Appendix S1

Flexible species distribution modelling methods perform well on spatially separated testing data. *Global Ecology and Biogeography.*

Roozbeh Valavi, Jane Elith, José Lahoz-Monfort and Gurutzeta Guillera-Arroita

2023-01-06

Contents

1- Covariates and TGB samples 1
2- Code and data availability 5
3- Modelling methods parameters 8
4- Evaluation metrics 10
5- Dispersion of AUC_{ROC} 10
6- Top rank methods in different evaluations 11
7- Rank of the Ensemble vs its components 13
8- Interactions in BRT and MaxEnt 14
9- Extrapolation in testing blocks 15
10- List of species used in modelling 17
11- Statistical tests 22
12- Explanation for poor performance of some flexible methods 24
13- References 25

Appendix S1 for *Valavi, R., Elith, J., Lahos-Monfort, J.J., Guillera-Arroita, G. (2023) Flexible species distribution modelling methods perform well on spatially separated testing data. Global Ecology and Biogeography. DOI: 10.1111/GEB.13639*

1- Covariates and TGB samples

To model the presence-only data with background samples and deal with biases in these data, we used Target-Group-Background (TGB) samples introduced by Phillips *et al.* (2009). TGB provide a background sample with similar biases to that of the presence records. For each species, a background sample is generated by collating all presence records from the same biological group and region (including the target species), with minor adjustments to avoid multiple records per grid cell. For details see Phillips *et al.* (2009) and for code see Elith *et al.* (2020).

Table **S1.1 to S1.6** below show the environmental covariates used in each region. These are the variables that do not have a pairwise correlation of more than 0.8. You can find a complete list of variables with more

details in Elith *et al.* (2020), and in the help of the `disdat` R package.

Table S1.1: Environmental vartiables for AWT region (80m spatial resolution).

Code	Description	Units	Type
bc04	Temperature seasonality	NA	Continuous
bc05	Max. temperature of warmest period	˚C	Continuous
bc06	Min. temperature of coldest period.	˚C	Continuous
bc12	Annual precipitation	mm	Continuous
bc15	Precipitation seasonality	NA	Continuous
slope	Mean slope (derived from 9 second spatial resolution elevation data)	percent	Continuous
topo	Topographic position	NA	Continuous
tri	Terrain ruggedness index	NA	Continuous

Table S1.2: Environmental vartiables for CAN region (1000m spatial resolution).

Code	Description	Units	Type
alt	Digital elevation	m	Continuous
asp2	Aspect – ranges from -1 to 1 (sin transformation)	number (dimensionless unit)	Continuous
ontprec	Annual Precipitation	mm	Continuous
ontslp	Slope	degrees	Continuous
onttemp	Annual mean temperature	˚C * 10	Continuous
ontveg	Vegetation, from Ontario Land Cover Database (OLC) vegetation map, derived from a mosaic of Landsat images.	number (dimensionless unit)	Categorical
watdist	Distance from Hudson Bay	m	Continuous

Table S1.3: Environmental vartiables for NSW region (100m spatial resolution).

Code	Description	Units	Type
cti	compound topographic index- a quantification of the position of a site in the local landscape	number (dimensionless index)	Continous
disturb	disturbance (clearing, logging etc)	number (ordinal)	Continous
mi	moisture index.	number (dimensionless index)	Continous
rainann	mean annual rainfall	Mm	Continous
raindq	mean rainfall of the driest quarter	Mm	Continous
rugged	ruggedness	number (dimensionless index)	Continous
soildepth	mean soil depth	m *1000	Continous
soilfert	soil fertility	number (ordinal)	Continous
solrad	annual mean solar radiation	MJm-2 day-1 * 10	Continous
tempann	annual mean temperature	°C * 10	Continous
topo	topographic position	m	Continous

Table S1.4: Environmental vartiables for NZ region (100m spatial resolution).

Code	Description	Units	Type
age	soil parent material: age since last major rejuvenation	number (category)	Categorical
deficit	mean October vapor pressure deficit at 0900 hours	kPa	Continous
hillshade	surrogate for slope and aspect	number (dimensionless index)	Continous
mas	mean annual solar radiation	MJm^-2 day^-1 * 100	Continous
mat	mean annual temperature	°C * 10	Continous

Table S1.4: Environmental vartiables for NZ region (100m spatial resolution). *(continued)*

Code	Description	Units	Type
r2pet	average monthly ratio of potential evapotranspiration	number (dimensionless index)	Continous
slope	slope	degrees	Continous
sseas	solar radiation seasonality	number (dimensionless index)	Continous
toxicats	toxic cations in soil	number (category)	Categorical
tseas	temperature seasonality	number (dimensionless index)	Continous
vpd	annual vapor pressure deficit	kPa	Continous

Table S1.5: Environmental vartiables for SA region (1000m spatial resolution).

Code	Description	Units	Type
sabio2	mean diurnal range (mean of monthly (max temp - min temp))	˚C*10	Continous
sabio4	temperature seasonality (standard deviation *100)	number (dimensionless index)	Continous
sabio5	max temperature of warmest month	˚C*10	Continous
sabio6	min temperature of coldest month	˚C*10	Continous
sabio12	annual precipitation	mm	Continous
sabio15	precipitation seasonality (coefficient of variation)	number (dimensionless index)	Continous
sabio17	precipitation of driest quarter	mm	Continous
sabio18	precipitation of warmest quarter	mm	Continous

Table S1.6: Environmental vartiables for SWI region (100m spatial resolution).

Code	Description	Units	Type
BCC	broadleaved continuous cover	% cover	Continous
CALC	bedrock strictly calcareous vs other type	1 (present) or 0 (absent)	Categorical
CCC	coniferous continuous cover	% cover	Continous
DDEG	growing degree-days above the threshold of 0°C	°C * days	Continous
NUTRI	soil nutrients index	D mval/cm2	Continous
PDAY	number of days with rainfall > 1 mm	ndays	Continous
PRECYY	average yearly precipitation sum	mm	Continous
SFROYY	summer frost frequency – number of days	days	Continous
SLOPE	slope	degrees x 10	Continous
SRADYY	potential yearly global radiation (daily average)	kJm-2day-1	Continous
SWB	site water balance	mm	Continous
TOPO	topographic position	number (dimensionless index)	Continous

2- Code and data availability

You can find species data in `disdat` R package, with the rasters available on OSF (Elith *et al.,* 2020). To run the models, you can use the codes and data provide in OSF repository.

To reproduce our results, you can use any of the following (details below):

- Using `renv` package recovery in local R environment
  - Use the `revn.lock` file in OSF repository and retrieve the R package versions in your local system with the `renv` package
- Using Docker containers
  - Use the pre-built Docker image (`rvalavi_image.tar` stored in OSF) and create a virtual environment with the same R packages (recommended)
  - Build a Docker image based on the `Dockerfile` provided in the OSF repository.

2.1- Using `renv` package recovery

First, create a new RStudio project and place the `renv.lock` file in it. Use the commands below to restore the R package for modelling.

```
install.packages("renve")

renv::init()

renv::restore()
```

You need to install a few other packages that are not listed in `renv` file.

```
install.packages("remotes")
install.packages("rJava")

remotes::install_github("meeliskull/prg/R_package/prg")
remotes::install_github("b0rxa/scmamp")
remotes::install_github("rvalavi/myspatial")

remotes::install_version("gam", version = "1.20", repos = "http://cran.us.r-project.org")
remotes::install_version("gbm", version = "2.1.5", repos = "http://cran.us.r-project.org")
```

2.2- Using Docker containers

Here we explain how to use Dockers for creating a virtual system to reproduce our results. Docker can be installed on different platforms. You can find the instructions for installing Docker on different operating systems on their website. Docker provides an RStudio installed and all the R and system packages required for running our analysis.

To use Docker, you can either A) load the pre-built image (recommended), or B) build a new image from `Dockerfile`.

`A)` To load the Docker image, first download the files from OSF, then use:

```
docker load --input rvalavi_image.tar
```

`B)` If you want to build the image in your local system, you need to download all the files in OSF (except "docker" folder). Then run the following terminal commands. In Linux systems you might need to use `sudo` before `docker` commands.

