

Ecography

E4596

Elith, J., Graham, C. H., Anderson, R. P., Dudík, M., Ferrier, S., Guisan, A., Hijmans, R. J., Huettmann, F., Leathwick, J. R., Lehmann, A., Li, J., Lohmann, L. G., Loiselle, B. A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J. McC., Peterson, A. T., Phillips, S. J., Richardson, K. S., Schachetti-Pereira, R., Schapire, R. E., Soberón, J., Williams, S., Wisz, M. S. and Zimmermann, N. E. 2006. Novel methods improve prediction of species' distributions from occurrence data. – Ecography 29: 129–151.

Table S1: Variables used in modelling.

	AWT	CAN	NSW	NZ	SA	SWI	reason for restrictions
BIOCLIM	all	all	all	all	all	all	NA
BRT	not annual temp, precip DQ, annual radiation, MI (moisture index) seas, MI of lowest quarter MI	not precip DQ, precip seas, temp seas, april temp	not min temp	not dem, rain	not annual temp, temp range, temp WQ	not av temp coldest month	high pairwise correlation (>0.85)
BRUTO	as for BRT	as for BRT minus veg	as for BRT minus veg	as for BRT minus age and toxicants	as for BRT	as for BRT minus calc	high pairwise correlation (>0.85); cannot use categories or 3 or less unique values
DOMAIN	all	all	all	all	all	all	NA
GAM, GLM	not temp WQ, temp CQ, precip DQ, mean rad, MI seas, MI of lowest quarter MI	not precip DQ, precip seas, april temp	not min temp	not dem, rain	not max temp, min temp, temp range, temp WQ	not annual temp	high pairwise correlation (>0.80 or 0.85)
DK-GARP	all	all	all	all	all	all	NA
OM-GARP	all	all	all	all	all	all	NA
GDM, GDM-SS	not max temp warmest quarter, precip dry quarter, annual radiation, moisture index seas	not precip DQ, temp seas, april temp, veg	not veg	not rain	not annual temp, temp WQ	not av temp coldest month	high pairwise correlation (>0.90); categories not modeled
LIVES	all	all	all	all	all	all	NA
MARS	as for BRT	as for BRT, minus veg	as for BRT, minus veg	as for BRT	as for BRT	as for BRT	high pairwise correlation (>0.85); categories not modeled
MAXENT	all	all	all	all	all	all	NA

Note: WQ is an abbreviation for wettest quarter – also coldest (CQ), driest (DQ). Seas = seasonality.

Table S2. Details of implementation.

Method	Available as code ± interface (and method used here, if either)	Computer specs for these runs	Time taken for all species predicting to sites	Estimated additional time taken to make maps for all species	Expertise of user wrt method	Data manipulations, and tests of settings	Anticipated gain with more experience or time (because..)
BRUTO	code	Pentium 4, 2.4 GHz CPU, 1 GB ram	17 min	not done, but possible. Time same as GAM	user (JE) new to method but experienced modeler ¹	highly correlated and categorical variables removed; also those with less than 3 unique values	some (tuning of parameters per region; using all variables (except correlated ones))
BRT	either (code)	Pentium 4, 2.4 GHz CPU, 1 GB ram	ca 80 h for this run; ca 8 h if optimised	not done, but possible. Time likely to be <1 d	user (JE) new to method but experienced modeler ¹	highly correlated variables removed	some (improved techniques and knowledge to select model complexity and learning rates and tuning per region)
BIOCLIM	either (code)	standard modern PC	5 h	20 h	experienced (CG and RH)	none	little
DOMAIN	either (code)	standard modern PC	7 h	20 h	experienced (CH and RH)	none	some (predictions could be scaled differently, to improve spread of predictions)
GAM	either (code)	Pentium 4, 2.4 GHz CPU, 1 GB ram	17 h	15 h	experienced (AL, AG and JE)	highly correlated variables removed	some (use modern selection methods eg Lasso; include interactions)
DK-GARP	interface	3 twin CPU PCs	ca 6 weeks on one processor	produced with site models	experienced	none	little
OM-GARP	either (code)	128-processor PC cluster	ca 200 h on one processor	months (time for one cell = time for one site)	experienced (author of new code, RP)	none	little

GDM (both versions)	interface (but many species at a time)	Pentium 4, 2.4 GHz CPU, 1 GB ram	21 h (GDM); 35 h GDM-SS. This will reduce substantially; code being optimised	40 h	new user (JE, experienced modeler), trained by authors of code (GM, SF)	highly correlated variables removed	little
GLM	either (code)	Pentium 4, 2.4 GHz CPU, 1 GB ram	17 h	15 h	experienced (AL, AG, JE)	highly correlated variables removed	as for GAMs
LIVES	code	standard modern PC	5 h	not done; possible but slow	experienced (JLi, author of code)	categorical variables converted to binary	little
MAXENT (both versions)	either (code)	Pentium 4 running Linux, 2.3.2 GHz CPUs, 8.75 GB ram	2.75 h for this run; 26 min on default settings; (using 1 processor)	4.25 h	experienced (SP, author of code)	tuned to modelling data, sample size used to inform regularization and use of features	some (tuning of regularization and feature selection)
MARS (all versions)	either (code)	Pentium 4, 2.4 GHz CPU, 1 GB ram	15 min (less for mars community)	12 h	user (JE) new to method but experienced modeler ¹	highly correlated variables and categorical variables removed	some (tuning of parameters per region, using categorical vars)

¹ In all cases, J. Elith and J. Leathwick worked together on code and exploring appropriate settings on other data. Also, we has some advice from authors of the code.

Table S3. Predictive performance for each species, summarized across methods (*cv = coefficient of variation).

Region	Species code	Species name	Max AUC	Max COR	Max KAPPA	Mean AUC	cv(%) * AUC
awt	BFMON	<i>Monarcha melanopsis</i>	0.760	0.349	0.399	0.677	7.3
awt	BHE	<i>Lichenostomus frenatus</i>	0.843	0.554	0.570	0.766	8.4
awt	BST	<i>Colluricincla boweri</i>	0.796	0.459	0.467	0.737	5.2
awt	CC	<i>Orthonyx spaldingii</i>	0.696	0.310	0.301	0.652	3.8
awt	FW	<i>Oreoscopus gutturalis</i>	0.709	0.262	0.293	0.655	4.9
awt	GHE	<i>Meliphaga gracilis</i>	0.876	0.614	0.603	0.766	9.2
awt	GHR	<i>Heteromyias albispecularis</i>	0.810	0.536	0.565	0.774	3.8
awt	GOLDBB	<i>Prionodura newtoniana</i>	0.874	0.300	0.325	0.794	7.5
awt	LBSW	<i>Sericornis magnirostris</i>	0.488	0.005	0.061	0.465	3.7
awt	LEWHE	<i>Meliphaga lewinii</i>	0.764	0.488	0.526	0.694	6.1
awt	MACHE	<i>Xanthotis macleayana</i>	0.577	0.121	0.148	0.518	9.0
awt	MTHORN	<i>Acanthiza katherina</i>	0.926	0.597	0.760	0.784	13.2
awt	PMON	<i>Arses kaupi</i>	0.582	0.036	0.138	0.450	14.3
awt	SFD	<i>Ptilinopus superbus</i>	0.688	0.312	0.327	0.531	13.6
awt	SMON	<i>Monarcha trivirgatus</i>	0.649	0.270	0.261	0.583	6.5
awt	TBBB	<i>Scenopoeetes dentirostris</i>	0.779	0.340	0.378	0.746	3.2
awt	VRIF	<i>Ptiloris victoriae</i>	0.595	0.172	0.178	0.543	4.4
awt	WOMP	<i>Ptilinopus magnificus</i>	0.594	0.135	0.159	0.532	5.9
awt	YSHE	<i>Meliphaga notata</i>	0.825	0.549	0.569	0.720	8.9
awt	YTSW	<i>Sericornis citreogularis</i>	0.723	0.310	0.294	0.673	4.7
awt	Argpol	<i>Argyrodendron polyandrum</i>	0.705	0.346	0.381	0.594	15.9
awt	Aspsim	<i>Asplenium simplicifrons</i>	0.704	0.340	0.392	0.569	15.9
awt	Ausbid	<i>Austromyrtus bidwillii</i>	0.738	0.364	0.408	0.628	10.0
awt	Ausele	<i>Austromatthaea elegans</i>	0.818	0.413	0.414	0.704	10.1
awt	Balaus	<i>Balanops australiana</i>	0.898	0.548	0.550	0.810	9.0
awt	Beiban	<i>Beilschmiedia bancroftii</i>	0.815	0.532	0.585	0.751	5.1
awt	Carsub	<i>Cardwellia sublimis</i>	0.823	0.566	0.665	0.703	10.0
awt	Clmaus	<i>Calamus australis</i>	0.750	0.407	0.470	0.663	11.2
awt	Covpoe	<i>Coveniella poecilophlebia</i>	0.782	0.465	0.476	0.567	24.5
awt	Cryliv	<i>Cryptocarya lividula</i>	0.860	0.505	0.552	0.778	6.1
awt	Cyareb	<i>Cyathea rebeccae</i>	0.838	0.558	0.623	0.744	9.0
awt	Flibou	<i>Flindersia bourjotiana</i>	0.781	0.457	0.484	0.643	10.1
awt	Gomaus	<i>Gomphandra australiana</i>	0.855	0.532	0.542	0.798	5.4
awt	Guiacu	<i>Guioa acutifolia</i>	0.581	0.120	0.235	0.446	16.4
awt	Heralb	<i>Hernandia albiflora</i>	0.926	0.603	0.605	0.848	5.4
awt	Littlee	<i>Litsea leefeana</i>	0.677	0.322	0.411	0.580	9.5
awt	Pulstu	<i>Pullea stutzeri</i>	0.816	0.465	0.550	0.753	5.4
awt	Rhobla	<i>Rhodamnia blairiana</i>	0.894	0.605	0.609	0.831	8.6
awt	Roubra	<i>Rourea brachyandra</i>	0.766	0.430	0.504	0.637	15.6
awt	Syzcor	<i>Syzygium cormiflorum</i>	0.751	0.410	0.370	0.624	8.2
can	alfl	Alder flycatcher	0.547	0.020	0.018	0.364	17.2
can	amcr	American crow	0.609	0.200	0.150	0.583	3.6
can	bhvi	Blue-headed vireo	0.668	0.091	0.070	0.487	17.2
can	blja	Blue jay	0.562	0.089	0.065	0.535	2.3
can	btnw	Black-throated green warbler	0.582	0.076	0.099	0.492	13.8
can	cogr	Common grackle	0.706	0.267	0.223	0.653	9.0
can	eame	Eastern meadowlark	0.697	0.204	0.125	0.670	2.8
can	eato	Eastern towhee	0.758	0.083	0.053	0.704	4.1
can	fisp	Field sparrow	0.715	0.105	0.058	0.666	3.6
can	gcki	Golden crowned kinglet	0.707	0.123	0.081	0.365	46.5
can	hosp	House sparrow	0.799	0.256	0.245	0.743	8.0
can	inbu	Indigo bunting	0.638	0.109	0.071	0.616	3.7
can	lefl	Least flycatcher	0.560	0.039	0.032	0.484	7.2
can	modo	Mourning dove	0.717	0.293	0.210	0.700	2.5
can	oven	Ovenbird	0.631	0.184	0.201	0.445	17.7
can	rbwo	Red-bellied woodpecker	0.899	0.115	0.085	0.871	4.3
can	vira	Virginia rail	0.660	0.020	0.007	0.527	10.9
can	wtsp	White-throated sparrow	0.709	0.242	0.200	0.533	15.7
can	ybfl	Yellow-bellied flycatcher	0.819	0.142	0.101	0.531	41.0

