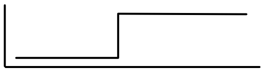
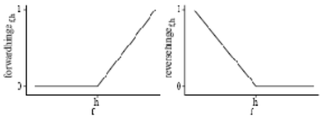
# Appendix 1: Details about features.

| Feature class | Description in relation to environmental variable | Constraint imposed on estimated distribution $\hat{P}$ | Ecological interpretation of the constraint |
|---|---|---|---|
| Linear (L) | Variable itself | The mean of variable under $\hat{P}$ should be close to its mean in the sample locations | The mean of the sample indicates average conditions for species presence |
| Quadratic (Q) | Square of variable | If used with L, variance of variable under $\hat{P}$ is close to its variance in the sample | The variation in that variable in the sample indicates the tolerance of the species for variation from suitable conditions |
| Product (P) | Product of 2 variables | If used with linear features for the 2 variables, that the covariance of the variables under $\hat{P}$ should be close to its covariance in the sample | The effect of one variable on species presence varies with the value of the other variable – i.e. there are interactions between the variables. |
| Threshold (T) | A step function that allows a different response below the threshold (the "knot") to that above it. Equivalent to a piecewise constant spline.  | The proportion of $\hat{P}$ that has values of this variable above the knot should be close to that proportion in the sample | Many threshold features can be used on the same variable, with different thresholds. These can add together to model an arbitrary stepped response to the variable. |
| Hinge (H) | Similar to the threshold feature, but the response above the knot (forward hinge; below left) or below the knot (reverse hinge; below right) is linear with a positive or negative coefficient (slope). Equivalent to a piecewise linear spline.   forwardhinge$_{c,h}$<br>1<br>0<br>h<br>reversehinge$_{c,h}$<br>1<br>0<br>h | The mean of the variable above the knot under $\hat{P}$ should be close to its mean above the knot in the sample locations | A model using only hinge features fits a piecewise linear response. If hinge features are used, linear features are redundant (a linear feature can be created from a hinge, with the knot at one extreme of the feature space). |
| Category (C) | A binary indicator showing membership in one class of a categorical variable. For a k-class categorical variable there will be k categorical features | The proportion of $\hat{P}$ that has values in this class should be close to that proportion in the sample | |

# Appendix 2: The transition from a geographic to environmental viewpoint.

Here we show the derivation for formula 2 in the main manuscript.

In Phillips *et al.* (2006) we see that the target is $\pi(x) = \Pr(x|y{=}1)$. This is a probability distribution on pixels $x$, and can be a difficult distribution to conceptualise. It is the probability, given the species is present (found), that it is found at pixel $x$.
MaxEnt produces the estimate

$$\begin{aligned}\hat{p}(x|y=1) &= q_\beta(x)\\ &= \frac{e^{\beta\cdot \boldsymbol{Z}(x)}}{Q_\beta}\end{aligned}$$

The notation above is as consistent with the notation in the main manuscript as possible:

$\boldsymbol{Z}(x)$ is a vector of environmental random variables defined over the sampling space of pixels

$\beta$ is a vector of weights (coefficients)

$q_\beta(x)$ is a probability distribution over $x$

$Q_\beta$ is a normalization constant that ensures $q_\beta$ sums to 1.

MaxEnt of course works with transformations of $\boldsymbol{Z}(x)$ – i.e. $h(\boldsymbol{Z}(x))$ – in its expanded feature set, but here for simplicity we will stick with $\boldsymbol{Z}(x)$. So given the pixel distribution $q_\beta(x)$, what is the induced distribution of $\boldsymbol{Z}(x)$ (the distribution of environmental variables, given species is present)? It is easier to do in discrete space in terms of sums:

$$\begin{aligned}f_1(z) &= \Pr(\boldsymbol{Z} = \boldsymbol{z}|y=1)\\ &= \sum_{x\in L:\boldsymbol{Z}(x)=\boldsymbol{z}} q_\beta(x)\\ &= \sum_{x\in L:\boldsymbol{Z}(x)=\boldsymbol{z}} \frac{e^{\beta\cdot\boldsymbol{Z}(x)}}{Q_\beta}\\ &= \frac{f(\boldsymbol{z})e^{\beta\cdot\boldsymbol{z}}}{Q_\beta/|L|}\\ &= f(\boldsymbol{z})e^{\alpha+\beta\cdot\boldsymbol{z}}\end{aligned}$$

where

$$f(\boldsymbol{z}) = \frac{\#[x\in L:\boldsymbol{Z}(x)=\boldsymbol{z}]}{|L|}$$

(i.e. the number of $x$ in L such that $\boldsymbol{Z}(x)$ equals $\boldsymbol{z}$, divided by the number of elements (pixels) in the landscape $L$) describes the distribution of $\boldsymbol{z}$ in the discrete landscape.

for a normalizing constant $\alpha$ that ensures that $f_1(\boldsymbol{z})$ integrates (sums) to 1.