Go to the directory of downloaded files from the OSF within terminal and run:

```
docker build -t rvalavi:4.0 .
```

Wait until the build is complete. Then check to see the images is created.

```
docker images
```

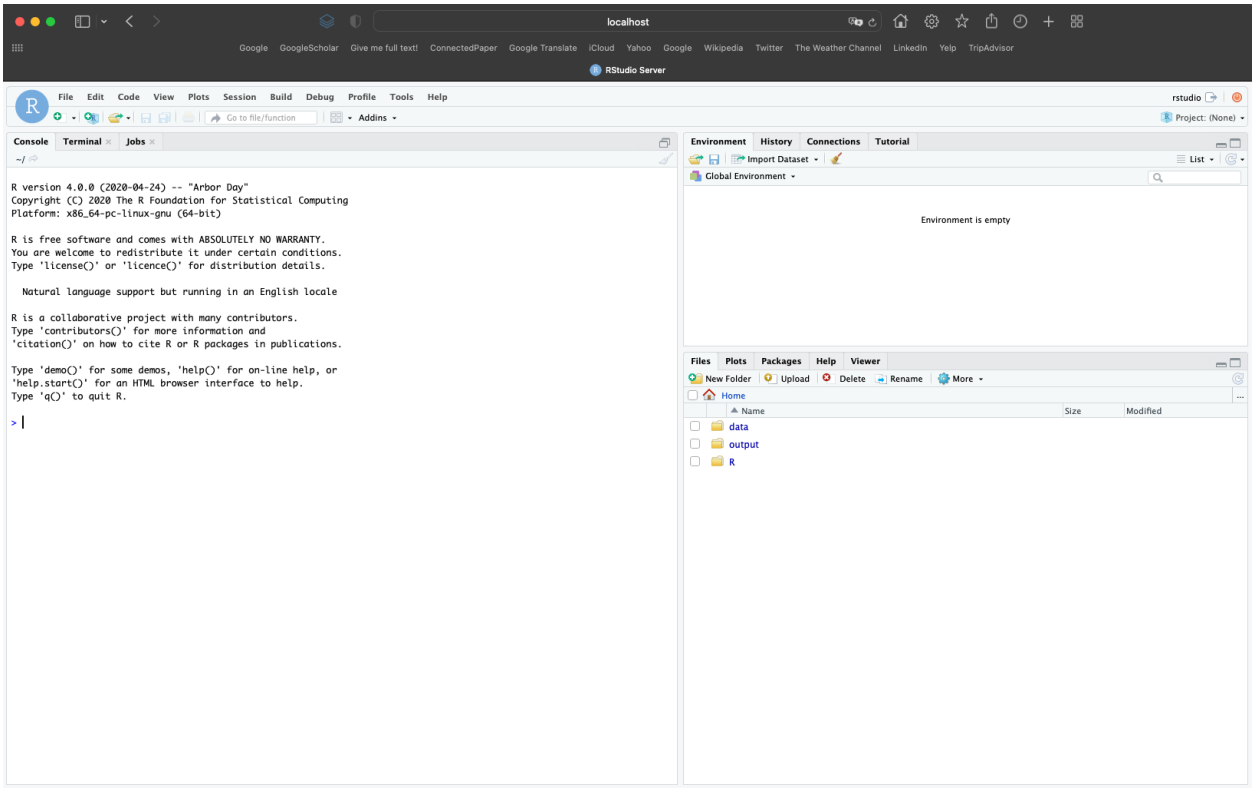
You should see `rvalavi` with TAG 4.0 listed as a Docker image.

After the image is loaded (A) or created (B), you need to run a container to get access to RStudio and the R packages. The Docker container is a live instance of the image. Use the following command to run a container to access RStudio.

```
docker run --name rstudio -p 8787:8787 -e PASSWORD=123 -d rvalavi:4.0
```

This code has several components:
`--name`: name of the container
`-p`: port on which container is running. We use this to connect to rstudio
`-e PASSWORD`: password for the rstudio server (you can choose any password)
`-v`: mapping a directory in the local system to a directory in the container (this was not used in the code above). This will allow you to save the generated files and code in local drive and also access to code/data inside your system.
`-d`: run the container in the background
`rvalavi:4.0`: name and tag of the Docker image

Now, run RStudio server from container. Open an Internet browser and go to `localhost:8787` to open RStudio. Use "**rstudio**" as username and the password you specified in the previous step (here password is "123") to open RStudio.



You can run the models by running the `R/nceas_modelling.R` script. The model predictions will be stored in `output/nceas_model_output` folder. To calculate evaluations, run `R/nceas_evaluation.R` script.

3- Modelling methods parameters

A summary of model implementation settings are presented in Table S1.7, below. The "parameters" column shows model arguments in R programming that are selected in the modelling process. The "values" column shows the value or ranges of values selected for model fitting and tuning for each R function. All models are fitted in R `v4.0.0`.

Some methods accept weights. GAM, GLMs, and BRT use case weights (i.e., there is a weight for each training sample), and SVM utilizes class weights (i.e., there is a weight for each class; here presence and background samples are the two classes so there are only two weights). The weights are generated by giving a weight of 1 to every presence point and giving the weights to the background in a way that the sum of the weights for the presence and background are equal. For class weights in SVM, an inverse proportional weight was used.

Table S1.7: Parameters used for implementing different modes.

Method	Parameter	Values	Description	R.packages
GAM	method	REML	smoothing parameter estimation method	mgcv v1.8-32
	k	10	the number of basis functions (for creating smoothing terms) specifies the possible maximum effective degree of freedom	
GLM-step *	direction	both	step-selection direction (forward & backward) based on AIC	gam v1.20
GLM-lasso *	alpha	1	lasso penalty	glmnet v4.0-2
MARS	nprune	2 to 20	number of terms	earth v5.1.2
	degree	1	degree of interaction (1 means no interaction allowed)	
MaxEnt (default)	args	nothreshold	auto select feature and exclude threshold feature	dismo v1.1-4 and maxent.jar v3.4.4
	betamultiplier	1	regularisation multiplier	
BRT	tree.complexity	1 or 5	depth of individual trees - two options depending on sample size	dismo v1.1-4 and gbm v2.1.5
	learning.rate	0.001	shrinkage or the weight applied to individual trees	
	bag.fraction	0.75	proportion of observations sampled to train each tree	
	n.folds	5	number of cross-validation folds	
RF-shallow	mtry	sqrt(p)	number of variables randomly selected at each split	ranger v0.12.1
	num.trees	2000	number of trees	

Table S1.7: Parameters used for implementing different modes. *(continued)*

Method	Parameter	Values	Description	R.packages
	splitrule	"hellinger"	tree splitting criterion	
	max.depth	2	maximum depth of each tree (forcing shallow trees)	
	probability	TRUE	fitting probability trees	
RF down-sampled	mtry	sqrt(p)	number of variables randomly selected at each split	randomForest v4.6-14
	sampsize	n. presences	number of bootstrap samples taken from each class	
	ntrees	1000	number of trees	
SVM	kernel	radial	radial basis kernel	e1071 v1.7-3
Ensemble			Rescale and average of individual modes implemented here: GAM, GLM-lasso, MaxEnt, BRT and RF down-sampled	

Note:
* GLM-step and GLM-lasso were fitted allowing linear and quadratic terms only, with no interactions.

Parameters of MaxEnt variants are presented in the following table.

Table S1.8: Parameters of MaxEnt model in different MaxEnt variants.

Method	Parameter	Values	Description
MaxEnt (default)	betamultiplier	1	regularization multiplier
	args	nothreshold	auto select feature and exclude threshold feature
MaxEnt noclamp			the same as MaxEnt (default) with clammping set to off for prediction
MaxEnt LQ	betamultiplier	1	regularization multiplier
	feature types	LQ	transformations of input covariates. L: linear, Q: quadratic
MaxEnt tuned and spatial-tuned	betamultiplier	0.5, 1, 2, 3, 4	regularization multiplier

Table S1.8: Parameters of MaxEnt model in different MaxEnt variants. *(continued)*

Method	Parameter	Values	Description
	feature types	L, LQ, H, LQH, LQHP	transformations of input covariates. L: linear, Q: quadratic, H: hinge, P: product

4- Evaluation metrics

We used AUC_{ROC}, AUC_{PRG} and COR for evaluating the models. AUC_{ROC} measures how well a model discriminates between presence and absence records in the test dataset. It can range from 0 to 1, with 1 indicating a model has perfect discrimination abilities and 0.5 showing discrimination is equivalent to that from random predictions (Pearce & Ferrier, 2000; Elith *et al.*, 2006). AUC_{PRG} is similar to AUC_{ROC}, but less commonly used in ecology. It puts more focus on correctly predicted presences (Flach & Kull, 2015). An AUC_{PRG} value of 1 shows perfect discrimination, 0 indicates random discrimination and negative denotes worse than random. Since there is no lower limit for negative values in AUC_{PRG}, we only estimated ranks, not mean performance, for this metric. COR is the correlation between model predictions and the presence-absence testing data (Elith *et al.*, 2006).