can	ybsa	Yellow-bellied sapsucker	0.704	0.161	0.126	0.586	11.2
nsw	basp1	<i>Chalinolobus gouldii</i>	0.553	0.043	0.066	0.477	11.5
nsw	basp2	<i>Falsistrellus tasmaniensis</i>	0.750	0.265	0.291	0.662	9.5
nsw	basp3	<i>Kerivoula papuensis</i>	0.707	0.140	0.179	0.601	11.1
nsw	basp4	<i>Nyctophilus bifax</i>	0.899	0.222	0.345	0.819	8.4
nsw	basp5	<i>Nyctophilus gouldi</i>	0.567	0.068	0.127	0.499	5.3
nsw	basp6	<i>Vespadelus darlingtoni</i>	0.796	0.338	0.445	0.668	11.4
nsw	basp7	<i>Vespadelus vulturnus</i>	0.704	0.256	0.278	0.583	12.8
nsw	dbsp1	<i>Ptilinopus regina</i>	0.840	0.236	0.281	0.748	6.9
nsw	dbsp2	<i>Calyptorhynchus lathami</i>	0.665	0.177	0.200	0.599	6.4
nsw	dbsp3	<i>Menura novaehollandiae</i>	0.813	0.436	0.434	0.764	5.5
nsw	dbsp4	<i>Lalage leucomela</i>	0.865	0.253	0.279	0.828	4.1
nsw	dbsp5	<i>Monarcha trivirgatus</i>	0.720	0.133	0.235	0.615	10.3
nsw	dbsp6	<i>Pachycephala olivacea</i>	0.974	0.366	0.725	0.937	4.3
nsw	dbsp7	<i>Myzomela sanguinolenta</i>	0.747	0.336	0.319	0.653	7.7
nsw	dbsp8	<i>Corvus tasmanicus</i>	0.767	0.084	0.141	0.611	13.0
nsw	nbsp1	<i>Ninox strenua</i>	0.671	0.186	0.170	0.534	14.7
nsw	nbsp2	<i>Tyto tenebricosa</i>	0.674	0.181	0.139	0.628	3.6
nsw	otsp1	<i>Angophora costata</i>	0.928	0.306	0.288	0.858	5.3
nsw	otsp2	<i>Corymbia gummifera</i>	0.743	0.170	0.125	0.661	8.9
nsw	otsp3	<i>Corymbia intermedia</i>	0.809	0.434	0.400	0.761	4.5
nsw	otsp4	<i>Eucalyptus blakelyi</i>	0.975	0.370	0.397	0.858	15.5
nsw	otsp5	<i>Eucalyptus carnea</i>	0.719	0.231	0.139	0.666	4.7
nsw	otsp6	<i>Eucalyptus fastigata</i>	0.986	0.529	0.579	0.848	17.9
nsw	otsp7	<i>Eucalyptus campanulata</i>	0.815	0.431	0.407	0.716	8.4
nsw	otsp8	<i>Eucalyptus nova-anglica</i>	0.966	0.237	0.219	0.918	6.9
nsw	ousp1	<i>Cassinia quinquefarina</i>	0.907	0.364	0.399	0.735	18.5
nsw	ousp2	<i>Lepidosperma laterale</i>	0.545	0.097	0.076	0.517	3.7
nsw	ousp3	<i>Glycine clandestina</i>	0.578	0.119	0.137	0.545	3.6
nsw	ousp4	<i>Marsdenia liisae</i>	0.884	0.122	0.262	0.695	22.4
nsw	ousp5	<i>Imperata cylindrica</i>	0.562	0.118	0.154	0.499	7.1
nsw	ousp6	<i>Poa sieberiana</i>	0.783	0.479	0.435	0.702	11.8
nsw	ousp7	<i>Eustrephus latifolius</i>	0.577	0.146	0.111	0.492	9.6
nsw	ousp8	<i>Acrotriche aggregata</i>	0.816	0.176	0.163	0.707	10.8
nsw	rtsp1	<i>Alectryon subdentatus</i>	0.587	0.018	0.094	0.436	27.3
nsw	rtsp2	<i>Cupaniopsis anacardiooides</i>	0.968	0.624	0.629	0.808	13.5
nsw	rtsp3	<i>Diploglottis australis</i>	0.582	0.132	0.179	0.527	8.3
nsw	rtsp4	<i>Heritiera actinophylla</i>	0.611	0.160	0.187	0.522	6.8
nsw	rtsp5	<i>Schizomeria ovata</i>	0.697	0.326	0.291	0.641	4.8
nsw	rtsp6	<i>Syzygium luehmanii</i>	0.724	0.138	0.115	0.549	20.7
nsw	rtsp7	<i>Syzygium luehmanii</i>	0.644	0.141	0.166	0.545	8.6
nsw	rusp1	<i>Corokia whiteana</i>	0.966	0.209	0.399	0.808	19.5
nsw	rusp2	<i>Cyathea leichhardtiana</i>	0.719	0.322	0.301	0.670	10.6
nsw	rusp3	<i>Desmodium acanthocladum</i>	0.991	0.613	0.748	0.959	4.0
nsw	rusp4	<i>Dicksonia antarctica</i>	0.858	0.514	0.428	0.744	15.3
nsw	rusp5	<i>Elatostema reticulatum</i>	0.619	0.178	0.132	0.565	5.6
nsw	rusp6	<i>Tasmannia purpurascens</i>	0.996	0.762	0.776	0.940	10.3
nsw	srsp1	<i>Cacophis kreftii</i>	0.889	0.245	0.477	0.729	12.7
nsw	srsp2	<i>Calyptotis scutirostrum</i>	0.796	0.377	0.363	0.749	3.4
nsw	srsp3	<i>Coeranoscincus reticulatus</i>	0.964	0.419	0.556	0.907	4.8
nsw	srsp4	<i>Egernia mcphee</i>	0.766	0.215	0.272	0.613	22.5
nsw	srsp5	<i>Eulamprus murrayi</i>	0.810	0.423	0.412	0.773	5.4
nsw	srsp6	<i>Ophioscincus truncatus</i>	0.923	0.517	0.565	0.849	5.8
nsw	srsp7	<i>Pseudochiropetes porphyriacus</i>	0.623	0.056	0.137	0.519	9.0
nsw	srsp8	<i>Saltuarius swaini</i>	0.984	0.177	0.499	0.738	17.1
nz	CLEFOR	<i>Clematis forsteri</i>	0.799	0.118	0.101	0.728	6.3
nz	COPARE	<i>Coprosma areolata</i>	0.796	0.066	0.135	0.649	19.1
nz	COPCOL	<i>Coprosma colensoi</i>	0.740	0.324	0.268	0.564	15.3
nz	COPLIN	<i>Coprosma linariifolia</i>	0.735	0.208	0.162	0.639	10.4
nz	COPPAR	<i>Coprosma parviflora</i>	0.545	0.006	0.062	0.430	15.2
nz	COPPRO	<i>Coprosma propinqua</i>	0.634	0.142	0.129	0.523	11.1
nz	COPRHA	<i>Coprosma rhamnoides</i>	0.729	0.269	0.286	0.625	9.9
nz	COPSPA	<i>Coprosma spathulata</i>	0.905	0.324	0.280	0.822	11.4
nz	DACDAC	<i>Dacrycarpus dacrydioides</i>	0.852	0.363	0.360	0.808	4.4

nz	DRALAT	<i>Dracophyllum latifolium</i>	0.961	0.423	0.426	0.917	5.9
nz	DRALON	<i>Dracophyllum longifolium</i>	0.838	0.354	0.357	0.736	15.4
nz	DRAMEN	<i>Dracophyllum menziesii</i>	0.932	0.238	0.231	0.892	5.0
nz	DRASUB	<i>Dracophyllum subulatum</i>	0.953	0.087	0.060	0.911	4.4
nz	DRATRA	<i>Dracophyllum traversii</i>	0.856	0.295	0.317	0.652	17.5
nz	DRAUNI	<i>Dracophyllum uniflorum</i>	0.918	0.262	0.359	0.811	13.6
nz	FUCEXC	<i>Fuchsia excorticata</i>	0.612	0.127	0.133	0.537	8.6
nz	HALBID	<i>Halocarpus bidwillii</i>	0.750	0.056	0.135	0.623	11.8
nz	HALBIF	<i>Halocarpus biformis</i>	0.811	0.275	0.228	0.711	13.1
nz	HEBCOR	<i>Hebe corriganii</i>	0.943	0.170	0.181	0.880	9.2
nz	HEBSAL	<i>Hebe salicifolia</i>	0.711	0.198	0.196	0.636	11.3
nz	HEBSTR	<i>Hebe stricta</i>	0.708	0.115	0.097	0.618	10.3
nz	LEPSCO	<i>Leptospermum scoparium</i>	0.642	0.114	0.111	0.542	10.2
nz	LIBBID	<i>Libocedrus bidwillii</i>	0.800	0.261	0.206	0.687	14.3
nz	LIBPLU	<i>Libocedrus plumosa</i>	0.905	0.127	0.196	0.860	5.3
nz	LOPOBC	<i>Lophomyrtus obcordata</i>	0.788	0.147	0.271	0.721	6.7
nz	MELMIC	<i>Melicytus micranthus</i>	0.753	0.057	0.045	0.694	9.5
nz	METALB	<i>Metrosideros albiflora</i>	0.913	0.262	0.379	0.803	12.3
nz	METCOL	<i>Metrosideros colensoi</i>	0.921	0.185	0.389	0.816	9.7
nz	METDIF	<i>Metrosideros diffusa</i>	0.798	0.422	0.394	0.707	8.7
nz	METEXC	<i>Metrosideros excelsa</i>	0.982	0.204	0.383	0.900	18.3
nz	METFUL	<i>Metrosideros fulgens</i>	0.900	0.533	0.564	0.825	10.0
nz	METPER	<i>Metrosideros perforata</i>	0.812	0.366	0.306	0.759	7.0
nz	METROB	<i>Metrosideros robusta</i>	0.881	0.446	0.413	0.815	4.1
nz	METUMB	<i>Metrosideros umbellata</i>	0.844	0.509	0.507	0.719	15.4
nz	MYRDIV	<i>Myrsine divaricata</i>	0.706	0.338	0.319	0.547	17.0
nz	NESCUN	<i>Nestegis cunninghamii</i>	0.856	0.218	0.193	0.747	11.8
nz	NESLAN	<i>Nestegis lanceolata</i>	0.878	0.270	0.231	0.788	11.4
nz	NOTFUS	<i>Nothofagus fusca</i>	0.758	0.309	0.290	0.676	11.1
nz	NOTMEN	<i>Nothofagus menziesii</i>	0.601	0.150	0.166	0.476	13.9
nz	NOTSOL	<i>Nothofagus solandri</i>	0.759	0.144	0.083	0.660	10.1
nz	PENCOR	<i>Nothofagus truncata</i>	0.788	0.225	0.218	0.731	7.6
nz	PHYALP	<i>Phyllocladus alpinus</i>	0.762	0.336	0.282	0.696	10.5
nz	PHYTRI	<i>Phyllocladus trichomanoides</i>	0.956	0.470	0.426	0.930	4.4
nz	PODHAL	<i>Podocarpus hallii</i>	0.664	0.265	0.247	0.570	11.2
nz	PODNIV	<i>Podocarpus nivalis</i>	0.875	0.245	0.185	0.773	16.6
nz	PODTOT	<i>Podocarpus totara</i>	0.673	0.096	0.091	0.563	14.4
nz	PRUFER	<i>Prumnopitys ferruginea</i>	0.821	0.531	0.497	0.776	5.1
nz	PRUTAX	<i>Prumnopitys taxifolia</i>	0.788	0.179	0.177	0.704	8.1
nz	RUBAUS	<i>Rubus australis</i>	0.677	0.130	0.146	0.641	3.8
nz	RUBSCH	<i>Rubus schmideliooides</i>	0.625	0.074	0.059	0.574	5.9
nz	SYZMAI	<i>Syzygium maire</i>	0.949	0.198	0.150	0.747	23.1
nz	WEIRAC	<i>Weinmannia racemosa</i>	0.836	0.541	0.531	0.758	9.7
sa	adenimpr	<i>Adenocalymma impressum</i>	0.964	0.569	0.702	0.873	5.8
sa	amphpani	<i>Amphilophium paniculatum</i>	0.656	0.157	0.238	0.585	9.5
sa	arraffi	<i>Arrabidaea affinis</i>	0.906	0.404	0.400	0.827	11.3
sa	arrabrac	<i>Arrabidaea brachypoda</i>	0.829	0.463	0.504	0.772	3.9
sa	arrachic	<i>Arrabidaea chica</i>	0.647	0.156	0.302	0.553	11.5
sa	arracinn	<i>Arrabidaea cinnomomea</i>	0.851	0.449	0.435	0.805	4.7
sa	arraplat	<i>Arrabidaea platyphylla</i>	0.954	0.472	0.551	0.903	5.2
sa	arrapulc	<i>Arrabidaea pulchra</i>	0.988	0.722	0.720	0.927	7.3
sa	arrasell	<i>Arrabidaea selloi</i>	0.919	0.398	0.442	0.883	4.0
sa	arratrip	<i>Arrabidaea triplinervia</i>	0.895	0.338	0.445	0.713	12.6
sa	calllati	<i>Callichlamys latifolia</i>	0.647	0.202	0.253	0.594	5.6
sa	ceratetr	<i>Ceratophytum tetragonolobum</i>	0.937	0.449	0.509	0.834	6.8
sa	clytsciu	<i>Clytostoma sciuripabulum</i>	0.817	0.331	0.473	0.677	13.9
sa	cusplate	<i>Cuspidaria lateriflora</i>	0.919	0.521	0.528	0.756	10.7
sa	cydiaequ	<i>Cydista aequinoctialis</i>	0.821	0.368	0.401	0.770	3.9
sa	distmagn	<i>Distictella magnoliifolia</i>	0.888	0.338	0.433	0.797	7.9
sa	fridspec	<i>Fridericia speciosa</i>	0.910	0.393	0.447	0.862	2.9
sa	lundcord	<i>Lundia cordata</i>	0.908	0.545	0.548	0.851	4.5
sa	lundvirg	<i>Lundia virginialis</i>	0.883	0.389	0.441	0.836	4.3
sa	macfungu	<i>Macfadyena unguis-cati</i>	0.769	0.371	0.413	0.671	12.3
sa	mansdiff	<i>Mansoa diffcilis</i>	0.922	0.418	0.482	0.864	3.7