# Appendix 3: More on the logistic output

Dudík and Phillips (2009) used a robust Bayesian approach to convert MaxEnt's raw output, which is an exponential-family model (formally resembling a GLM with a log link) into a logistic model, i.e., one using a logit link. Here we summarize that derivation. In order to estimate $\Pr(y{=}1|\mathbf{z})$, we first imagine the more general problem of making an estimate of $f(\mathbf{z},y)$, the joint density of covariates and the response variable in the landscape. The accuracy of such an estimate (which we denote $p$) can be measured by the expected relative log likelihood (relative to a null model) that $p$ assigns to random test data drawn from the study region, i.e.

$$E\ [\ln\ (p(\mathbf{z},y)/v(\mathbf{z},y)] \qquad\qquad ..........(A)$$

where $v(z,\mathrm{y})$ is a null model for $p(\mathbf{z},y)$: $v(\mathbf{z}) = f(\mathbf{z})$ and $v(y{=}1|\mathbf{z}){=}\ \tau$. However, we do not know $f(\mathbf{z},y)$ exactly, as we only have some samples drawn from $f(\mathbf{z})$ and $f_1(\mathbf{z})$. A good estimate should therefore have high expected log likelihood for any distribution $\pi$ that is consistent with the available data, i.e., it should maximize:

$$\min_\pi\ \mathrm{E}_\pi\ [\ln\ (p(\mathbf{z},y)/v(\mathbf{z},y))] \qquad\qquad ..........(B)$$

where the minimization is over densities $\pi(\mathbf{z},y)$ satisfying a set of constraints suggested by the available data; the true density $f(\mathbf{z},y)$ is just one of the densities considered in this minimization. Grünwald and Dawid (2004) proved that this formulation, which is sometimes called a "robust Bayesian" formulation, is equivalent (for certain types of constraints on $\pi$) to a maximum entropy formulation:

$$\min_p\ \mathrm{RE}(p\,||\,v) \qquad\qquad ..........\ (C)$$

where the minimization is over distributions $p$ satisfying the same constraints as in Eqn. B. Dudík and Phillips (2009) prove that when the constraints are derived from occurrence data (i.e., they only constrain $p(\mathbf{z}|y{=}1)$), the solution to Eqn. C satisfies:

$$p(y = 1|\mathbf{z})\ =\ \tau\, e^{\eta(\mathbf{z})-r}\ /\ ((1\text{-}\ \tau) + \tau\, e^{\eta(\mathbf{z})-r}) \qquad\qquad ..........(D)$$

where $\eta(\mathbf{z})$is the linear score from Eqn. 2 (main manuscript), such that $f(\mathbf{z})\ e^{\eta(\mathbf{z})}$ is the MaxEnt estimate (using the same constraints) of $f_1(\mathbf{z})$ and $r$ is the relative entropy of the estimate of $f_1(\mathbf{z})$ with respect to $f(\mathbf{z})$. The right hand side of Eqn. D is MaxEnt's logistic output. The term $e^{-r}$ in Eqn. D incorporates prevalence in a reasonable way since $f_1(\mathbf{z})$ has high relative entropy with respect to $f(\mathbf{z})$ only if the species is a specialist, which implies low prevalence. By the definition of relative entropy, $r$ is the average of $\eta(\mathbf{z})$ under $f_1(\mathbf{z})$. Therefore, sites with $\eta(\mathbf{z}) = r$ have "typical" conditions for the presence of the species; for such sites, Eqn. D simplifies to $p(y{=}1|\mathbf{z}) = \tau$.

The default logistic output from MaxEnt corresponds to $\tau = 0.5$. This constant can be adjusted, as discussed in the main manuscript. In the example of the jaguar (Box 2, main manuscript), if we chose to define the temporal and spatial scale of an observation such that probability of presence in a typical site within the jaguar's range is only 0.1, then we would use $\tau = 0.1$.

We note that the only constraints placed on $\pi$ for deriving MaxEnt's logistic output are constraints on $\pi(\mathbf{z}|y{=}1)$. In particular, $\pi(\mathbf{z})$ is not constrained to be equal to $f(\mathbf{z})$; this results in the estimate $f_0(\mathbf{z}){=}\ f(\mathbf{z})$ which is clearly incorrect but may be reasonable when the prevalence is very low. Additional constraints can be placed on $\pi$ to avoid this issue (Dudík and Phillips 2009), but the resulting models have not been found to improve empirical performance. Further research on this issue would be useful.