5- Dispersion of AUC_{ROC}

To further highlight the difference between the performance of models (dispersion of validation metrics), we calculated, for each species, the difference between the AUC_{ROC} of each method and the average AUC_{ROC} of all modelling methods for that species (Fig S1.1). Values higher than zero indicate AUC_{ROC} higher than average.

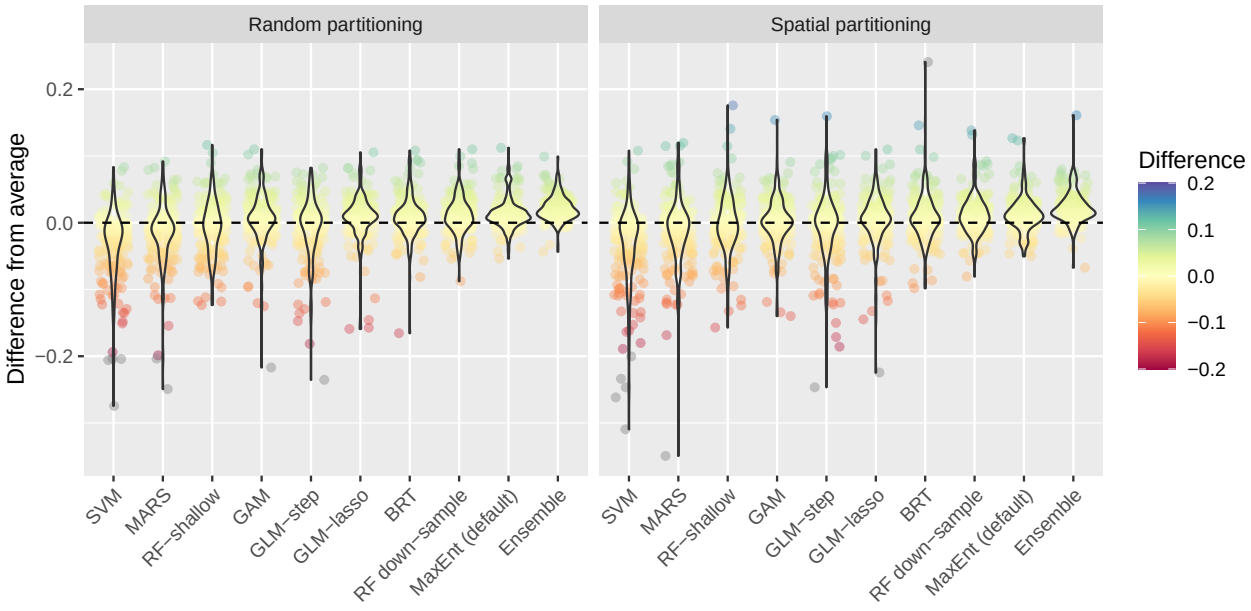


Fig. S1.1: Difference from average AUC_{ROC}.

6- Top rank methods in different evaluations

In the main text, the mean performance and the average rank of performance metrics (i.e., AUC_{ROC}, AUC_{PRG}, and COR) were used to assess and compare models. However, this approach does not show how frequently a method was the top method or whether it was among the top 2 or 3 methods. Here we calculated the percentage of species (171 species in our study) for which a method was in the top 1, top 2, or top 3 methods (Fig S1.2 and S1.3). For example, for AUC_{ROC} and random partitioning (Fig S1.2), the Ensemble was within the top 3 methods for 63.7% of the species.

Notice that all methods performed the best (top 1) for at least a few species in both random and spatial partitioning. A noticeable result here is that while some methods like GLM-step, MARS, or RF-shallow are average performers overall, they may be top methods more frequently than better average performers (for example, GLM-step vs BRT in the top 1 methods for random partitioning; or RF-shallow vs RF down-sample or MaxEnt in top 1 method for spatial partitioning).

Another highlight is that although the Ensemble was not the top performer for many species, it was among the top 3 methods for more than half of them in both partitioning strategies (Fig S1.2 and S1.3).

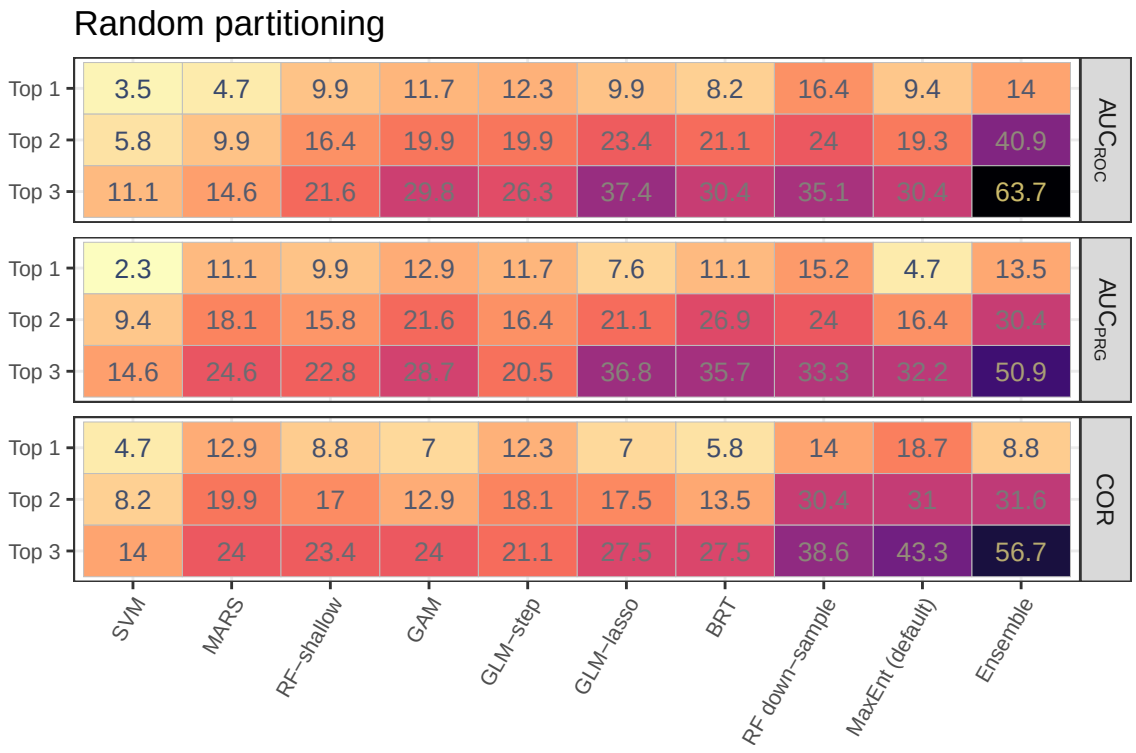


Fig. S1.2: Top rank methods in random partitioning.