sa	mansverr	<i>Mansoa verrucifera</i>	0.811	0.433	0.495	0.737	10.5
sa	martobov	<i>Martinella obovata</i>	0.767	0.208	0.241	0.645	13.3
sa	mellquad	<i>Mellooa quadrivalvis</i>	0.797	0.282	0.403	0.760	4.0
sa	parapyra	<i>Paragonia pyramidata</i>	0.713	0.326	0.378	0.641	8.2
sa	phrycory	<i>Phryganocydia corymbosa</i>	0.890	0.455	0.528	0.826	7.9
sa	pithcruc	<i>Pithecoctenium crucigerum</i>	0.727	0.280	0.340	0.568	16.7
sa	pleomeli	<i>Pleonotoma melioides</i>	0.963	0.583	0.646	0.912	3.7
sa	tananoct	<i>Tanaecium nocturnum</i>	0.813	0.288	0.369	0.691	15.8
sa	tynaschu	<i>Tynanthus schumannianus</i>	0.917	0.493	0.598	0.808	8.0
swi	abialb	<i>Abies alba</i>	0.787	0.469	0.404	0.727	4.8
swi	acecam	<i>Acer campestre</i>	0.876	0.199	0.155	0.836	4.9
swi	acepla	<i>Acer platanoides</i>	0.850	0.174	0.174	0.788	7.2
swi	acepse	<i>Acer pseudoplatanus</i>	0.727	0.276	0.233	0.682	6.4
swi	alnglu	<i>Alnus glutinosa</i>	0.796	0.146	0.134	0.768	2.8
swi	alninc	<i>Alnus incana</i>	0.667	0.109	0.073	0.587	8.6
swi	betpen	<i>Betula pendula</i>	0.772	0.247	0.253	0.695	8.3
swi	carbet	<i>Carpinus betulus</i>	0.916	0.297	0.237	0.891	2.9
swi	cassat	<i>Castanea sativa</i>	0.989	0.752	0.729	0.956	5.2
swi	fagsyl	<i>Fagus sylvatica</i>	0.836	0.573	0.504	0.779	5.0
swi	fraexc	<i>Fraxinus excelsior</i>	0.777	0.317	0.242	0.730	4.3
swi	lardec	<i>Larix decidua</i>	0.813	0.479	0.466	0.709	11.7
swi	ostcar	<i>Ostrya carpinifolia</i>	0.990	0.545	0.544	0.968	2.6
swi	picabi	<i>Picea abies</i>	0.799	0.483	0.428	0.701	4.9
swi	pincem	<i>Pinus cembra</i>	0.973	0.514	0.517	0.948	3.4
swi	pinmug	<i>Pinus mugo</i>	0.865	0.128	0.243	0.794	5.7
swi	pinsyl	<i>Pinus sylvestris</i>	0.808	0.359	0.287	0.759	7.9
swi	pinunc	<i>Pinus uncinata</i>	0.738	0.099	0.095	0.652	8.9
swi	popnig	<i>Populus nigra</i>	0.957	0.107	0.131	0.882	8.2
swi	poptre	<i>Populus tremula</i>	0.693	0.118	0.129	0.646	7.5
swi	pruavi	<i>Prunus avium</i>	0.778	0.148	0.087	0.736	3.7
swi	quepet	<i>Quercus petraea</i>	0.865	0.325	0.285	0.827	4.6
swi	quepub	<i>Quercus pubescens</i>	0.952	0.162	0.136	0.881	10.9
swi	querob	<i>Quercus robur</i>	0.853	0.266	0.202	0.828	1.7
swi	salalb	<i>Salix alba</i>	0.752	0.060	0.085	0.648	6.7
swi	sorari	<i>Sorbus aria</i>	0.763	0.165	0.130	0.715	8.0
swi	sorauc	<i>Sorbus aucuparia</i>	0.772	0.144	0.107	0.695	7.9
swi	tilcor	<i>Tilia cordata</i>	0.871	0.236	0.196	0.818	7.1
swi	tilpla	<i>Tilia platyphyllos</i>	0.866	0.167	0.127	0.810	7.7
swi	ulmgla	<i>Ulmus glabra</i>	0.811	0.208	0.150	0.749	8.4
	min		0.488	0.005	0.007	0.364	1.7
	max		0.996	0.762	0.776	0.968	46.5
	mean		0.788	0.292	0.310	0.701	9.40

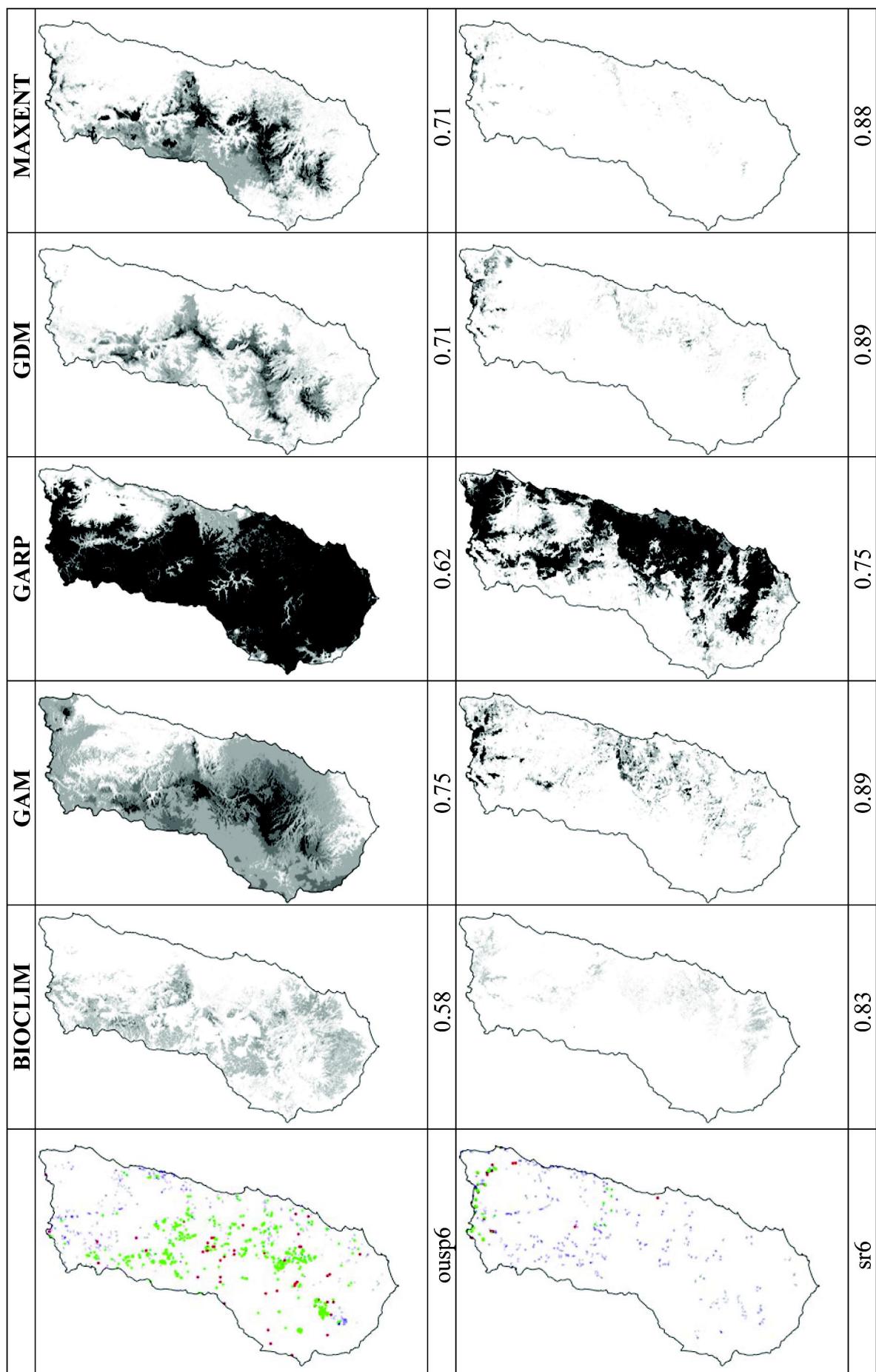
Table S4. Average maximum KAPPA values summarized over methods and regions.