# Appendix 4: Case study 1.

The covariates used in the first case study included five climate variables from the set available from ANUCLIM (ANU, 2009), at 0.01degree (~1km) resolution. These variables are named in the main manuscript : isothermality (ISOTHERM), mean temperature of the wettest quarter (TEMPWETQ), mean temperature of the warmest quarter (TEMPWARMQ), annual precipitation (RAIN) and precipitation of the driest quarter (RAINDRYQ), but you will also see them referred to here as (respectively): clim03, clim08, clim10, clim12, clim17. These data were kindly provided by the Department of Environment, Heritage, Water and the Arts (DEWHA) and were originally supplied by Janet Stein (Fenner School of Environment and Study, ANU). From the limited soils data available Australia-wide, we selected an estimate of the solum plant-available water holding capacity (SOLWHC; derived by Janet Stein, and based on Western & McKenzie, 2006) as the most likely relevant variable to this species. All variables were available at a 0.01 degree (~ 1km) cell size, though the underlying data for the soils has variable precision across Australia (Janet Stein, pers.comm.). We present this as a demonstration study only, and recognize that, for rigorous application in this region, better soils data and predictors representing land transformation are needed for more precise predictions (Yates *et al.*, 2010). The future environment is represented by changes predicted under the A1FI scenario for 2070 estimated over the ensemble of 23 GCMs in IPCC report 4 (Solomon *et al.*, 2007); the SOLWHC was assumed to remain as it is now.

If you are new to MaxEnt, and particularly to running it from batch files, you might want to look at case study 2 first – for it, we supply example data, explain batch files, and annotate the code in some detail. It gives useful preliminary information not given in this Appendix. We do not provide the data for this case study, but here list our code from our batch files, so you can see how things were done. The ideas should be easily transferable to other data. First, though, we explain cross-validation, which will be used here.

## Cross-validation

Cross-validation is a straightforward, quick and useful method for resampling data for training and testing models. In *k*-fold cross-validation (Kohavi, 1995; Hastie *et al.*, 2009) the data are divided into a small number (*k*, usually 5 or10) of mutually exclusive subsets. Model performance is assessed by successively removing each subset, so you have one subset omitted and *k*-1 retained. You then fit your model on the retained data, and predict to the omitted data. You cycle through all the possibilities *k* times (the "folds"; see box below). By the end every site has been used in model fitting *k*-1 times (but in *k*-1 different combinations with other sites; see box below), and each site has been used for evaluation just once, and only when the model was not trained on that site. This is the elegance of cross-validation: it is structured, and you know that any evaluation will be performed on held-out data. Cross-validation is a widely used resampling method, and is fast and performs well. See Hastie et al. (2009) for further information, and explanations of alternative resampling approaches. Resampling methods that test on withheld data are appropriate tests of a model's predictive performance on data from within the same general set of environmental conditions as those used to fit the data – e.g., within the same geographic region, or the same time.

Box: the details of a ten-fold cross-validation

```
Loop 1:  Train model on folds: 2 3 4 5 6 7 8 9 10; Test against fold: 1
Loop 2:  Train model on folds: 1 3 4 5 6 7 8 9 10; Test against fold: 2
Loop 3:  Train model on folds: 1 2 4 5 6 7 8 9 10; Test against fold: 3
Loop 4:  Train model on folds: 1 2 3 5 6 7 8 9 10; Test against fold: 4
Loop 5:  Train model on folds: 1 2 3 4 6 7 8 9 10; Test against fold: 5
Loop 6:  Train model on folds: 1 2 3 4 5 7 8 9 10; Test against fold: 6
Loop 7:  Train model on folds: 1 2 3 4 5 6 8 9 10; Test against fold: 7
Loop 8:  Train model on folds: 1 2 3 4 5 6 7 9 10; Test against fold: 8
Loop 9:  Train model on folds: 1 2 3 4 5 6 7 8 10; Test against fold: 9
Loop 10: Train model on folds: 1 2 3 4 5 6 7 8 9;  Test against fold: 10
```

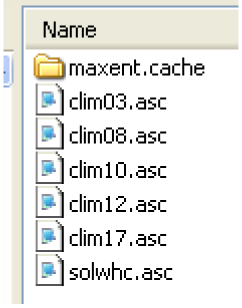
MaxEnt uses the cross-validation in 2 ways. First, it records the response curve and the predictions fitted under each of the *k* different training sets, and uses those to give an indication of variability in the results. It also uses the predictions to the omitted (with-held) sites for estimating test performance (e.g. for the reported test AUC). For AUC, MaxEnt not only needs predictions to sites with observed occurrence, but also to background sites. For background, it uses the same background data for training and evaluation of all replicate runs.

Now we give some examples of our data set-up for case study 1:

1. An example of the SWD format: the first few rows of the background (atlas sites) data file, which is comma separated:

```
id,x,y,clim03,clim08,clim10,clim12,clim17,solwhc
atlas,117.94874,-35.12101671,0.52,13.1,18.8,931,79,133
atlas,117.9473511,-35.1154611,0.51,12.6,18.4,937,82,133
atlas,117.9431843,-35.11268334,0.51,12.6,18.4,937,82,133
atlas,117.9404065,-35.10712776,0.51,12.8,18.5,927,81,154
```

2. An example of the files in the "current" and "future" directories. Both contain grids with identical names; the climate data are different, of course, for the future. Note that the names match those in the header of the species file (with the exception of the ".asc" suffix). MaxEnt has created the MaxEnt.cache directory, into which it has written the MaxEnt format of the ascii grids.