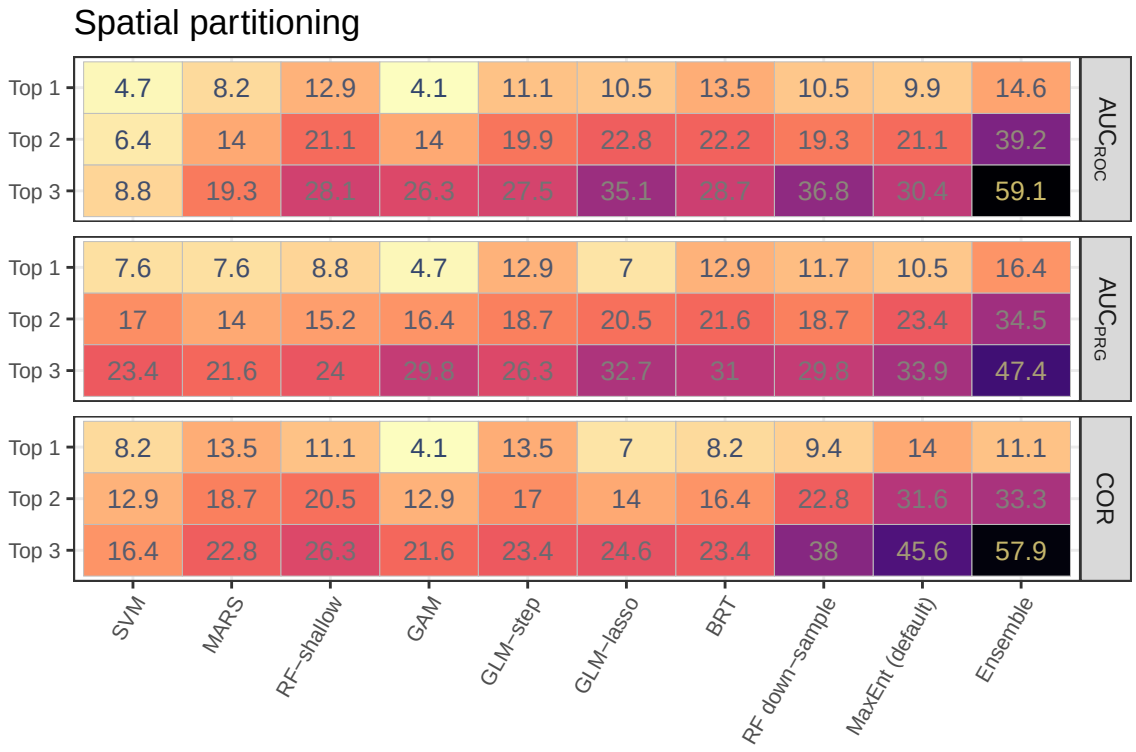


Fig. S1.3: Top rank methods in spatial partitioning

An interesting result is that although Ensemble is the most frequent top performer in terms of AUC_{ROC} and AUC_{PRG} when predicting spatially separated testing data, and second best in terms of COR in spatial partitioning. Under random partitioning, Ensemble was the second or third best average performer. We explored further how Ensemble is performing compared to its component models in the following section.

7- Rank of the Ensemble vs its components

Here we calculated the same plots but only for the Ensemble and its component models i.e, GLM-lasso, GAM, MaxEnt (default), BRT, and RF down-sample (Fig S1.4). The Ensemble appeared in the top 2 and 3 models for more species than its component in both random and spatial partitioning. For top 1 methods, it was only best for AUC_{ROC} (along with RF down-sample) in random partitioning, but the best for both AUCs and the second-best for COR in spatial partitioning. The fact that Ensemble appears better than its component in spatial partitioning may be evidence that ensembling of tuned models can lead to better generalisation.

Fig. S1.4: Top rank methods among Ensemble and its components.

To assess whether the Ensemble improves with respect to the best method in the set or not, we further explored (Fig S1.5) and realised that in all cases of "Top 1", Ensemble actually somewhat outperformed its component (rather than being just as good as the best of its components). This could be an indication that the ensemble gains by combining "complementary" predictions.

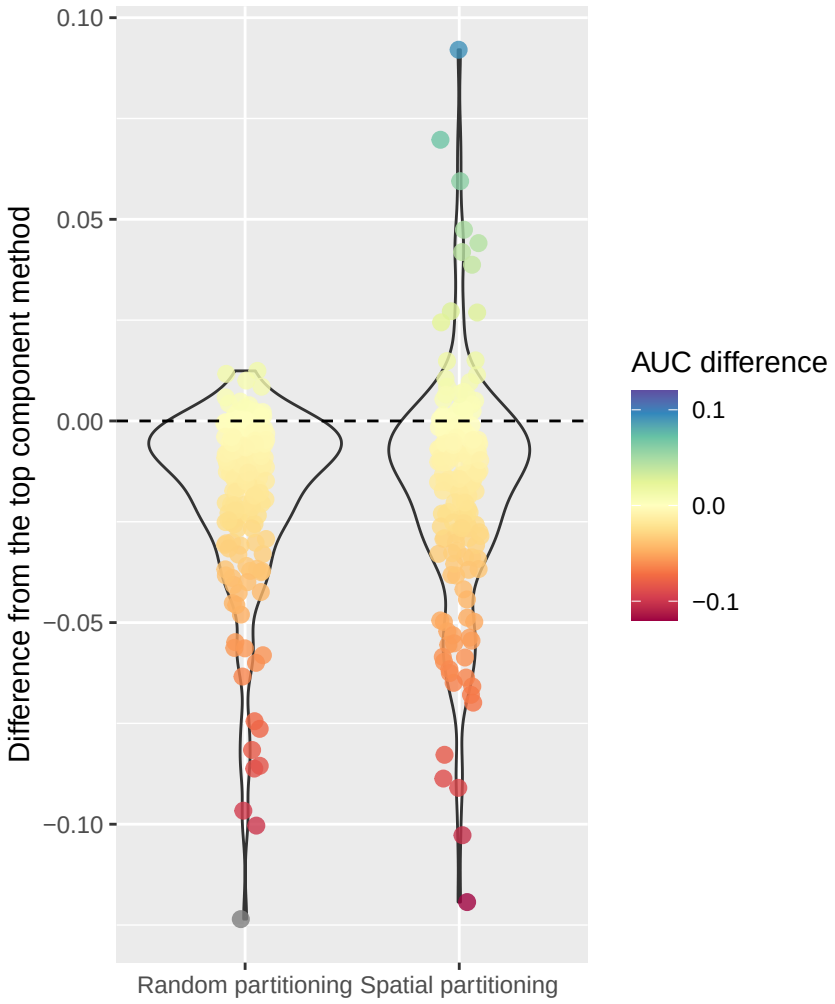


Fig. S1.5: The differnce between AUC_{ROC} of Ensemble and AUC_{ROC} of the best component method.

8- Interactions in BRT and MaxEnt

To assess the impacts of interactions in flexible methods, we compared two of our top methods, BRT and MaxEnt, with and without interactions (Fig S1.6). We implemented a BRT model with a set tree-complexity of `1` also known as **stump**. The main difference between the BRT-stump and the default BRT (with tree-complexity 1 or 5) is that the default BRT is allowed to fit a higher level of interaction between the covariates if there are more than 50 species records in the training data (57% of cases). The implemented BRT method (Elith *et al.* 2008) utilized internal cross-validation to find the best number of trees for the model. Thus, by limiting tree-complexity to 1, the model adds more trees to find a similar balance in the fitted model as the BRT with tree-complexity 5.

MaxEnt is also presented as two variants here, the MaxEnt with enforced LQ features and one with enforced LQP features. The main difference between these two is that MaxEnt-LQP accommodates interaction as the product of the linear features.

The main BRT was modelled with a tree-complexity of 1 (stump) for 43% of the times.

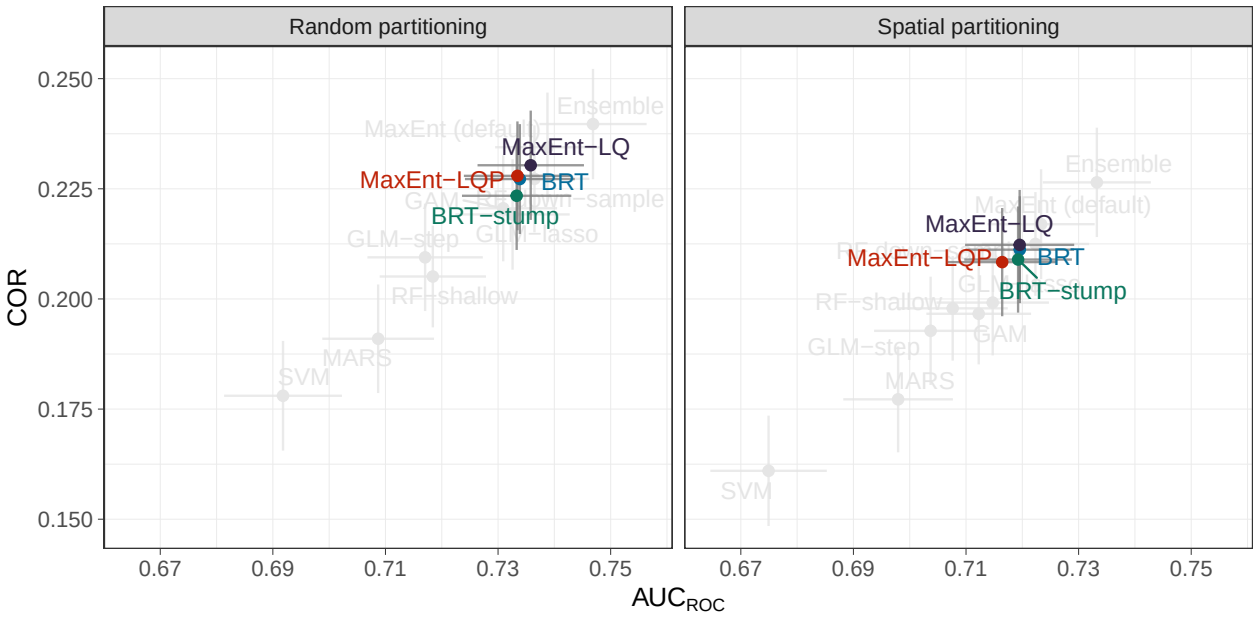


Fig. S1.6: Perfromance of implementation of BRT and MaxEnt with forced limited flexibility.

