bio04 + bio08 + bio16	10	0.176	10.122	2	0.00634
5	bio06 + bio02 + bio04 + bio08 + bio17	9	0.175	4.951	1	0.0261

[8] selected model

round	variables	m	Dsq	Chisq	df	P
6	bio06 + bio02 + bio04 + bio08 + bio12 + bio13	12	0.178	4.179	2	0.124
6	bio06 + bio02 + bio04 + bio08 + bio12 + bio16	12	0.178	2.058	2	0.357

Model parameters

A vector of model predictions in Probability Ratio Output, denoted $\dot{q}$, is calculated as follows:

$$\dot{q_i} = N \cdot e^{\alpha + \sum_{k=1}^{m} \beta_k x_{ik}}$$

where N is the number of background locations (presences plus uninformed background locations), α is a normalizing constant, β is a vector of coefficients, and x is a matrix of derived variables.

The names of the coefficients in the table below are made up of the name of the explanatory variable followed by an extension that denotes the type of transformation applied to obtain the derived variable. The extensions follow defaults in Vollering et al. 2018.

parameter	value
N	10243
α	-5.180681688
bio06_HF17	5.035540643
bio06_D2	-9.506398283
bio02_HF10	-6.766011518
bio02_T3	-1.727938294
bio04_HR8	-4.657031056
bio08_D05	-1.653936235
bio08_HF18	-939.904812
bio08_T14	0.607510014
bio12_HR4	0.243793567
bio12_D2	-6.494875284