Name
maxent.cache
clim03.asc
clim08.asc
clim10.asc
clim12.asc
clim17.asc
solwhc.asc

3. Making background samples that take into account unequal cell areas across Australia: As explained in the main manuscript, there are several ways to approach this. We made our own background samples, and used code supplied by Robert Hijmans (UC Davis, CA) that utilizes his "raster" package for R (http://www.r-project.org/). This worked on a new laptop (admittedly slowly) for all of Australia at 1km grid cells (13.8 million grid cells). The following is JE's adaptation of RH's R code which could be adapted for other data. Comments after hash symbols explain the code. If you are new to R, you will need to find a tutorial or book that helps you understand it - many are available. Please understand that this code is provided as a helper, but the authors will not be able to write additional code for your particular set-up.

```
#load the raster library
library(raster)

#inform R about the raster locations:
r1 <- raster("c:\\wa2\\current\\clim03.asc")
r2 <- raster("c:\\wa2\\current\\clim08.asc")
r3 <- raster("c:\\wa2\\current\\clim10.asc")
r4 <- raster("c:\\wa2\\current\\clim12.asc")
r5 <- raster("c:\\wa2\\current\\clim17.asc")
r6 <- raster("c:\\wa2\\current\\solwhc.asc")

#make these into a stack for faster processing
st <- stack(r1, r2, r3, r4, r5, r6)
```

```
#make a sum of the stack to make a raster with NA's that represent missing data in any or all of the rasters r1 to r6
(the raster package knows to interpret missing data as NA). Note this is slow for large data sets. The alternative is to
take a larger sample than required then strip out the NA's at the end
r <- sum(st)

# set sample size
n <- 20000

# locate the cells with data for all rasters
cells <- which(!is.na(getValues(r)) )
# which rows are that?
rows <- rowFromCell(r, cells)
# what is the latitude?
y <- yFromRow(r, rows)
# what is the 'width' of a cell? This will be used to create the weights for sampling, in the subsequent row
dx <- pointDistance(cbind(0, y), cbind(xres(r), y), 'GreatCircle')

selected <- sample(cells, n, prob=dx)

#what are the coordinates of the selected cells? Plot them on a map.
xy <- xyFromCell(r, selected)
plot(r)
points(xy)

#now create a dataframe suitable for MaxEnt SWD format, inserting the cell values for the covariates
bg20k <- data.frame(rep("bg", 20000), xy[1:20000], xy[20001:40000], cellValues(st, selected))

names(bg20k) <- c("sp", "x", "y", "clim03", "clim08", "clim10", "clim12", "clim17", "solwhc")

#write out the data for use in MaxEnt:
write.table(bg20k, "c:\\wa2\\sites\\bg20k.csv", sep=",",row.names=F)
```

## Batch file code for the MaxEnt model runs

The runs 1 to 3 below run models 1 to 3, described in our manuscript. We annotate the code (green indented text) to explain those parts not covered in case study 2.

> **Introductory comments:** The code is in pairs; the a part (output to eg directory run1a) runs the model, does jackknife measures of variable importance (see MaxEnt's tutorial for details on these), and makes predictions to grids. The b part does a 10-fold cross-validation to get estimates of uncertainty for the response curves, and a cross-validated estimate of predictive performance. For this latter part we don't use a jackknife and tend to use SWD format. This saves time.
>
> **Commentary on first runs:** As for case study 2, this uses SWD format for the species data (pr.csv) and the background sites (atlas.csv). In this case, though, we project to grids in directories "current" and "future". Note use of 2 different projection directories (2 filepaths, separated by a comma). The flags used here and not in case study 2 are:
>
> nowarnings    MaxEnt will not produce a window with a report if something is amiss, but will just write the warning to the log file.

-r tells MaxEnt not to ask about overwriting existing files, but just to do it

adjustsampleradius a method for changing the size of the white squares used, on the maps, to show the occurrence records. A value of -10 makes them very small.