There is a small and non significant difference between the average performance of the BRT vs BRT-stump and between MaxEnt-LQP vs MaxEnt-LQ. The MaxEnt-LQP had a lower mean AUC_{ROC} and COR in both partitioning methods which could be because enforced interaction in the model is too complex for some species with a very low number of presence records. On the other hand, BRT performed better than BRT-stump, implying that the additional flexibility of interactions is beneficial.

9- Extrapolation in testing blocks

It is useful to know whether extrapolation occurs when models are used to predict to spatially separated points. Extrapolation occurs when the testing/predicting sites have environmental values outside of the range of environmental conditions used in the training samples. To measure the amount of extrapolation in testing sites we used Multivariate Environmental Similarity Surface (**MESS**) introduced by Elith *et al.* (2010). We modified the `mess` function in the **dismo** R package to compute MESS values for points, not rasters. We estimated MESS values for the records in the presence-absence evaluation dataset, using the training presence-TGB as the reference sites. We used only continuous covariates for this. For more explanation on MESS, see Elith *et al.* (2010).

In Fig S1.7, we summed the number of testing points with extrapolation (negative MESS values) for each species. This gives a good sense of how frequently species from a region experience extrapolation when predicting.

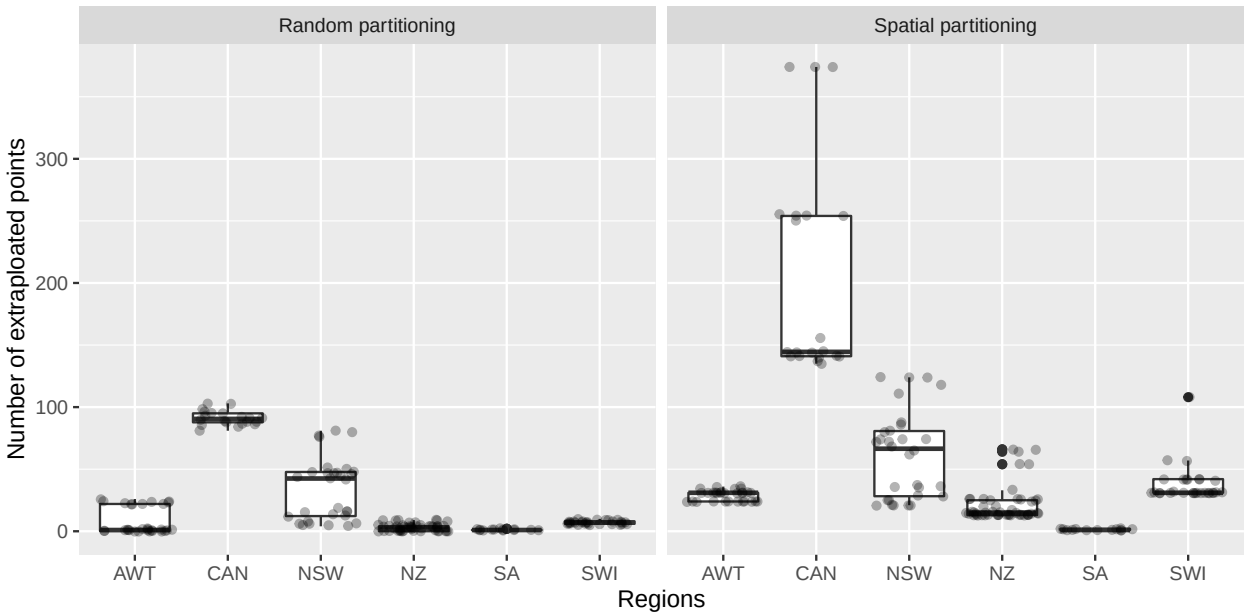


Fig. S1.7: The sum of extrapolated points of each species in each region.

Fig S1.8 shows the number of extrapolated sites when using spatial partitioning compared to random partitioning.

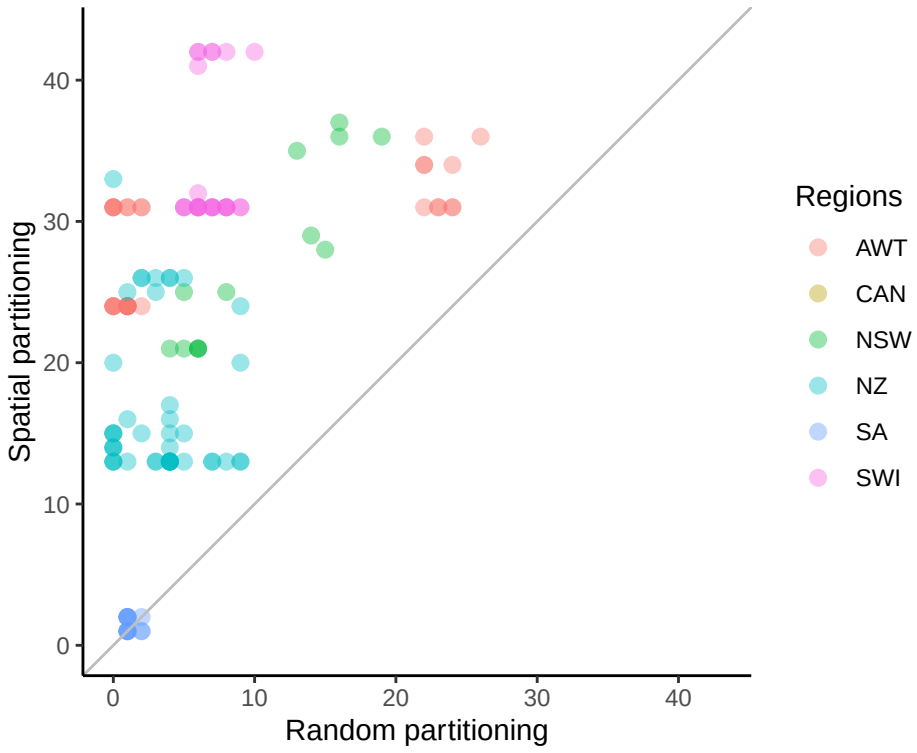


Fig. S1.8: The sum of the number of extrapolated points in each species in random vs spatial partitioning.

10- List of species used in modelling

For creating spatial blocks some species did not have enough data to fit and evaluate models. Here, we provide the list of species we used in our study (Table S1.9). You can see the location of each region on the world map in Fig S1.9. Read detailed explanation of this dataset in Elith *et al* (2020).