	AWT	CAN	NSW	NZ	SA	SWI
BIOCLIM	0.27	0.07	0.13	0.08	0.31	0.14
BRT	0.29	0.07	0.20	0.16	0.35	0.22
BRUTO	0.27	0.07	0.18	0.17	0.25	0.20
DKGARP	0.27	0.02	0.12	na	0.18	0.07
DOMAIN	0.26	0.06	0.19	0.10	0.30	0.13
GAM	0.28	0.06	0.16	0.16	0.27	0.20
GDM	0.31	0.08	0.23	0.16	0.34	0.17
GDMSS	0.33	0.07	0.19	0.16	0.32	0.19
GLM	0.29	0.06	0.16	0.15	0.25	0.18
LIVES	0.26	0.06	0.18	0.09	0.31	0.14
MARS	0.29	0.06	0.17	0.17	0.30	0.20
MARS-COMM	0.30	0.09	0.21	0.17	0.29	0.22
MARS-INT	0.28	0.06	0.16	0.15	0.31	0.20
MAXENT	0.30	0.06	0.19	0.17	0.31	0.21
MAXENT-T	0.31	0.06	0.19	0.16	0.30	0.22
OMGARP	0.30	0.04	0.14	0.11	0.27	0.15
MEAN	0.29	0.06	0.18	0.14	0.29	0.18

Table S5. Rankings of methods on a regional basis.

	AWT	CAN	NSW	NZ	SA	SWI						
MARS-COMM	8.0	9	5.2	1	5.6	1	6.3	3	7.7	8	3.3	1
BRT	7.8	5	5.3	4	7.3	5	6.9	4	4.8	1	2.8	2
MAXENT-T	6.5	3	7.2	6	6.1	3	6.7	5	7.9	7	4.5	3
MAXENT	7.5	6	7.7	7	6.5	4	6.5	2	7.3	4	4.9	4
GDM-SS	6.8	1	8.5	11	7.6	6	7.3	7	7.2	3	7.5	7
GDM	8.8	7	8.7	9	5.9	2	7.0	1	6.3	2	10.1	11
GAM	8.9	14	10.0	15	9.1	10	6.6	6	10.0	11	6.0	5
GLM	9.0	10	9.1	8	8.3	8	7.1	10	11.3	15	8.6	9
DOMAIN	8.6	8	8.6	5	7.3	7	9.2	13	7.4	5	12.0	13
BRUTO	9.9	16	9.5	13	9.1	11	6.7	8	10.8	14	7.9	8
MARS	9.4	12	9.5	12	10.5	12	7.2	9	9.3	12	7.6	6
OM-GARP	7.1	2	10.9	14	9.3	9	9.1	12	7.9	6	10.2	12
MARS-INT	9.7	15	9.9	16	11.5	15	8.4	11	10.0	16	8.4	10
LIVES	9.6	11	8.9	3	9.3	14	11.0	14	8.5	9	13.8	14
DK-GARP	8.4	4	10.5	10	10.2	13	NA	NA	9.6	13	14.6	16
BIOCLIM	9.1	13	6.5	2	11.1	16	13.1	15	9.0	10	13.7	15

For each region, the first column shows the mean of the per species AUC ranks, and the second column shows the rank of the mean AUC over all species. Best performance = smallest rank. Methods sorted by mean AUC rank over all regions, as for Table 2 of the manuscript.



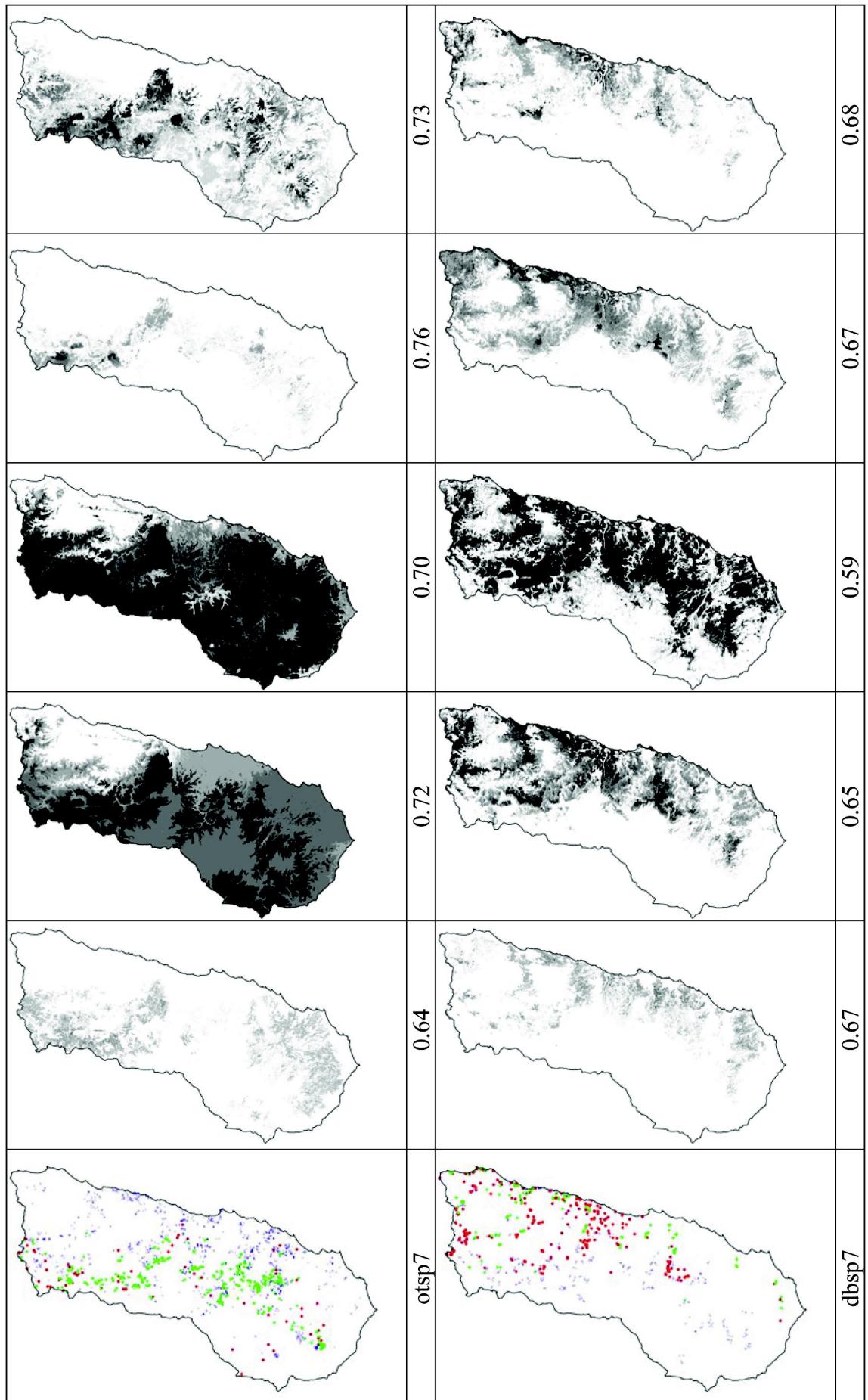


Fig. S1. Maps for four species from NSW for each of five selected techniques: ousp6, *Poa sieberiana* (53 records for modelling and 512 presence/797 absence for evaluation); srsp6 *Ophiocincus truncatus* (79 model, 74/932 eval); otsp7, *Eucalyptus campanulata* (69 model, 400/1636 eval); dbsp7, *Myzomela sanguinolenta* (315 model, 161/541 eval). The first column shows modelling sites (red) and evaluation sites: presence = green, absence = blue. The numbers are AUC scores.

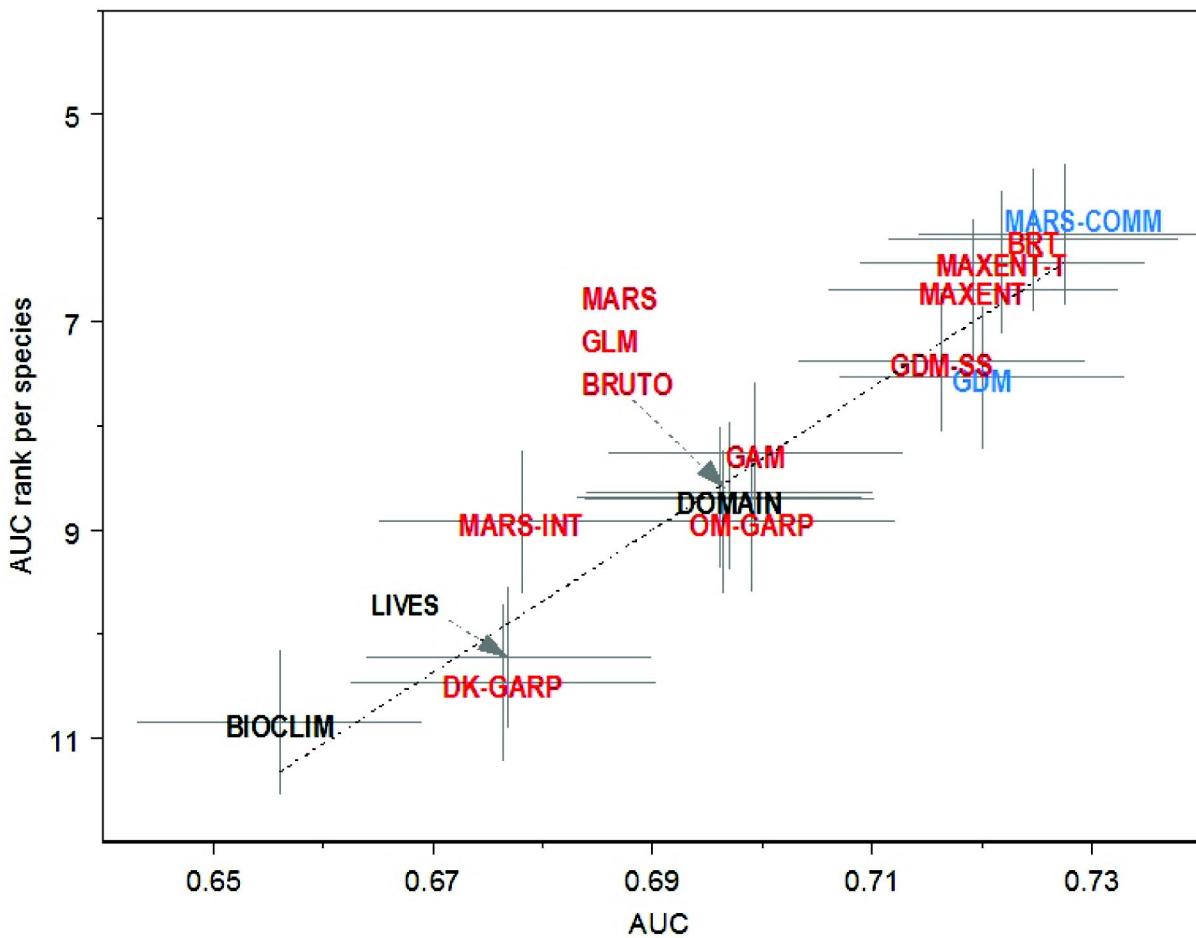


Fig. S2. Mean AUC vs the rank of the method when AUCs were assessed on a per-species basis. Low ranks report methods that are consistently one of the best; ranks compare methods without referring to the actual differences in AUC value. Grey bars designate standard errors for an average species in an average region, as estimated in a generalized linear mixed model. The dotted black line is the line of best fit between the mean AUC and mean AUC rank. The colours are broad classifications of the methods: black = only use presence data, red = use presence and background samples , blue = community methods.