We only use hinge features, which we achieve by turning off the other feature types (nolinear, etc). **If you wanted to use all feature types** you would delete this part of the code:

nolinear noquadratic nothreshold noproduct

```
java -mx1500m -jar MaxEnt.jar -s C:\wa2\sites\pr.csv -e C:\wa2\sites\atlas.csv -o C:\wa2\run1a -j
C:\wa2\current,C:\wa2\future -P -J nowarnings nolinear noquadratic nothreshold noproduct -r -a adjustsampleradius=-
10
```

**Commentary on following:** The new flag here is:

nooutputgrids MaxEnt doesn't save the output prediction grids from each of the *k* cross-validation models; instead it just returns summaries: the average, max, min, median and std deviation.

```
java -mx1500m -jar MaxEnt.jar -s C:\wa2\sites\pr.csv -e C:\wa2\sites\atlas.csv -o C:\wa2\run1b -j
C:\wa2\current,C:\wa2\future -P nowarnings nolinear noquadratic nothreshold noproduct -r -a replicates=10
crossvalidate nooutputgrids
```

**Commentary on following:** Same as run1, but using different background data. Again, using SWD format (it is faster than asking MaxEnt to sample the grids)

```
java -mx1500m -jar MaxEnt.jar -s C:\wa2\sites\pr.csv -e C:\wa2\sites\bg10kSW.csv -o C:\wa2\run2a -j
C:\wa2\current,C:\wa2\future -P -J nowarnings nolinear noquadratic nothreshold noproduct -r -a adjustsampleradius=-
10
java -mx1500m -jar MaxEnt.jar -s C:\wa2\sites\pr.csv -e C:\wa2\sites\bg10kSW.csv -o C:\wa2\run2b -j
C:\wa2\current,C:\wa2\future -P nowarnings nolinear noquadratic nothreshold noproduct -r -a replicates=10
crossvalidate nooutputgrids
```

And the all-Australia background data, again in SWD format:

```
java -mx1500m -jar MaxEnt.jar -s C:\wa2\sites\pr.csv -e C:\wa2\sites\bg20kAUS.csv -o C:\wa2\run3a -j
C:\wa2\current,C:\wa2\future -P -J nowarnings nolinear noquadratic nothreshold noproduct -r -a adjustsampleradius=-
10
java -mx1500m -jar MaxEnt.jar -s C:\wa2\sites\pr.csv -e C:\wa2\sites\bg20kAUS.csv -o C:\wa2\run3b -j
C:\wa2\current,C:\wa2\future -P nowarnings nolinear noquadratic nothreshold noproduct -r -a replicates=10
crossvalidate nooutputgrids
```

## Running limiting factors and MESS maps

See the 2010 MaxEnt tutorial, and Elith et al. (2010, and particularly the online supplement) for an explanation of both of these and instructions for running them. Since Elith et al. (2010) was published, MaxEnt has been updated. A link to MESS maps is now automatically provided in the html output, for the data / projection combination used in the model.

## Code for five-fold cross-validation

The different models (1 to 3, see main manuscript) use the same presence sites but different backgrounds for training. Therefore, the AUC's estimated in MaxEnt's cross-validation ($\text{AUC}_{\text{Maxent}}$) are not comparable, because they use different background data sets. For evaluation we want a consistent dataset across models, but also one that was not used for model training. We could do one split of the training data, and just have one training set and one test set, but cross-validation is more thorough, tending to give a lower variance estimate than split samples (Hastie et al. 2009). Hence we used a "manual" five-fold cross-validation (i.e. one where we pre-specified the data sets, and calculated the evaluation statistics outside of MaxEnt; $\text{AUC}_{\text{Manual}}$) to ensure consistent evaluation data sets across all four models.

We divided the presence records into 5 mutually exclusive subsets (using the same ideas as presented in the cross-validation section earlier in this appendix). For each training run (each fold), 4 of these were combined into a training set, and the 5th set aside as the data to which the model was predicted. We also need absence or background data. In this case we used the atlas sites at which *Banksia prionotes* was not recorded, and treated them as absences. So, the "absence" atlas sites were also divided into 5 subsets, and each combined with a presence subset to give 5 independent evaluation data sets. Note that for model 1, which uses atlas sites as background, we ensured that the "testing" (evaluation) sites for any given fold were not used as background in the training data, to guarantee that each model's training set was independent of the testing set. In other words, for model 1 we had different subsets of background data for each cross-validation fold.

We then ran 15 MaxEnt models (3 models each on five datasets). Example code for the first "fold" is given below. Note the directory set-up for the output directories: we have 5 separate directories for the five folds, with subdirectories within that for each model type. This made it easy then to write R code, using loops to run through the data and read it into R.