Response curves

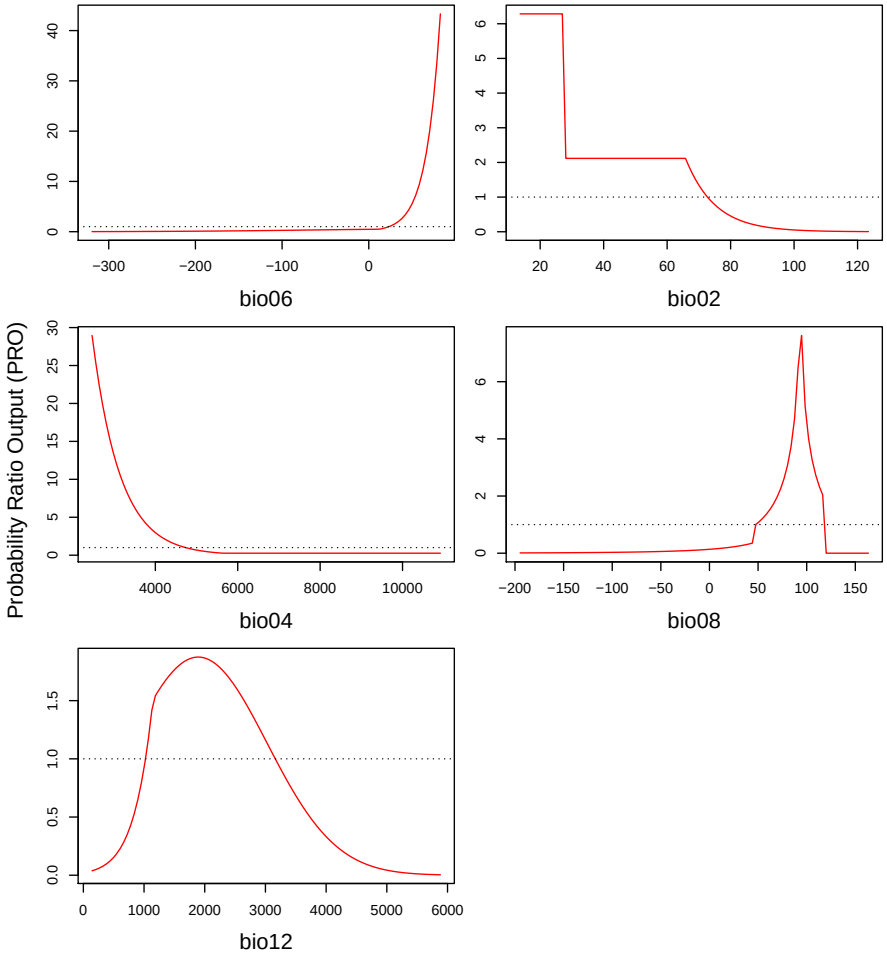

Figure A.20: Single-effect response curves of the model

E Sitka spruce case study: predictions using WorldClim

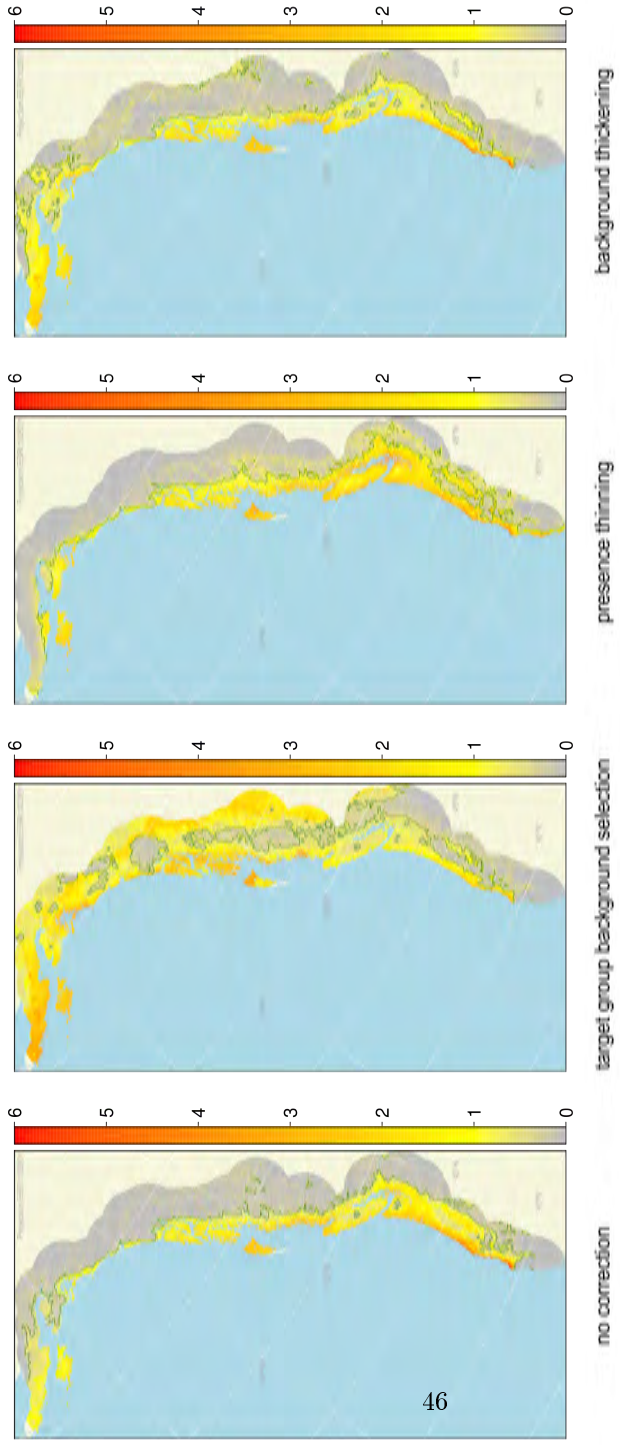

Figure A.21: Relative presence probability of Sitka spruce, predicted by models using WorldClim climate data and four different sampling bias approaches. Predictions are given in probability ratio output (PRO) units, which means the value of one is the expected relative presence probability in a randomly drawn training data location — i.e. an "average" training data location has PRO=1 (Halvorsen 2013). To visualize differences between large or small values, the color scale represents log2(PRO+1). The dark green lines show the 5 % omission threshold in model predictions — i.e. the value above which 95 % of presences occur.