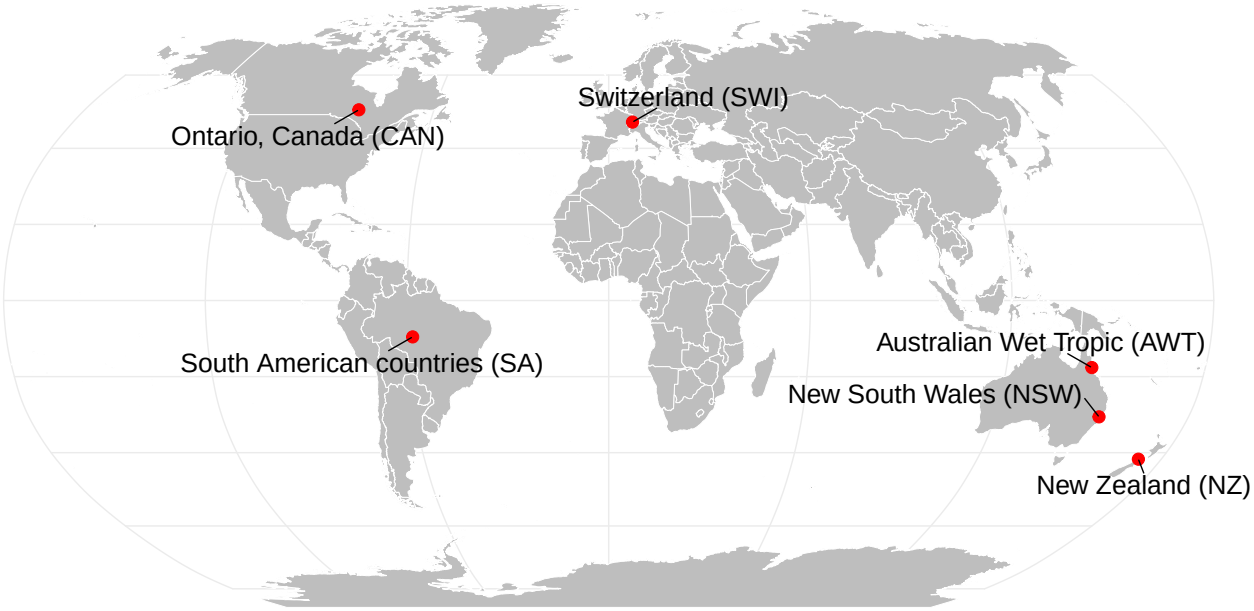


Fig. S1.9: Location of each region in the world.

Table S1.9: List of species used for modelling. PO is the number of presence-only recods in the tarining dataset, TGBs is the number of Target-Group-Background samples, and Presence/Absence are the number of records in the evaluation dataset.

	Region	Species	Group	PO	TGBs	Presence	Absence
1	AWT	**awt01**	Birds	178	714	92	248
2	AWT	**awt02**	Birds	149	714	105	235
3	AWT	**awt03**	Birds	219	714	129	211
4	AWT	**awt04**	Birds	216	714	158	182
5	AWT	**awt05**	Birds	114	714	47	293
6	AWT	**awt06**	Birds	265	714	187	153
7	AWT	**awt07**	Birds	158	714	81	259
8	AWT	**awt08**	Birds	231	714	137	203
9	AWT	**awt09**	Birds	112	714	46	294
10	AWT	**awt10**	Birds	122	714	102	238
11	AWT	**awt11**	Birds	82	714	13	327
12	AWT	**awt12**	Birds	32	714	16	324
13	AWT	**awt13**	Birds	122	714	102	238

Table S1.9: List of species used for modelling. PO is the number of presence-only recods in the tarining dataset, TGBs is the number of Target-Group-Background samples, and Presence/Absence are the number of records in the evaluation dataset. *(continued)*

	Region	Species	Group	PO	TGBs	Presence	Absence
14	AWT	**awt14**	Birds	99	714	65	275
15	AWT	**awt15**	Birds	254	714	214	126
16	AWT	**awt16**	Birds	144	714	64	276
17	AWT	**awt17**	Birds	184	714	103	237
18	AWT	**awt18**	Birds	242	714	137	203
19	AWT	**awt19**	Birds	78	714	64	276
20	AWT	**awt20**	Birds	104	714	80	260
21	AWT	**awt21**	Plants	17	379	20	82
22	AWT	**awt25**	Plants	12	379	24	78
23	AWT	**awt27**	Plants	41	379	27	75
24	AWT	**awt31**	Plants	14	379	62	40
25	AWT	**awt33**	Plants	18	379	24	78
26	AWT	**awt35**	Plants	44	379	52	50
27	AWT	**awt36**	Plants	56	379	51	51
28	AWT	**awt37**	Plants	28	379	21	81
29	AWT	**awt38**	Plants	42	379	16	86
30	AWT	**awt39**	Plants	20	379	32	70
31	CAN	**can01**	Birds	16	3298	230	14341
32	CAN	**can02**	Birds	740	3298	2682	11889
33	CAN	**can03**	Birds	165	3298	703	13868
34	CAN	**can04**	Birds	42	3298	887	13684
35	CAN	**can05**	Birds	138	3298	846	13725
36	CAN	**can06**	Birds	27	3298	53	14518
37	CAN	**can07**	Birds	221	3298	333	14238
38	CAN	**can08**	Birds	322	3298	2600	11971
39	CAN	**can09**	Birds	119	3298	163	14408
40	CAN	**can10**	Birds	234	3298	1251	13320
41	CAN	**can11**	Birds	478	3298	4512	10059
42	CAN	**can12**	Birds	312	3298	2045	12526
43	CAN	**can13**	Birds	39	3298	735	13836
44	CAN	**can14**	Birds	18	3298	400	14171
45	CAN	**can15**	Birds	721	3298	3025	11546
46	CAN	**can16**	Birds	57	3298	316	14255
47	CAN	**can17**	Birds	313	3298	2758	11813

Table S1.9: List of species used for modelling. PO is the number of presence-only recods in the tarining dataset, TGBs is the number of Target-Group-Background samples, and Presence/Absence are the number of records in the evaluation dataset. *(continued)*

	Region	Species	Group	PO	TGBs	Presence	Absence
48	CAN	**can18**	Birds	612	3298	1184	13387
49	CAN	**can19**	Birds	109	3298	24	14547
50	CAN	**can20**	Birds	380	3298	893	13678
51	NSW	**nsw01**	Bats	26	147	120	450
52	NSW	**nsw03**	Bats	12	147	21	549
53	NSW	**nsw05**	Bats	22	147	40	530
54	NSW	**nsw07**	Bats	28	147	44	526
55	NSW	**nsw08**	Diurnal birds	129	1351	20	682
56	NSW	**nsw09**	Diurnal birds	426	1351	77	625
57	NSW	**nsw11**	Diurnal birds	48	1351	19	683
58	NSW	**nsw12**	Diurnal birds	139	1351	27	675
59	NSW	**nsw13**	Diurnal birds	155	1351	25	677
60	NSW	**nsw14**	Diurnal birds	315	1351	161	541
61	NSW	**nsw15**	Diurnal birds	236	1351	123	579
62	NSW	**nsw16**	Nocturnal birds	120	1351	143	994
63	NSW	**nsw17**	Nocturnal birds	148	1351	141	996
64	NSW	**nsw18**	Open-forest trees	69	569	440	1635
65	NSW	**nsw23**	Open-forest trees	60	569	480	1595
66	NSW	**nsw24**	Open-forest trees	68	569	184	1891
67	NSW	**nsw25**	Open-forest trees	23	569	87	1988
68	NSW	**nsw27**	Open-forest understorey vascular plants	23	569	445	864
69	NSW	**nsw28**	Open-forest understorey vascular plants	53	569	512	797
70	NSW	**nsw31**	Open-forest understorey vascular plants	24	569	653	656
71	NSW	**nsw32**	Open-forest understorey vascular plants	28	569	22	1287
72	NSW	**nsw33**	Open-forest understorey vascular plants	15	569	513	796
73	NSW	**nsw39**	Rainforest trees	16	569	414	622
74	NSW	**nsw40**	Rainforest trees	14	569	359	677
75	NSW	**nsw43**	Rainforest understorey vascular plants	42	569	216	693
76	NSW	**nsw45**	Rainforest understorey vascular plants	22	569	138	771

Table S1.9: List of species used for modelling. PO is the number of presence-only recods in the tarining dataset, TGBs is the number of Target-Group-Background samples, and Presence/Absence are the number of records in the evaluation dataset. *(continued)*