> **Commentary on the following:** Model 1 (i.e. presence sites, atlas sites). Uses the first fold's training data (train1_pres.csv, which is 80% of the presence records), the first fold's background data (train1_all.csv; this is only 80% of the atlas sites, and contains no sites in the testing set); and predicts to the first testing data set (test1.csv).

```
java -mx1500m -jar MaxEnt.jar -s C:\wa2\sites\train1_pres.csv -e C:\wa2\sites\train1_all.csv -o C:\wa2\cv1\atlasall -j C:\wa2\sites\test1.csv -P -J nowarnings nowarnings nolinear noquadratic nothreshold noproduct  -r –a
```

> **Commentary on the following:** Model 2. As above, but uses all 10000 background points for each fold of the cross-validation. Similarly, for the all-Australia background in the third model..

```
java -mx1500m -jar MaxEnt.jar -s C:\wa2\sites\train1_pres.csv -e C:\wa2\sites\bg10kSW.csv -o C:\wa2\cv1\10khinge -j C:\wa2\sites\test1.csv -P -J nowarnings nolinear noquadratic nothreshold noproduct -r -a
```

```
java -mx1500m -jar MaxEnt.jar -s C:\wa2\sites\train1_pres.csv -e C:\wa2\sites\bg20kAUS.csv -o C:\wa2\cv1\20khinge -j C:\wa2\sites\test1.csv -P -J nowarnings nolinear noquadratic nothreshold noproduct -r -a
```

For all these, we had called the species banpri (in column 1 of the training presence files), so predictions would be in a file called e.g. banpri_test1.csv, in column 3. We kept track of the observations (presence or absence) for each site in the testing data sets, so could at the end make a final file where:

- rows were sites, and all five testing sets were merged.
- columns were: (1) observation; (2 to 5) predictions for models 1 through 3.

We then calculated the area under the receiver operating characteristic curve ($\text{AUC}_{\text{Manual}}$) and point biserial correlation coefficient (COR) for all models, using custom code in R. See Elith et al. (2006) for further explanations of these statistics, and the main manuscript and Appendix 5 for results and discussion.

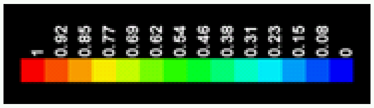
## Appendix 5: Case study 1- model summaries

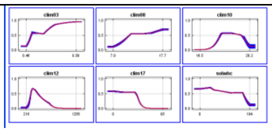
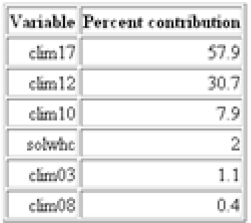
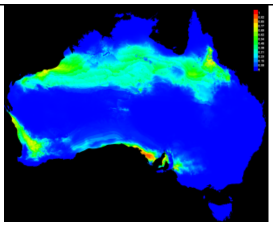
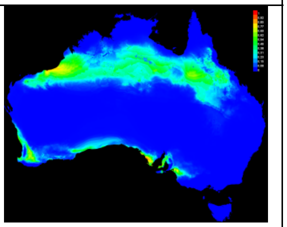
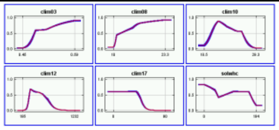
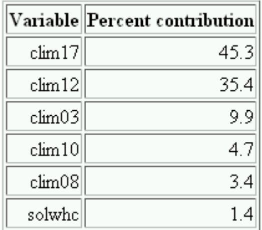
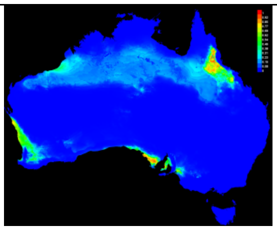
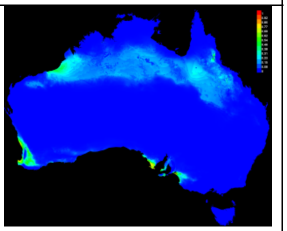
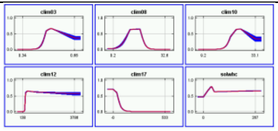
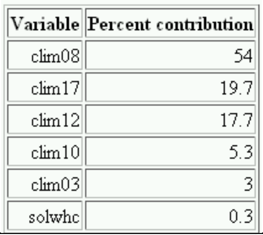
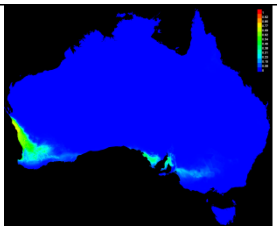
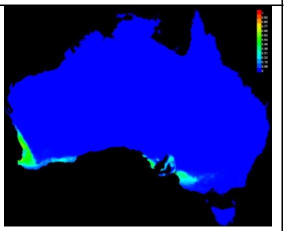
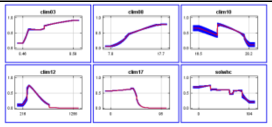
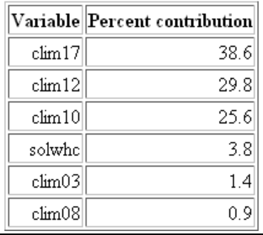
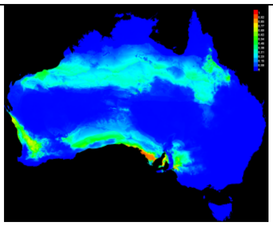
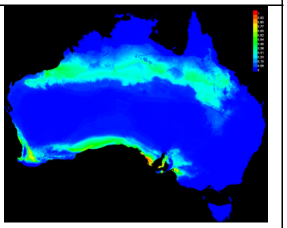
Here we present more information on the modeled responses, variable importance and mapped predictions. We have noted in Appendix 4 that clim03, clim08, clim10, clim12 and clim17 are, respectively, isothermality (ISOTHERM), mean temperature of the wettest quarter (TEMPWETQ), mean temperature of the warmest quarter (TEMPWARMQ), annual precipitation (RAIN) and precipitation of the driest quarter (RAINDRYQ).