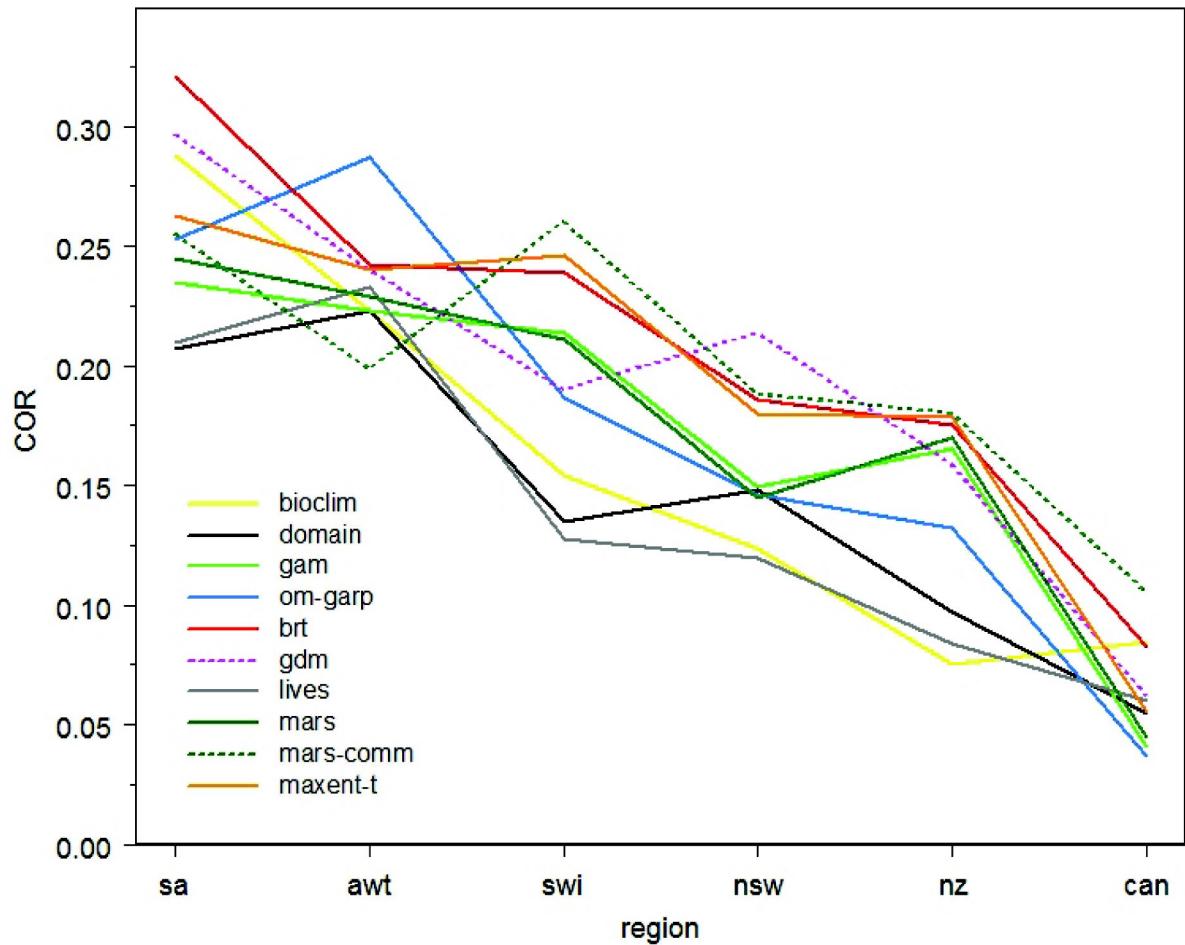


Fig. S3. Predictive success measured by COR, across regions, for 10 methods. Regions are sorted by the mean COR across all 16 methods and all species per region.

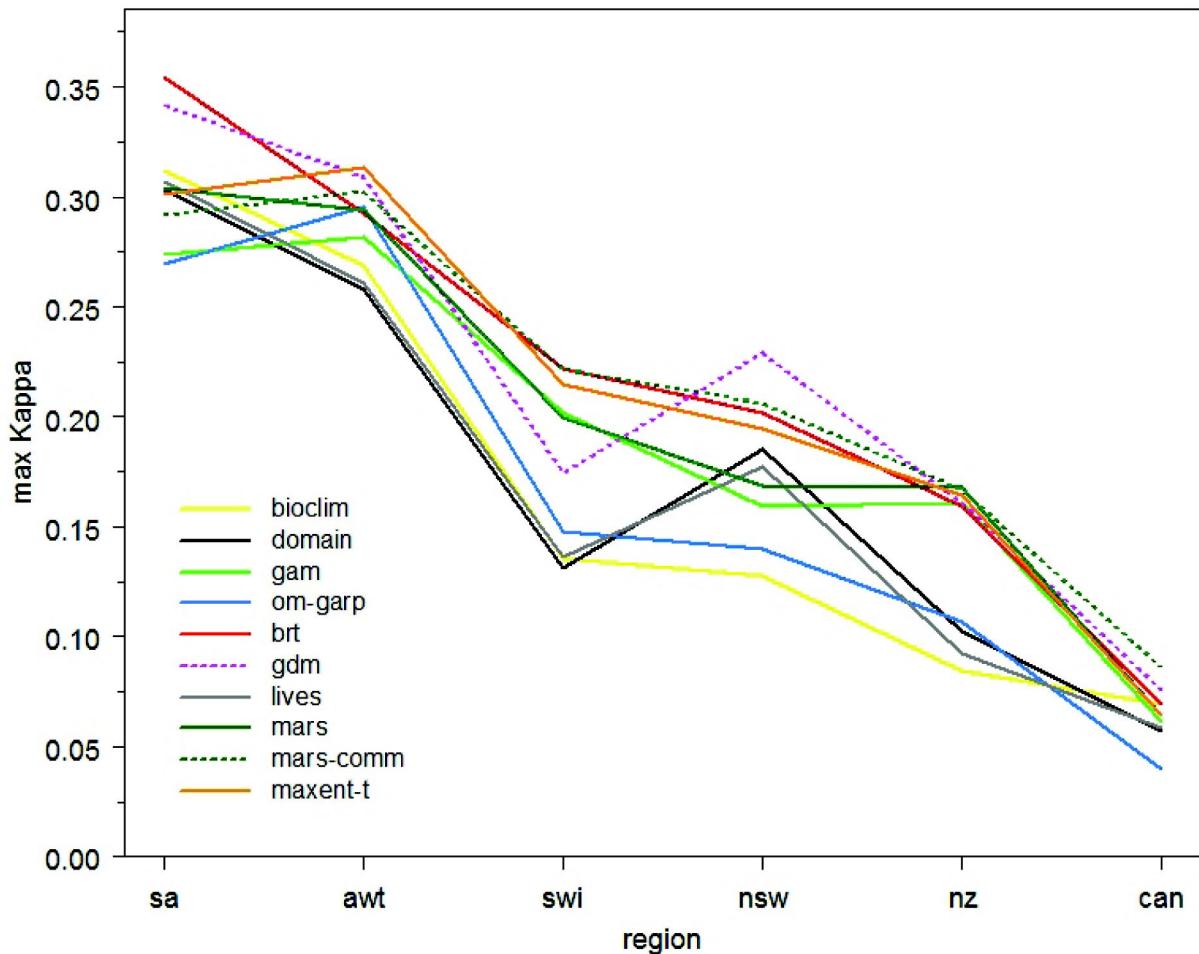


Fig. S4. Predictive success measured by KAPPA, across regions, for 10 methods. Regions are sorted by the mean KAPPA across all 16 methods and all species per region.

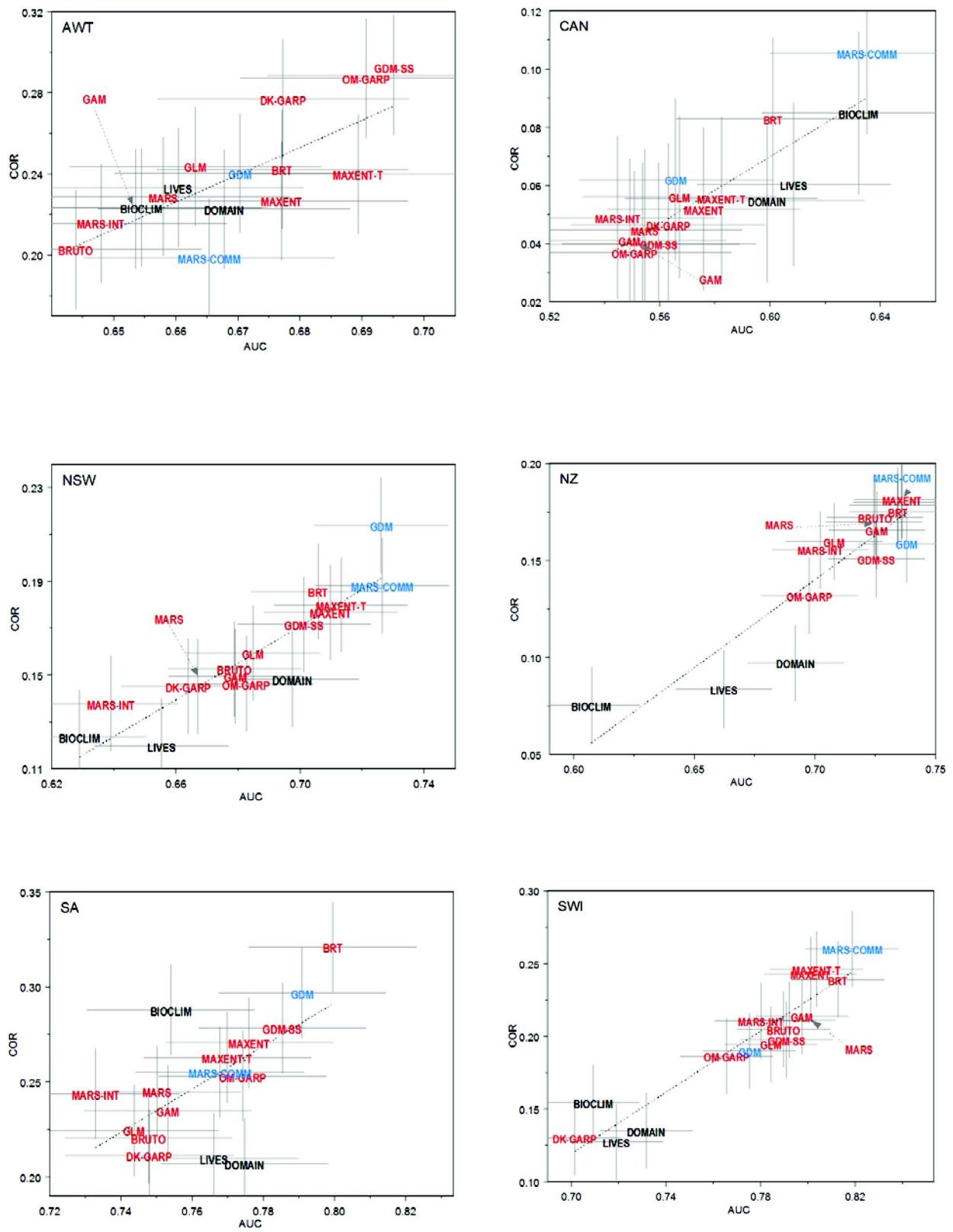


Fig. S5. Mean AUC vs mean COR, on a regional level. Format follows Fig. S2. Note that the axes are scaled differently between regions.