	Region	Species	Group	PO	TGBs	Presence	Absence
77	NSW	**nsw49**	Small reptiles	110	530	169	839
78	NSW	**nsw52**	Small reptiles	186	530	165	843
79	NSW	**nsw53**	Small reptiles	34	530	27	981
80	NSW	**nsw54**	Small reptiles	118	530	13	995
81	NZ	**nz01**	Plants	23	2503	66	19054
82	NZ	**nz02**	Plants	80	2503	5130	13990
83	NZ	**nz03**	Plants	32	2503	217	18903
84	NZ	**nz04**	Plants	48	2503	1776	17344
85	NZ	**nz05**	Plants	211	2503	3233	15887
86	NZ	**nz06**	Plants	35	2503	1374	17746
87	NZ	**nz07**	Plants	65	2503	42	19078
88	NZ	**nz08**	Plants	124	2503	1767	17353
89	NZ	**nz09**	Plants	36	2503	216	18904
90	NZ	**nz10**	Plants	27	2503	1032	18088
91	NZ	**nz11**	Plants	21	2503	741	18379
92	NZ	**nz12**	Plants	25	2503	3727	15393
93	NZ	**nz13**	Plants	21	2503	2745	16375
94	NZ	**nz14**	Plants	19	2503	6458	12662
95	NZ	**nz17**	Plants	94	2503	5537	13583
96	NZ	**nz18**	Plants	43	2503	510	18610
97	NZ	**nz19**	Plants	127	2503	800	18320
98	NZ	**nz20**	Plants	33	2503	477	18643
99	NZ	**nz21**	Plants	21	2503	842	18278
100	NZ	**nz22**	Plants	130	2503	689	18431
101	NZ	**nz23**	Plants	23	2503	3119	16001
102	NZ	**nz24**	Plants	22	2503	60	19060
103	NZ	**nz25**	Plants	113	2503	489	18631
104	NZ	**nz26**	Plants	22	2503	238	18882
105	NZ	**nz27**	Plants	40	2503	1102	18018
106	NZ	**nz28**	Plants	18	2503	69	19051
107	NZ	**nz29**	Plants	21	2503	3382	15738
108	NZ	**nz30**	Plants	101	2503	7490	11630
109	NZ	**nz32**	Plants	105	2503	1119	18001
110	NZ	**nz33**	Plants	19	2503	781	18339

Table S1.9: List of species used for modelling. PO is the number of presence-only recods in the tarining dataset, TGBs is the number of Target-Group-Background samples, and Presence/Absence are the number of records in the evaluation dataset. *(continued)*

	Region	Species	Group	PO	TGBs	Presence	Absence
111	NZ	**nz34**	Plants	42	2503	534	18586
112	NZ	**nz35**	Plants	24	2503	10581	8539
113	NZ	**nz36**	Plants	170	2503	6848	12272
114	NZ	**nz37**	Plants	22	2503	2496	16624
115	NZ	**nz38**	Plants	147	2503	1351	17769
116	NZ	**nz39**	Plants	21	2503	33	19087
117	NZ	**nz40**	Plants	19	2503	779	18341
118	NZ	**nz41**	Plants	20	2503	40	19080
119	NZ	**nz42**	Plants	27	2503	5845	13275
120	NZ	**nz43**	Plants	137	2503	47	19073
121	NZ	**nz44**	Plants	65	2503	2790	16330
122	NZ	**nz45**	Plants	26	2503	301	18819
123	NZ	**nz46**	Plants	36	2503	22	19098
124	NZ	**nz47**	Plants	87	2503	2536	16584
125	NZ	**nz48**	Plants	43	2503	889	18231
126	NZ	**nz49**	Plants	37	2503	125	18995
127	NZ	**nz50**	Plants	131	2503	959	18161
128	NZ	**nz51**	Plants	42	2503	562	18558
129	NZ	**nz52**	Plants	174	2503	555	18565
130	SA	**sa01**	Plants	120	1221	15	137
131	SA	**sa02**	Plants	150	1221	23	129
132	SA	**sa09**	Plants	49	1221	10	142
133	SA	**sa10**	Plants	99	1221	29	123
134	SA	**sa12**	Plants	203	1221	15	137
135	SA	**sa15**	Plants	88	1221	15	137
136	SA	**sa17**	Plants	37	1221	11	141
137	SA	**sa18**	Plants	123	1221	19	133
138	SA	**sa20**	Plants	54	1221	10	142
139	SA	**sa21**	Plants	27	1221	13	139
140	SA	**sa22**	Plants	57	1221	11	141
141	SA	**sa24**	Plants	138	1221	21	131
142	SA	**sa26**	Plants	216	1221	27	125
143	SWI	**swi01**	Trees	482	11429	107	9906
144	SWI	**swi02**	Trees	1245	11429	298	9715

Table S1.9: List of species used for modelling. PO is the number of presence-only recods in the tarining dataset, TGBs is the number of Target-Group-Background samples, and Presence/Absence are the number of records in the evaluation dataset. *(continued)*

	Region	Species	Group	PO	TGBs	Presence	Absence
145	SWI	**swi03**	Trees	291	11429	142	9871
146	SWI	**swi04**	Trees	710	11429	119	9894
147	SWI	**swi06**	Trees	5822	11429	6953	3060
148	SWI	**swi07**	Trees	857	11429	222	9791
149	SWI	**swi08**	Trees	1452	11429	477	9536
150	SWI	**swi09**	Trees	937	11429	306	9707
151	SWI	**swi10**	Trees	2830	11429	1366	8647
152	SWI	**swi11**	Trees	749	11429	104	9909
153	SWI	**swi12**	Trees	37	11429	20	9993
154	SWI	**swi13**	Trees	3357	11429	3326	6687
155	SWI	**swi14**	Trees	2142	11429	978	9035
156	SWI	**swi15**	Trees	297	11429	134	9879
157	SWI	**swi16**	Trees	734	11429	395	9618
158	SWI	**swi17**	Trees	458	11429	308	9705
159	SWI	**swi18**	Trees	382	11429	26	9987
160	SWI	**swi19**	Trees	36	11429	19	9994
161	SWI	**swi20**	Trees	613	11429	271	9742
162	SWI	**swi21**	Trees	426	11429	224	9789
163	SWI	**swi22**	Trees	560	11429	182	9831
164	SWI	**swi23**	Trees	986	11429	1493	8520
165	SWI	**swi24**	Trees	293	11429	278	9735
166	SWI	**swi25**	Trees	279	11429	238	9775
167	SWI	**swi26**	Trees	89	11429	22	9991
168	SWI	**swi27**	Trees	468	11429	391	9622
169	SWI	**swi28**	Trees	5528	11429	4246	5767
170	SWI	**swi29**	Trees	154	11429	100	9913
171	SWI	**swi30**	Trees	2800	11429	1520	8493

11- Statistical tests

11.1- Statistical test for random partitioning

Here the statistical test on the differences between methods in random partitioning are presented (Fig S1.10). The plots are AUC_{ROC}, AUC_{PRG}, and COR from top to bottom, respectively. The number on the top of

the x-axis shows the range of the ranks of the models. The average rank of each model is indicated by the thin line connected to the axis. The lines (methods) that are connected by the horizontal thick line are not statistically different at 0.05 significance level.

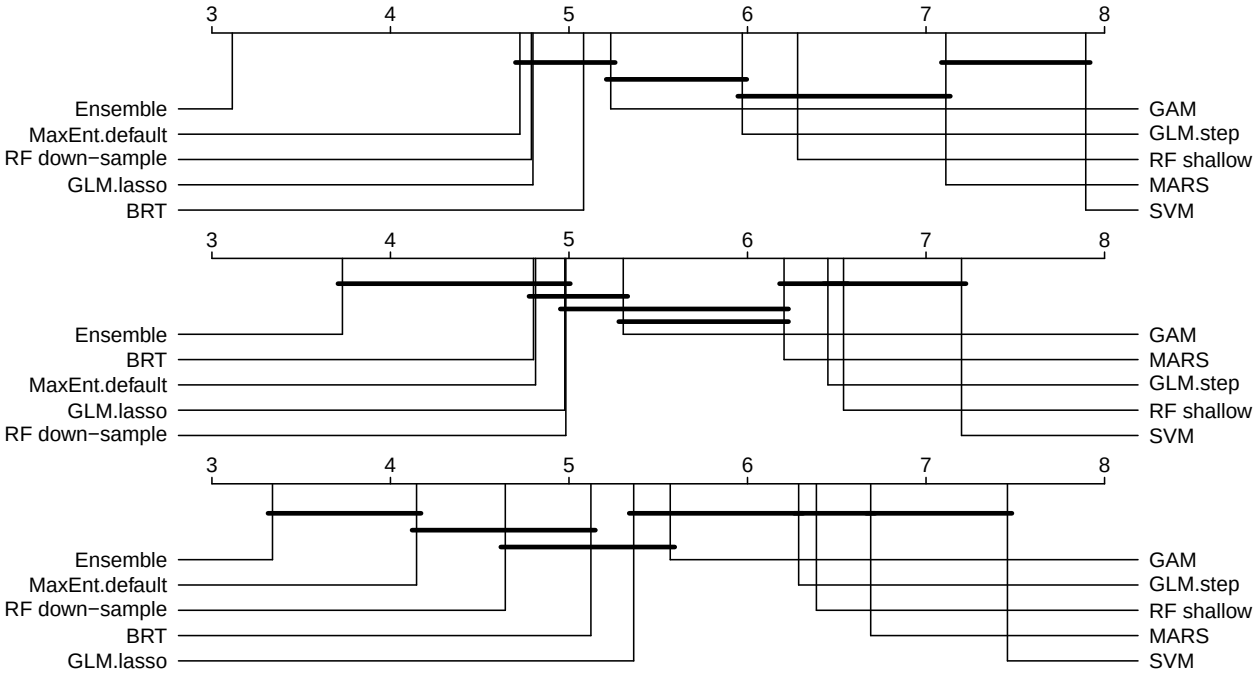


Fig. S1.10: Average rank and statistical difference of the methods in random partitioning.