**Table S5-1 – A repeat of Table 3 from the main manuscript, with on additional row for model 4.**

| Model (background; feature classes) | Variable importance | | | | | | $AUC_{Maxent}$ (10fold CV but varying data sets) | $AUC_{Manual}$; COR (5fold CV on atlas data) |
|---|---|---|---|---|---|---|---|---|
| | RAIN DRYQ | RAIN | TEMP-WARM Q | TEMP-WETQ | ISO-THERM | SOL-PWHC | | |
| 1 (atlas; hinge) | 57.9 | 30.7 | 7.9 | 0.4 | 1.1 | 2.0 | 0.92 | 0.96; 0.62 |
| 2 (southwest; hinge) | 45.3 | 35.4 | 4.7 | 3.4 | 9.9 | 1.4 | 0.90 | 0.93; 0.52 |
| 3 (Australia; hinge) | 19.7 | 17.7 | 5.3 | 54.0 | 3.0 | 0.3 | 0.99 | 0.91; 0.45 |
| 4 (atlas; all) | 38.6 | 29.8 | 25.6 | 0.9 | 1.4 | 3.8 | 0.92 | 0.96; 0.64 |

**Table S5-2 – More details on the models.** Legend for maps is the default MaxEnt one:



| Model | Modelled responses | Variable importance | Current predictions | Future predictions |
|---|---|---|---|---|
| Model 1. Atlas background, only hinge features] ($AUC_{Maxent}$ = 0.92) |  | <br>Variable Percent contribution<br>clim17 57.9<br>clim12 30.7<br>clim10 7.9<br>solwhc 2<br>clim03 1.1<br>clim08 0.4 |  |  |
| Model 2. Background in south-west western Australia. Only hinge features. ($AUC_{Maxent}$ = 0.90) |  | <br>Variable Percent contribution<br>clim17 45.3<br>clim12 35.4<br>clim03 9.9<br>clim10 4.7<br>clim08 3.4<br>solwhc 1.4 |  |  |
| Model 3. Background across all Australia, only hinge features. ($AUC_{Maxent}$ = 0.99) |  | <br>Variable Percent contribution<br>clim08 54<br>clim17 19.7<br>clim12 17.7<br>clim10 5.3<br>clim03 3<br>solwhc 0.3 |  |  |
| Model 4. Atlas background, all feature types ($AUC_{Maxent}$ = 0.92) |  | <br>Variable Percent contribution<br>clim17 38.6<br>clim12 29.8<br>clim10 25.6<br>solwhc 3.8<br>clim03 1.4<br>clim08 0.9 |  |  |

## Effect of background data and evaluation dataset

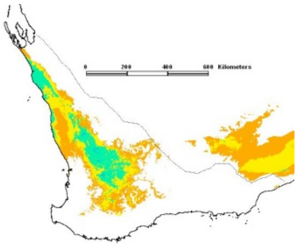
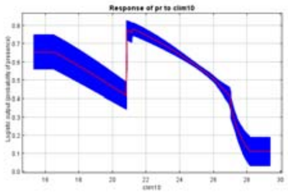
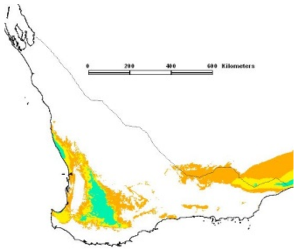
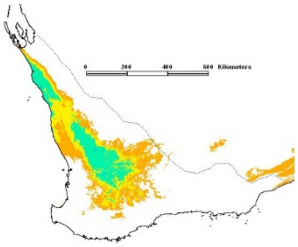
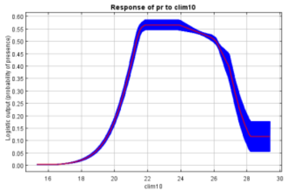
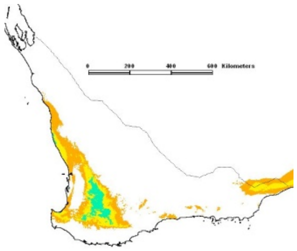
The results presented above highlight the fact that choice of data sets – here, background data and evaluation data – influence what is modelled and perceptions about the results. See previous comments on 5-fold cross-validation compared with AUC$_{\text{Maxent}}$ results provided by MaxEnt (Appendix 4, and main ms) for background information on the evaluation datasets. AUC$_{\text{Maxent}}$ results use background data supplied for the model, and these data differ across our set of models, so the statistics are not comparable across our models. The key problem in comparing the results is that the range of environments vary across the datasets (the differing number of sites is not much of an issue here). Table S5-1 shows that if you used the AUC$_{\text{Maxent}}$ as a guide, Model 3 appears best. It is estimated on background across Australia, so in the Model 3 dataset there are many more unsuitable sites for the species than in the data for Models 1 and 2, which are restricted to the South West Australia Floristic Region (SWAFR). A dataset with many unsuitable sites for a species will tend to give higher AUC than one with more subtle distinctions (providing a reasonable model is possible for both) because the prediction task is easier. In contrast, the 5-fold CV evaluation (AUC$_{\text{Manual}}$ ) on Banksia Atlas data uses consistent data to test across models. The Banksia Atlas data used here are restricted to the SWAFR, so the test on that data asks how well the models predict in the SWAFR. It ranks models 1 and 2 more highly than model 3.