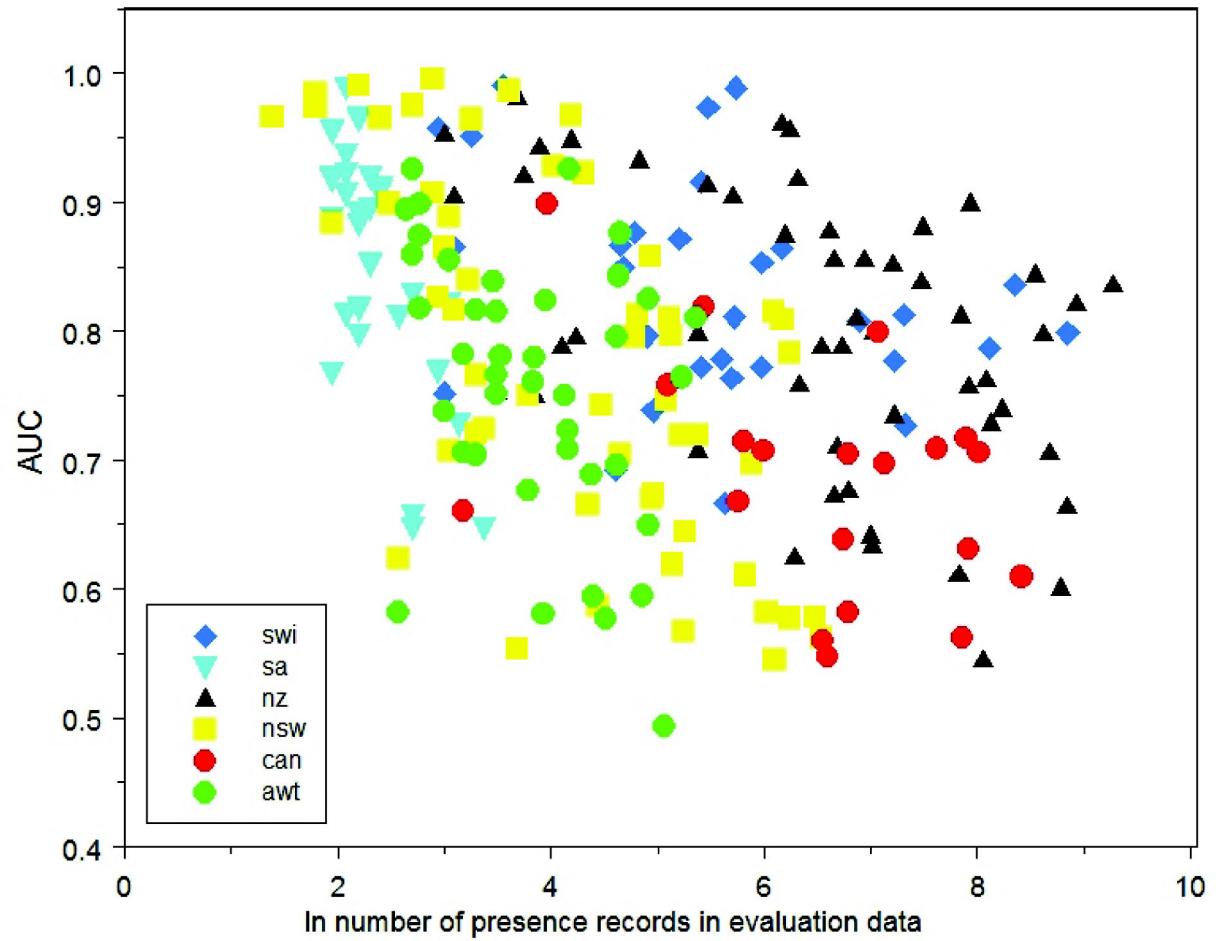


Fig. S6. Variation in maximum AUC with (log) number of presences in evaluation data. Colours identify the regions, and each point represents a species.

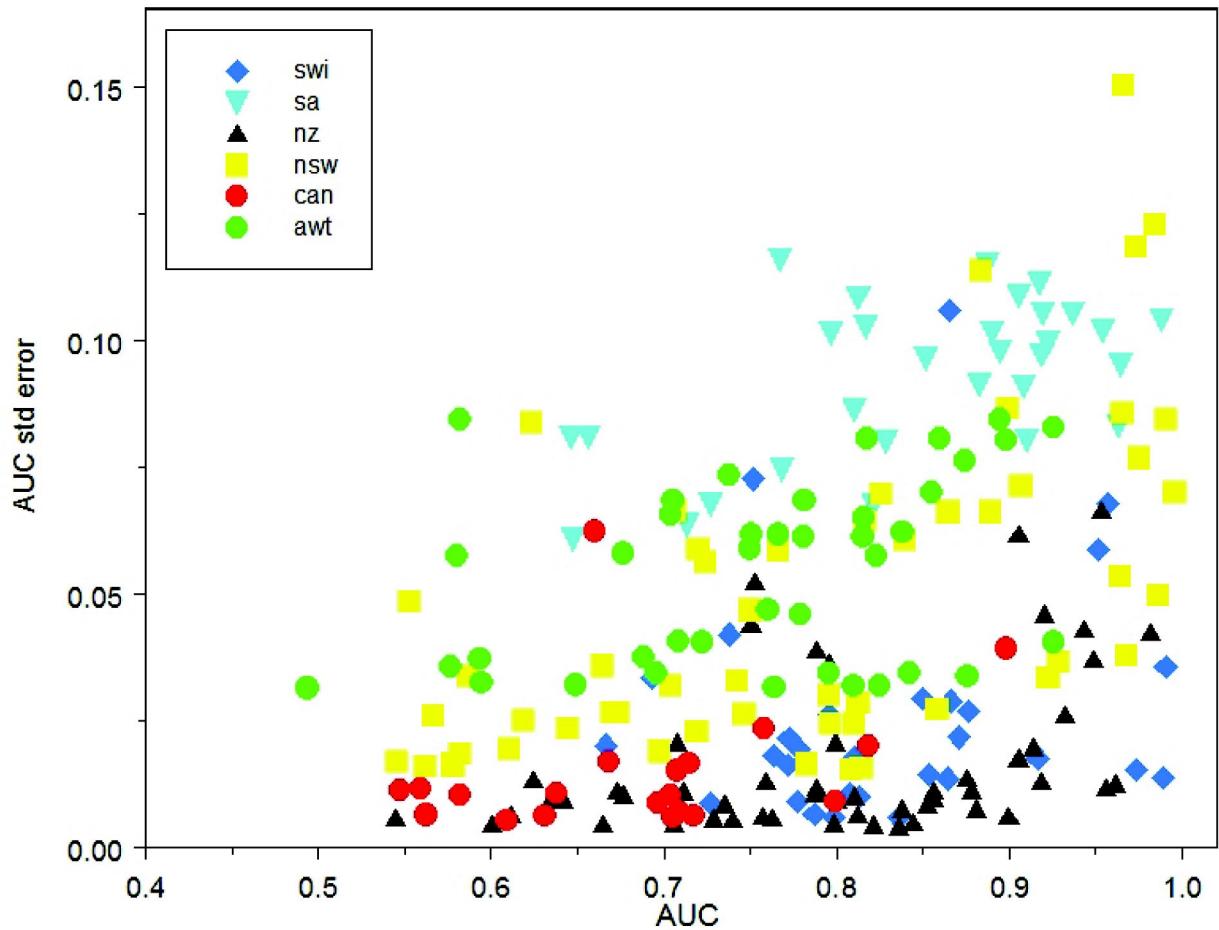


Fig. S7. Maximum AUC vs AUC standard error. Colours identify the regions, and each point represents a species.

Text S1. Details of methods and their application.

Full name of method: bioclimatic envelope model

Abbreviation: BIOCLIM

Alternative names: climate envelope

Implementations in this study: one only

Key references: Busby 1991

Examples of implementation in ecology: Lindenmayer et al. 1991, Hughes et al. 1996, Kadmon et al. 2003

Brief description: BIOCLIM is a profile matching method. It uses species presence records without reference to the background or to any form of absence. The species profile summarizes how the known presences are distributed with respect to the environmental variables. With several environmental variables, the aggregated profile forms a multidimensional space (a hyper-rectangle or “environmental envelope”) that defines the environmental domain of the species. This envelope specifies the model in terms of percentiles or upper and lower tolerances, and does not allow for regions of absence (i.e. “holes”) within the envelope. The concept is one of extremes and cores. A habitat map can be produced from the model by ranking each location according to its position in the species’ environmental profile. Commonly these maps are grid-based and classify each cell into one of several ranked classes of environmental suitability for the species. The DIVA-GIS (Hijmans et al. 2004) version is an implementation of the BIOCLIM method that can use all predictor variables (not just climate ones), and that produces predictions as percentiles.

Software used: DIVA-GIS

Settings: Default BIOCLIM settings

Specifics of data manipulations for modelling: all variables used

Predictions (range, increments): 1:50, continuous

Full name of method: boosted regression trees

Abbreviation: BRT

Alternative names: Stochastic gradient boosting

Implementations in this study: one only

Key references: Friedman et al. 2000, Friedman 2001, 2002, Schapire 2003

Examples of implementation in ecology: Leathwick et al. in press. **Brief description:** Boosted regression trees combine two algorithms: “boosting” is a method for developing multiple models and combining them; “regression trees” are single models that partition the predictor space into disjoint regions and predict a separate constant value in each of them (Friedman and Meulman 2003). Boosting is used to overcome the inaccuracies of a single model, and makes it possible to model a complex response surface. Regression trees can use continuous and categorical predictor variables, allow for missing data, are not sensitive to outliers, tend to exclude irrelevant variables, and model interactions.

BRT are described in different ways in different disciplines. The foremost interpretation from the machine learning community is that it is a method for finding many rough rules of thumb (i.e. many regression trees) that, when combined, are more accurate than any single rule. The boosting algorithm calls the regression tree algorithm repeatedly, each time giving it a re-weighted version of the data that emphasizes the records that were misclassified in the last round. Finally the suite of trees are combined by weighted averaging (Schapire 2003). Statisticians have reinterpreted it as a method for developing a regression model in a forward stage-wise fashion, adding small modifications across the model space (via trees) to fit the data better (Hastie et al. 2001). The final model has numerous terms, each term being a regression tree. Whatever the interpretation, the focus in model development is the same. As boosting proceeds, the model complexity increases until eventual-

ly it over-fits the data. In the gradient boosted methods (Friedman 2002) the aim is to maximize the log-likelihood, and updates are based on its gradient. The number of trees in the boosted model is a natural measure of complexity, and is chosen by measuring prediction accuracy on independent data. This identifies the most complex model that still predicts well, and is based on the trade-off between training error and generalization error.

The two main parameters to be set are the shrinkage parameter (learning rate), which controls the amount of re-weighting at each step, and the size of each tree – one partition (an additive model) or two or more splits. BRT is implemented in gbm (see below) for several response types, including binomial families. To model presence-only data we used the random background samples in place of “absence” records.

Software used: R version 2.0.1; gbm library version 1.5 (author: Greg Ridgeway); extra code written to run all species in one batch

Settings: learning rate = 0.001, interaction depth = 5, select number of trees via 5-fold cross-validation up to a maximum of 10000, weight pseudo-absences so total weight for absences = total weight for presences.

Specifics of data manipulations for modelling: excluded highly correlated variables (Table S2)

Predictions (range, increments): 0 to 1, continuous.

Full name of method: BRUTO

Abbreviation: na

Alternative names: flexible discriminant analysis

Implementations in this study: one only

Key references: Hastie and Tibshirani 1996

Examples of implementation in ecology: Leathwick et al. unpubl.