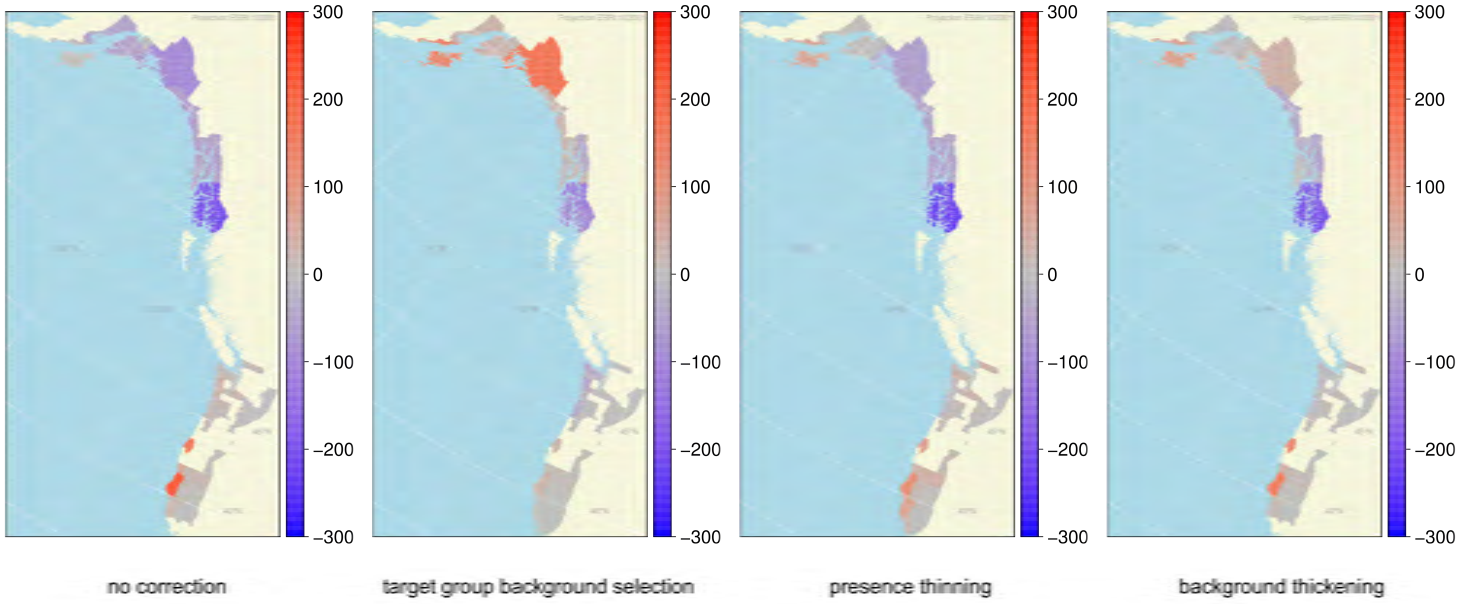

Figure A.22: Overprediction and underprediction of relative presence probability of Sitka spruce, compared to surveyed presence in the Forest Inventory and Analysis (FIA) program. Predictions are from models using WorldClim climate data and four different sampling bias approaches. Plotted values represent the difference between the actual number of surveyed Sitka spruce presences in each county (sum = 1831), and the fraction of 1831 presences expected to occur in each county based on model predictions (red shows overprediction while blue shows underprediction). The number of surveyed presences is considered only in U.S. counties with equal plot density in the FIA database.

F Sitka spruce case study: model projection

In the case study, the performance of each sampling bias correction approach was also evaluated in the context of model projection — when the predictions are made outside the spatial or temporal range of the training data (Halvorsen 2012). Model projection requires that environmental conditions in the spatial or temporal range of interest are not outside the range of conditions used to train the model (Elith et al. 2010; Merow et al. 2013). We were interested in projecting models of Sitka spruce to Norway, because it poses a high ecological risk there (Gederaas et al. 2012; Elven et al. 2018), so we assessed how well Norwegian conditions were represented in the training data produced by each bias correction approach. The comparison was performed multivariately using Mahalanobis distance, which takes into account the correlation structure between variables (Mesgaran et al. 2014). For each model in the case study, we calculated the proportion of Norwegian locations beyond the 100th, 99th, and 95th percentile of conditions in the training data set. Locations beyond the 100th percentile of training data conditions require model extrapolation. The comparison was performed once with all 19 BIOCLIM variables, and once with only the variables selected in the relevant model.

Mahalanobis distances showed that there were no strictly novel climatic conditions in the projection area, compared to any set of training data (Table A.17). However, target group background selection caused large proportions of Norwegian locations to occur beyond the 95th or even 99th percentile of multivariate training data conditions, while the other correction methods caused much smaller proportions of Norwegian locations to be extreme compared to training data.

Model		All variables			Selected variables		
Correction method	Climate data product	100^{th}	99^{th}	95^{th}	100^{th}	99^{th}	95^{th}
no correction	WorldClim	0	2.3	9.9	0	0	0.2
no correction	CHELSA	0	9.6	27.7	0	2.2	10.6
target group background selection	WorldClim	0	11.2	60.1	0	1.4	58.4
target group background selection	CHELSA	0	36.7	73.6	0	4.7	31.8
presence thinning	WorldClim	0	2.2	10.8	0	0	0.5
presence thinning	CHELSA	0	9.3	24.7	0	0	0.4
background thickening	WorldClim	0	1.9	11.8	0	0	1.3
background thickening	CHELSA	0	7.8	33.3	0	0.4	2.7

Table A.17: The percentage of Norwegian locations beyond the 100th, 99th, and 95th percentile of environmental conditions in the eight training data sets used to model Sitka spruce's native distribution. For example, no Norwegian locations had conditions that extended beyond the most extreme training data conditions. Values calculated using all 19 BIOCLIM variables are contingent only on the locations of the training data and the climate data product, and therefore directly comparable across rows. Values calculated using selected variables only are additionally contingent on the complexity of the model, and are relevant for justifying model projection.

G Sitka spruce case study: sources of uncertainty

We wanted to assess whether the choice between background thickening and presence thinning or the choice between WorldClim and CHELSA was the larger source of uncertainty (variability) in model predictions. To measure the magnitude of variability between models, we needed a measure of dispersion that is comparable across relative probabilities ranging from from zero to infinity, so we used the "proportional variability" index (Heath 2006; Heath and Borowski 2013). We partitioned uncertainty between correction approach and data product by calculating the mean proportional variability among pairs of models differing only in one respect.