11.2- Statistical test for MaxEnt variants in spatial partitioning

The Friedman's Aligned Rank test indicated no significant difference among MaxEnt variants for AUC_{PRG}. Thus here we only plot the result of pairwise test for AUC_{ROC} and COR (Fig S1.11). In general the differences between models are also insignificant for these statistics.

```
##
##  Friedman's Aligned Rank Test for Multiple Comparisons
##
## data:  roc_maxents
## T = 13.704, df = 4, p-value = 0.008303

##
##  Friedman's Aligned Rank Test for Multiple Comparisons
##
## data:  prg_maxents
## T = 7.7372, df = 4, p-value = 0.1017

##
##  Friedman's Aligned Rank Test for Multiple Comparisons
##
## data:  cor_maxents
```

```
## T = 16.766, df = 4, p-value = 0.002146
```

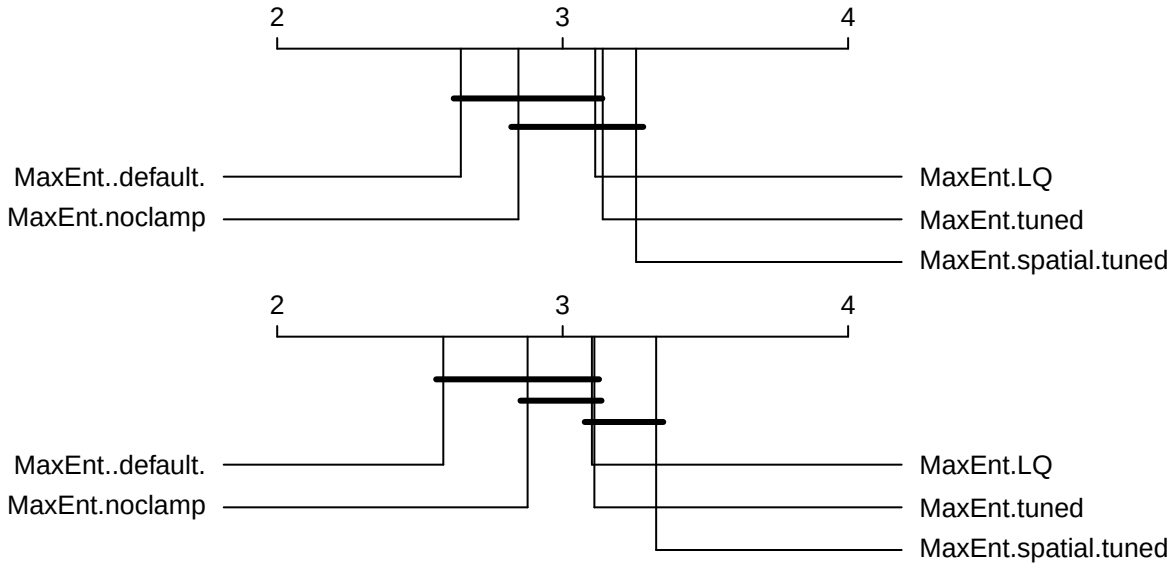


Fig. S1.11: Average rank and statistical difference of the MaxEnt variants in spatial partitioning.

12- Explanation for poor performance of some flexible methods

One result of interest is that some flexible methods (SVM, MARS and RF-shallow) did not perform particularly well (Fig. 3-5 in the main text). So why do these perform less well? It could be that they are not flexible enough: they have no or very limited interactions (Table 1 in the main text). To explore this idea, we did explore how much interactions affect results for BRT and MaxEnt (Section 8). Impacts are small and variable – for BRT, allowing interactions for species with `> 50` data points improves performance slightly, and for MaxEnt, enforcing product features decreases it, probably because enforced feature classes perform less well on species with fewer records. Other issues may be at play for SVM, MARS and RF-shallow. It is possible that our implementation of them was not optimal, since we have more experience with the methods that did well than with some of these. SVM was fitted with defaults. Tuning options are available, so it may be worth testing this further. SVM did not perform as well in the "top 3" analysis (see Section 6) as it did in Valavi *et al.* (2022), where it was fitted with the same settings but on presence-random background data (rather than presence-TGB), and using more species and more data per species (because the current test-train experiment requires subsampling). These nuances could be further explored. MARS was fitted with the `earth` package in R. In Elith *et al.* (2006) MARS, fitted there using the `mda` R package, performed relatively better than in Valavi *et al.* (2022) and this study. We used `earth` because `mda` failed to fit MARS with many background samples for many species. Perhaps different implementations of this model would perform better. Finally, RF-shallow performed better on spatial partitioning than random partitioning (Figs 3-5 in the main text, and Section 6 here), and may benefit from some tuning of tree depth parameter (e.g., Valavi *et al.* 2021). This is an experimental approach for RF, and still needs to be explored further.

13- References

- Elith, J., Graham, C.H., Anderson, R.P., Dudík, M., Ferrier, S., Guisan, A., J Hijmans, R., Huettmann, F., R Leathwick, J., Lehmann, A., Li, J., G Lohmann, L., Overton, J.McC.M., Peterson, A.T., Phillips, S.J., Richardson, K., Scachetti-Pereira, R., Schapire, R.E., Soberón, J., Williams, S., Wisz, M.S. & Zimmermann, N.E. (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, 29, 129–151.
- Elith, J., Graham, C., Valavi, R., Abegg, M., Bruce, C., Ferrier, S., Ford, A., Guisan, A., Hijmans, R.J., Huettmann, F., Lohmann, L., Loiselle, B., Moritz, C., Overton, J., Peterson, A.T., Phillips, S., Richardson, K., Williams, S., Wiser, S.K., Wohlgemuth, T. & Zimmermann, N.E. (2020) Presence-only and Presence-absence Data for Comparing Species Distribution Modeling Methods. *Biodiversity Informatics*, 15, 69–80.
- Elith, J., Kearney, M. & Phillips, S. (2010) The art of modelling range-shifting species. *Methods in ecology and evolution*, 1, 330–342.
- Elith, J., Leathwick, J.R. & Hastie, T. (2008) A Working Guide to Boosted Regression Trees. *Journal of Animal Ecology*, 77, 802–813.
- Flach, P. & Kull, M. (2015) Precision-Recall-Gain Curves: PR Analysis Done Right. *Advances in Neural Information Processing Systems*, pp. 838–846. Curran Associates, Inc.
- Pearce, J. & Ferrier, S. (2000) Evaluating the predictive performance of habitat models developed using logistic regression. *Ecological modelling*, 133, 225–245.
- Phillips, S.J., Dudík, M., Elith, J., Graham, C.H., Lehmann, A., Leathwick, J. & Ferrier, S. (2009) Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological applications*, 19, 181–197.
- Valavi, R., Elith, J., Lahoz-Monfort, J.J. & Guillera-Arroita, G. (2021) Modelling species presence-only data with random forests. *Ecography*, 44, 1731–1742.
- Valavi, R., Guillera-Arroita, G., Lahoz-Monfort, J.J. & Elith, J. (2022) Predictive performance of presence-only species distribution models: a benchmark study with reproducible code. *Ecological Monographs*, 92, 1–27.