Brief description: BRUTO (available in the mda library for both S-Plus and R) fits a generalized additive model (GAM, see below) using an adaptive back-fitting procedure with smoothing splines. In large data sets it is ca 100 times faster at fitting a model than a GAM (Leathwick et al. unpubl.). In addition to identifying which variables to include in the final model, BRUTO identifies the optimal degree of smoothing for each variable. BRUTO also allows specification of a penalty parameter that is applied to the addition of extra variables in the model. The model selection is based on an approximation to the generalized cross-validation (GCV) criterion, which is used at each step of the back-fitting procedure. Once the selection process stops, the model is backfit using the chosen amount of smoothing. However, because BRUTO can only be used to fit models assuming Gaussian errors, model parameters describing the selected variables and their degree of smoothing were extracted and used to specify a model of identical form but allowing for binomial errors, and this was fitted using the standard GAM function (“gam”) in Splus. To model presence-only data we used the random background samples in place of “absence” records. Currently BRUTO code does not allow use of categorical variables.

Software used: Splus, mda (bruto function) and gam libraries; extra code written to link the bruto output to the gam, and to allow modelling of all species in one batch. We attempted to use bruto in R but could not get the code available in Dec 2004 to run properly.

Settings: The default penalty parameter (2) was used; weight background samples (“absences”) so total weight for absences = total weight for presences.

Specifics of data manipulations for modelling: excluded highly correlated and categorical variables, plus variables with 3 or fewer unique values (Table S2).

Predictions (range, increments): 0:1, continuous.

Full name of method: DOMAIN

Abbreviation: na

Alternative names: na

Implementations in this study: one only

Key references: Carpenter et al. 1993

Examples of implementation in ecology: Carpenter et al. 1993, Loiselle et al. 2003

Brief description: DOMAIN estimates the environmental similarity (the complement of the distance) between a site of interest and the nearest presence record in environmental space. It uses species presence records without reference to the background or to any form of absence. DOMAIN uses the Gower metric, a distance measure that standardizes each variable by its range over all presence sites to equalise the contribution of all variables. DOMAIN can be used to specify an environmental envelope by selecting a minimum threshold of similarity, or it can be used to map similarities on a continuous scale. We used an implementation in DIVA-GIS rather than the original program

Software used: DIVA-GIS

Settings: default

Specifics of data manipulations for modelling: All variables used

Predictions (range, increments): ≤100, continuous

Full name of method: Generalized additive models

Abbreviation: GAMs

Alternative names: na

Implementations in this study: one only

Key references: Hastie and Tibshirani 1990 (GAMs), Lehmann, et al. 2003 (GRASP)

Examples of implementation in ecology: Yee and Mitchell 1991, Bio et al. 1998

Brief description: GAMs are multiple regression models (see GLMs) in which non-parametric smooth functions are used to model non-linear relationships. They share a number of features with GLMs, including: able to deal with categorical data; can include a mixture of linear and non-linear fitted functions; can model a variety of response types, including binomial and poisson. A range of alternative smoothers are available. GAMs are usually fitted through a back-fitting algorithm with a Newton-Raphson procedure, and in ecology the most common model selection method involves a stepwise procedure where successively simpler fits are compared with a measure such as Akaike's Information Criterion (AIC). To model presence-only data we used the random background samples in place of "absence" records.

Software used: S-PLUS v 6.x, with GRASP package

Settings: Predictor data set first reduced to variables not too highly correlated (Table S2) then models selected with both directions stepwise search, starting from full model. Allowed steps for fitted functions for continuous variables were: smoothed (cubic β -spline) with 4 degrees of freedom (df), linear fit, omitted. Categorical variables used as factors. No interactions modeled. AIC used as stopping criterion. The 10000 background samples ("absence") weighted so total weight for presence = total weight for absence.

Specifics of data manipulations for modelling: excluded highly correlated variables (Table S2) on this basis: CAN and AWT (Modeler A. Lehmann) and SWI and SA (Modeler A. Guisan): uncorrelated variables selected by removing correlated ones ($r < 0.80$) from right to left in the order of the original dataset; NSW and NZ (Modeler J. Elith): uncorrelated variables were selected by removing correlated ones ($r < 0.85$) that were judged by expert knowledge to be the least proximal ones.

Predictions (range, increments): 0:1, continuous.

Full name of method: genetic algorithm for rule-set prediction

Abbreviation: GARP

Alternative names: none

Implementations in this study: Desktop GARP (DK-GARP), Open Modeler GARP (OM-GARP)

Key references: Stockwell and Noble 1992, Stockwell and Peters 1999

Examples of implementation in ecology: Anderson et al. 2002, Peterson et al. 2004, 2006

Brief description: GARP represents an implementation of a genetic algorithm for identifying associations between known occurrences and a set of raster GIS coverages that summarize aspects of the environment. GARP uses a suite of four tools to produce initial hypotheses, including BIOCLIM rules and two related set-based rule types, as well as a very simple logistic regression analogue. These initial rules are modified in an "evolutionary" process, in which elements of rules are modified at random. The algorithm runs through 10^2 – 10^3 iterations of modification until further changes to rules do not improve rule fitness. When this "convergence" occurs, the model is used to characterize the entire landscape as to being within the modeled niche or not. To take into account the model to model variation that enters owing to the random selection of data for rule training and rule evaluation, as well as because of the random-walk nature of the genetic algorithm, many replicate models are produced, and the most useful models identified using the "best subsets" procedure (Anderson et al. 2003).

The OM-GARP algorithm used for this research is still in its testing phase, and not generally available to the public. An OM version of the Desktop GARP algorithm is publicly available, but was not tested here.

Software used: DesktopGARP version 1.1.6; <<http://www.lifemapper.org/desktopgarp>>.

Settings: All default settings used for model development; best subsets functionality activated, 20% soft threshold for omission, 50% commission threshold.

Specifics of data manipulations for modelling: Geographic data were processed into "GARP data sets" using the GARP Dataset Manager module that is available with the program.

Predictions (range, increments): DK=GARP: as integers from 0 to 10. OM-GARP: 0–100, continuous.

Full name of method: Generalized dissimilarity modelling

Abbreviation: GDM

Alternative names: na

Implementations in this study: community model (GDM), single species model (GDM-SS)

Key references: Ferrier 2002, Ferrier et al. 2002

Examples of implementation in ecology: Ferrier et al. 2004

Brief description: GDM models spatial turnover in community composition (i.e. "compositional dissimilarity", quantified with a Bray-Curtis measure) between pairs of sites as a function of environmental differences between these sites. GDM is an extension of matrix regression that addresses the problem of realistically modelling the non-linear responses common in ecological data. The first type of non-linearity is that the relationship between ecological separation and compositional dissimilarity is curvilinear, so a GLM with appropriate link and variance functions (rather than ordinary linear regression) is used within the matrix regression. The second non-linearity relates to the rate of compositional change, or "turnover", along environmental gradients. In ordinary matrix regression this rate of change is assumed constant along the gradient; in GDM it is allowed to be non-linear through use of monotonic I-splines. The splines are used to fit a transforming

function to each environmental variable that maximizes the reduction in deviance achieved by its inclusion. For predicting species distributions, an additional kernel regression algorithm (Lowe 1995) is applied within the transformed environmental space generated by GDM, to estimate likelihoods of occurrence of a given species at all sites.

Two versions of this approach were applied in the current study: 1) "GDM" in which a single GDM was fitted to the combined data for all species in a given biological group, such that the output from this GDM was then used as a common basis for all of the subsequent kernel regression analyses; and 2) "GDM-SS" in which a separate GDM was fitted to the data for each species alone, such that kernel regression analysis for each species was based on the output from a GDM tailored specifically to that species. Note that the first uses the data for broad functional groups (eg all birds in a region) and assigns absence to a site if a species is not recorded there – i.e. it uses "community" data. This is different to what other single-species methods used. However, to make it as comparable to single-species implementations as possible, we used the random background samples in the kernel regression stage, rather than the absences in the community data. The second implementation, GDM-SS, used only single species presence records plus random background samples.

Software used: Scripts written by Manion and Ferrier (Ferrier unpubl.), and run through ArcView and S-PLUS.

Settings: see description. No Euclidean distances used. Sub-sample of 2000 site pairs used in matrix regression. Sub-sample of 1000 of the 10000 random points used for kernel regression stage of GDM and for GDM-SS. No weighting for the Bray-Curtis measure.

Specifics of data manipulations for modelling: excluded highly correlated and categorical variables (Table S2). Used the functional groups listed in Table 1 for community models.

Predictions (range, increments): 0:1, continuous.

Full name of method: Generalized linear models

Abbreviation: GLMs

Alternative names: logistic regression, poisson regression etc.

Implementations in this study: one only

Key references: McCullagh and Nelder 1989

Examples of implementation in ecology: Austin et al. 1983, Winetle et al. 2005

Brief description: GLMs are a broad class of statistical models that include linear regression and analysis of variance. All GLMs have a response (the species data for models of distribution), one or more predictors (the explanatory variables, commonly environmental data) and a link function that describes the relationship between the expected value of the response and the predictors. Species distribution models are often constructed from presence-absence species data modeled with logistic regression – i.e. a GLM for data with a binomial distribution, with a logit link function. However, a wide variety of data can be accommodated by specifying different distributions for the response and different link functions. GLMs are able to model relationships of varying complexity between the response and a predictor variable by specifying linear, beta, polynomial or other functions. Categorical predictors can be included as factor variables. GLMs are fitted with Maximum Likelihood Estimation (Hastie et al. 2001), and in ecology the most common model selection method involves a stepwise procedure where successively simpler fits are compared with a measure such as Akaike's Information Criterion (AIC). To model presence-only data we used the random background samples in place of "absence" records.

Software used: S-PLUS v 6.x, with GRASP package

Settings: as for GAMs, except the allowed steps for continuous variables were: cubic polynomial, linear fit, omitted.

Specifics of data manipulations for modelling: as for GAMs

Predictions (range, increments): 0:1, continuous

Full name of method: Limiting Variable and Environmental Suitability

Abbreviation: LIVES

Alternative names: na

Implementations in this study: one only

Key references: Li and Hilbert unpubl.

Examples of implementation in ecology: Li and Hilbert unpubl.

Brief description: The ecological basis for LIVES is limiting factor theory (LFT) that postulates that the occurrence of a species is only determined by the factor that most limits its distribution. Unlike niche theory, LFT only considers the occurrence of a species rather than its abundance or frequency, so LIVES uses species presence records without reference to the background or to any form of absence. LIVES assumes: 1) all environmental factors are equally important and their effects on a species' distribution are determined by the magnitude of their difference between the grid cell for which a prediction is desired and the sites where presences are recorded. This can be measured using a similarity index; 2) the limiting factor of the species is defined as the environmental factor that has the minimum similarity (or maximum variation) between the predicted site and the presence sites for all environmental factors considered in the model; 3) the limiting factor is considered as the most important factor that determines the suitability of a site to a species, i.e. the distribution of the species; and 4) the lower and upper limits of the environmental gradient are assumed to be equally important. LIVES uses a modified form of the Gower metric as the similarity measure.

Software used: Scripts written by Li and colleagues, and run through R/S-PLUS.

Settings: na

Specifics of data manipulations for modelling: none. All variables used. Categorical variables were turned into binary variables.

Predictions (range, increments): habitat suitability (0 to 1, continuous)

Full name of method: Multivariate Adaptive Regression Splines

Abbreviation: MARS

Alternative names: na

Implementations in this study: single species models (MARS), single species models with one-way interactions allowed (MARS-INT); community models (MARS-COMM)

Key references: Friedman 1991, Hastie and Tibshirani 1996

Examples of implementation in ecology: Moisen and Frescino 2002, Yen et al. 2004, Leathwick et al. 2005

Brief description: MARS is a hybrid between conventional regression and recursive partitioning methods. MARS uses piece-wise linear basis functions to define the modeled relationship. Basis functions are defined in pairs, using a knot to define inflection points, and coefficients to quantify the slopes of the non-zero sections. More than one knot (i.e. more than one pair of basis functions) can be specified for a predictor variable, allowing complex non-linear relationships to be fitted. When fitting a MARS model, knots are chosen in a forward stepwise procedure. Candidate knots can be placed at any position within the range of each predictor variable to define a pair of basis functions. At each step, the model selects the knot and its corresponding pair of basis functions that give the greatest decrease in the residual sum of squares. Knot selection proceeds until some maximum model size is reached, after which a backwards-pruning procedure is applied and those basis

functions that contribute least to model fit are progressively removed. At this stage, a predictor variable can be dropped from the model completely if none of its basis functions contribute meaningfully to predictive performance. The sequence of models generated from this process is then evaluated using generalized cross-validation, and the model with the best predictive fit is selected.