There was greater variability in predictions between models using WorldClim and models using CHELSA than there was between models using background thickening and models using presence thinning; mean proportional variability was 0.634 and 0.608, respectively.

References

Austin, M. (2007). Species distribution models and ecological theory: A critical assessment and some possible new approaches. *Ecological Modelling*, 200(1-2):1–19.

Austin, M. P. (2002). Spatial prediction of species distribution: An interface between ecological theory and statistical modelling. *Ecological Modelling*, 157:101–118.

Bobrowski, M. and Schickhoff, U. (2017). Why input matters: Selection of climate data sets for modelling the potential distribution of a treeline species in the Himalayan region. *Ecological Modelling*, 359:92–102.

Budic, L., Didenko, G., and Dormann, C. F. (2016). Squares of different sizes: Effect of geographical projection on model parameter estimates in species distribution modeling. *Ecology and Evolution*, 6(1):202–211.

Daly, C. (2006). Guidelines for assessing the suitability of spatial climate data sets. *International Journal of Climatology*, 26(6):707–721.

Elith, J., Kearney, M., and Phillips, S. (2010). The art of modelling range-shifting species. *Methods in Ecology and Evolution*, 1(4):330–342.

Elven, R., Hegre, H., Solstad, H., Pedersen, O., Pedersen, P. A., Åsen, P. A., Bjureke, K., and Vandvik, V. (2018). Picea sitchensis, vurdering av økologisk risiko. https://artsdatabanken.no/Fab2018/N/537.

Fick, S. E. and Hijmans, R. J. (2017). WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, 37(12):4302–4315.

Fithian, W. and Hastie, T. (2013). Finite-sample equivalence in statistical models for presence-only data. *The Annals of Applied Statistics*, 7(4):1917–1939.

Gederaas, L., Moen, T. L., Skjelseth, S., and Larsen, L.-K. (2012). Alien species in Norway - with the Norwegian Black List 2012. Technical report, The Norwegian Biodiversity Information Centre, Trondheim, Norway.

Halvorsen, R. (2012). A gradient analytic perspective on distribution modelling. *Sommerfeltia*, 35:1–165.

Halvorsen, R. (2013). A strict maximum likelihood explanation of MaxEnt, and some implications for distribution modelling. *Sommerfeltia*, 36(1):1–132.

Halvorsen, R., Mazzoni, S., Bryn, A., and Bakkestuen, V. (2015). Opportunities for improved distribution modelling practice via a strict maximum likelihood interpretation of MaxEnt. *Ecography*, 38(2):172–183.

Halvorsen, R., Mazzoni, S., Dirksen, J. W., Næsset, E., Gobakken, T., and Ohlson, M. (2016). How important are choice of model selection method and spatial autocorrelation of presence data for distribution modelling by MaxEnt? *Ecological Modelling*, 328:108–118.

Heath, J. P. (2006). Quantifying temporal variability in population abundances. *Oikos*, 115(3):573–581.

Heath, J. P. and Borowski, P. (2013). Quantifying proportional variability. *PLoS ONE*, 8(12):e84074.

Jaynes, E. T. (1957a). Information theory and statistical mechanics. *Physical review*, 106(4):620.

Jaynes, E. T. (1957b). Information theory and statistical mechanics. II. *Physical review*, 108(2):171.

Karger, D. N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R. W., Zimmermann, N. E., Linder, H. P., and Kessler, M. (2017). Climatologies at high resolution for the earth's land surface areas. *Scientific Data*, 4:170122.

Merow, C., Smith, M. J., and Silander, J. A. (2013). A practical guide to MaxEnt for modeling species' distributions: What it does, and why inputs and settings matter. *Ecography*, 36(10):1058–1069.

Mesgaran, M. B., Cousens, R. D., and Webber, B. L. (2014). Here be dragons: A tool for quantifying novelty due to covariate range and correlation change when projecting species distribution models. *Diversity and Distributions*, 20(10):1147–1159.

Morales-Barbero, J. and Vega-Álvarez, J. (2018). Input matters matter: Bioclimatic consistency to map more reliable species distribution models. *Methods in Ecology and Evolution*, 10(2):212–224.

Phillips, S. J., Anderson, R. P., and Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, 190(3-4):231–259.

Reineking, B. and Schröder, B. (2006). Constrain to perform: Regularization of habitat models. *Ecological Modelling*, 193(3-4):675–690.

Vollering, J., Mazzoni, S., and Halvorsen, R. (2018). MIAmaxent: A modular, integrated approach to maximum entropy distribution modeling.

Whittaker, R. H. (1967). Gradient analysis of vegetation. *Biological Reviews*, 42(2):207–264.