Interactions between variables can be fitted, but rather than fitting a global interaction between a pair of variables, these are specified for only part of the environmental range using basis functions. The R implementation of MARS also allows for the fitting of multiple response variables ("community" models). In this case knots are selected based on their ability to reduce the residual sum of squares, averaged across all species. The final MARS model then uses a common set of basis functions for all species, but individual regressions are used to calculate unique coefficients for each basis function for each species.

The current implementation of MARS in R uses least squares fitting appropriate for data with normally distributed errors. To constrain predicted values within the range 0–1, as appropriate for presence-absence data, we first fitted a MARS model using the standard R code. We then extracted the basis functions from this model and computed a GLM model(s) that related these to the presence/absence of each species. To model presence-only data we used the random background samples in place of "absence" records.

Software used: R, with mda library; extra code written to model binomial responses properly (wrapping basis functions inside a GLM) and to allow modelling of all species in one batch.

Settings: Interactions (where fitted) depth 2; used the functional groups listed in Table 1 for community models. The 10000 background samples ("absence") weighted so total weight for presence = total weight for absence (single species) or total weight for community sites = total weight for absences (community model).

Specifics of data manipulations for modelling: excluded highly correlated and categorical variables

Predictions (range, increments): 0:1, continuous

Full name of method: Maximum entropy modelling

Abbreviation: MAXENT

Alternative names: na

Implementations in this study: MAXENT, MAXENT-T

Key references: Phillips et al. 2006

Examples of implementation in ecology: Phillips et al. 2006

Brief description: Maxent is a general-purpose method for making predictions or inferences from incomplete information. The basic idea is that if we need to estimate an unknown probability distribution, we should find the probability distribution of maximum entropy, subject to the constraints that represent our incomplete information about the unknown distribution. This is known as the maximum-entropy principle (Jaynes 1957).

Entropy is a fundamental concept in information theory: in the paper that originated that field, Shannon (Shannon 1948) described entropy as "a measure of how much 'choice' is involved in the selection of an event". Thus, a distribution with higher entropy involves more choices, i.e. it is less constrained. Therefore, the maximum entropy principle can be interpreted as saying that no unjustified constraints should be placed on our estimate of the unknown distribution.

The information available about the unknown distribution often presents itself as a set of real-valued variables, called "features", and the constraints are that the expected value of each feature should match its empirical average (the average value for a set of sample points taken from the target distribution). When Maxent is applied to presence-only species distribution modelling, the pix-

els of the study area make up the space on which the unknown probability distribution is defined, pixels with known species occurrence records constitute the sample points, and the features are climatic variables, elevation, soil category, vegetation type or other environmental variables, and functions thereof. The unknown probability distribution is proportional to probability of occurrence.

Maxent can also be seen as a maximum-likelihood method. The theory of convex duality can be used to show that if the features are $f_1 \dots f_k$, then the maxent distribution has the form

$$\exp(c_1 f_1(x) + c_2 f_2(x) + \dots + c_k f_k(x)) / Z$$

for some constants c_1, \dots, c_k . Here Z is a normalizing constant, which ensures that the distribution sums to 1. Distributions of this form are called "Gibbs distributions". The maxent distribution is always equal to the Gibbs distribution that maximizes the probability of the sample points. If the constraints are not equalities, but rather that the expected value of each feature is within some error bounds around the empirical average, then the maxent distribution is the Gibbs distribution that minimizes a penalized log loss, i.e., the negative log probability of the sample points plus a penalty term involving the absolute values of the coefficients c_1, \dots, c_k . This is called " L_1 -regularization" or a "lasso".

Software used: MaxEnt, written in Java by Phillips, Schapire and Dudik. It uses L_1 -regularization, with the error bounds depending on the observed standard deviation of each feature. Because entropy is a convex function, it can be efficiently optimized. The MaxEnt software guarantees convergence to the maxent distribution.

Settings: The width of the error bounds has a multiplier that depends on the number and type of features used. The multiplier was tuned on the presence-only data, and the results of the tuning were chosen for the default settings.

Specifics of data manipulations for modelling: na

Predictions (range, increments): Either 0:1 continuous (raw output) or 0:100 continuous (cumulative output). Raw output is proportional to predicted probability of occurrence. For cumulative output, a threshold of x excludes x% of the predicted distribution.

References

- Anderson, R. P., Peterson, A. T. and Gómez-Laverde, M. 2002. Using niche-based GIS modeling to test geographic predictions of competitive exclusion and competitive release in South American pocket mice. – Oikos 98: 3–16.
- Anderson, R. P., Lew, D. and Peterson, A. T. 2003. Evaluating predictive models of species' distributions: criteria for selecting optimal models. – Ecol. Modell. 162: 211–232.
- Austin, M. P., Cunningham, R. B. and Good, R. B. 1983. Altitudinal distribution in relation to other environmental factors of several eucalypt species in southern New South Wales. – Aust. J. Ecol. 8: 169–180.
- Bio, A. M. F., Alkemade, R. and Barendregt, A. 1998. Determining alternative models for vegetation response analysis – a non-parametric approach. – J. Veg. Sci. 9: 5–16.
- Busby, J. R. 1991. BIOCLIM – a bioclimate analysis and prediction system. – In: Margules, C. R. and Austin, M. P. (eds), Nature conservation: cost effective biological surveys and data analysis. CSIRO, pp. 64–68.
- Carpenter, G., Gillison, A. N. and Winter, J. 1993. DOMAIN: a flexible modelling procedure for mapping potential distributions of plants and animals. – Biodiv. Conserv. 2: 667–680.
- Ferrier, S. 2002. Mapping spatial pattern in biodiversity for regional conservation planning: where to from here? – Syst. Biol. 51: 331–363.
- Ferrier, S. et al. 2002. Extended statistical approaches to modelling spatial pattern in biodiversity: the north-east New South Wales experience. II. Community-level modelling. – Biodiv. Conserv. 11: 2309–2338.

- Ferrier, S. et al. 2004. Mapping more of terrestrial biodiversity for global conservation assessment. – Bioscience 54: 1101–1109.
- Friedman, J. H. 1991. Multivariate adaptive regression splines (with discussion). – Ann. Stat. 19: 1–141.
- Friedman, J. H. 2001. Greedy function approximation: a gradient boosting machine. – Ann. Stat. 29: 1189–1232.
- Friedman, J. H. 2002. Stochastic gradient boosting. – Comput. Stat. Data Anal. 38: 367–378.
- Friedman, J. H. and Meulman, J. J. 2003. Multiple additive regression trees with application in epidemiology. – Stat. Med. 22: 1365–1381.
- Friedman, J. H., Hastie, T. and Tibshirani, R. 2000. Additive logistic regression: a statistical view of boosting. – Ann. Stat. 28: 337–407.
- Hastie, T. and Tibshirani, R. 1990. Generalized additive models. – Chapman and Hall.
- Hastie, T. and Tibshirani, R. J. 1996. Discriminant analysis by gaussian mixtures. – J. R. Stat. Soc. Ser. B 58: 155–176.
- Hastie, T., Tibshirani, R. and Friedman, J. H. 2001. The elements of statistical learning: data mining, inference, and prediction. – Springer.
- Hijmans, R. J. et al. 2004. DIVA-GIS, ver. 4. A geographic information system for the analysis of biodiversity data. – Manual, available at <<http://www.diva-gis.org>>.
- Hughes, L., Cawsey, E. M. and Westoby, M. 1996. Climatic range sizes of eucalypt species in relation to future climate change. – Global Ecol. Biogeogr. Lett. 5: 23–29.
- Jaynes, E. T. 1957. Information theory and statistical mechanics. – Phys. Rev. 106: 620–630.
- Kadmon, R., Farber, O. and Danin, A. 2003. A systematic analysis of factors affecting the performance of climatic envelope models. – Ecol. Appl. 13: 853–867.
- Leathwick, J. R. et al. 2005. Using multivariate adaptive regression splines to predict the distributions of New Zealand's freshwater diadromous fish. – Freshwater Biol. 50: 2034–2052.
- Leathwick, J. R. et al. in press. Variation in demersal fish species richness in the oceans surrounding New Zealand: an analysis using boosted regression trees. – Mar. Ecol. Prog. Ser.
- Lehmann, A., Overton, J. M. and Leathwick, J. R. 2003. GRASP: generalized regression analysis and spatial prediction. – Ecol. Modell. 160: 165–183.
- Lindenmayer, D. B. et al. 1991. The conservation of Leadbeater's possum, *Gymnobelideus leadbeateri* (McCoy): a case study of the use of bioclimatic modelling. – J. Biogeogr. 18: 371–383.
- Loiselle, B. A. et al. 2003. Avoiding pitfalls of using species distribution models in conservation planning. – Conserv. Biol. 17: 1591–1600.
- Lowe, D. G. 1995. Similarity metric learning for a variable-kernel classifier. – Neural Comput 7: 72–85.
- McCullagh, P. and Nelder, J. A. 1989. Generalized linear models. – Chapman and Hall.
- Moisen, G. G. and Frescino, T. S. 2002. Comparing five modeling techniques for predicting forest characteristics. – Ecol. Modell. 157: 209–225.
- Peterson, A. T., Pereira, R. S. and Fonseca de Camargo-Neves, V. L. 2004. Using epidemiological survey data to infer geographic distributions of leishmania vector species. – Rev. Soc. Bras. Med. Trop. 37: 10–14.
- Peterson, A. T. et al. 2006. Geographic potential for outbreaks of Marburg hemorrhagic fever. – Am. J. Trop. Med. Hyg., in press.
- Phillips, S. J., Anderson, R. P. and Schapire, R. E. 2006. Maximum entropy modeling of species geographic distributions. – Ecol. Modell. 190: 231–259.
- Schapire, R. 2003. The boosting approach to machine learning – an overview. – In: Denison, D. D. et al. (eds), MSRI Workshop on Nonlinear Estimation and Classification, 2002.
- Shannon, C. E. 1948. A mathematical theory of communication. – The Bell System Technical Journal 27: 379–423 and 623–656.
- Stockwell, D. R. B. and Noble, I. R. 1992. Induction of sets of rules from animal distribution data: a robust and informative method of data analysis. – Math. Comput. Simul. 33: 385–390.
- Stockwell, D. and Peters, D. 1999. The GARP modelling system: problems and solutions to automated spatial prediction. – Int. J. Geogr. Inform. Sci. 13: 143–158.
- Wintle, B. A., Elith, J. and Potts, J. 2005. Fauna habitat modelling and mapping in an urbanising environment; A case study in the Lower Hunter Central Coast region of NSW. – Aust. Ecol. 30: 729–748.
- Yee, T. W. and Mitchell, N. D. 1991. Generalized additive models in plant ecology. – J. Veg. Sci. 2: 587–602.
- Yen, P., Huettmann, F. and Cooke, F. 2004. Modelling abundance and distribution of marbled murrelets (*Brachyramphus marmoratus*) using GIS, marine data and advanced multivariate statistics. – Ecol. Modell. 171: 395–413.