Clearly, we could extend this approach and could also set up a continent-wide set of consistent data to test the models at a continental scale, if we were interested in that question. We haven't done that, but let us consider the mapped distributions continent-wide. (Note: we reported on the mapped distributions across the SWAFR in the main manuscript, and noted that predictions of Models 2 and 3 reflect the biases in the survey locations. We will not consider those local – but important – differences here). The maps of current distribution in Table S5-2 suggest Model 3 works well continent-wide because it restricts predictions to the correct general area, and doesn't predict habitat in areas where *B. prionotes* is known not to occur. The reason that Model 3 can restrict predictions mostly to the south-west is that the model is based on background across all of Australia. So, the model has successfully found features that discriminate between the occupied sites in the south-west, and the rest of Australia. However, as discussed in the manuscript, there is some doubt about whether there is good ecological justification for using continent-wide background. Models 1 and 2 are modelling something different – they are finding the set of features that help to discriminate between occupied environments and those in atlas (model 1) or SWAFR (model2) background sites. Appendix 6 shows that – if used to predict across the continent - models 1 and 2 are extrapolating to new, unsampled conditions in the north of Australia, even for current climates. The mapped information about novel conditions could be used as a guide to prediction uncertainty, or at least as a warning that these models are being projected into novel environmental space. In summary, the selected background needs to make sense ecologically, but also with thought to the required application of the model, and knowledge of novel environments in the projection space.

## All feature types vs only hinge features

In addition to the three models presented in our manuscript, we also ran another version of Model 1 but with all feature types (as described in the code supplied in Appendix 4). We called that Model 4. Results are presented in Tables S5.1, S5.2 and S5.3. Compared with using all feature types (Model 4), using only hinge features (Model 1) simplified the fitted functions (see Table S5.3) without substantially affecting the predicted distribution in the inhabited region. There is no obvious basis for deciding which model (4 or 1, i.e. all feature types or only hinge) is more correct. Both models predict current and future distributions that are sensible ecologically and not remarkably different. Both have similar predictive performance on held-out data (Table S5.1). There is some rearrangement of variable importance between the two models (Table S5.2), though this is most likely because the two most highly correlated variables (TEMPWARMQ and RAINDRYQ, r = -0.83) are substituting for each other (see the MaxEnt tutorial available for download with the program for further discussion of this general issue). This substitution probably leads to the (reasonably subtle) differences in the future predicted distributions within the SWFRWA (Table S5.3). None of these issues are particular to MaxEnt; models fitted on correlated variables need careful interpretation for all methods (Dormann et al. unpubl. manuscript).

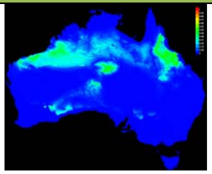
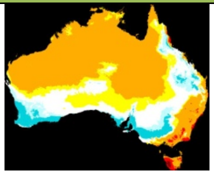
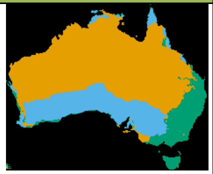
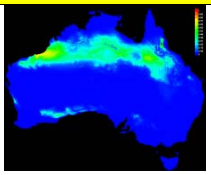
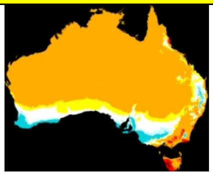
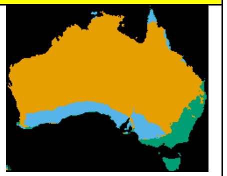
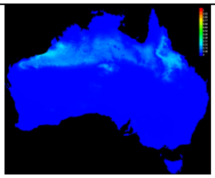
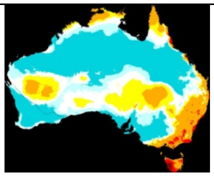
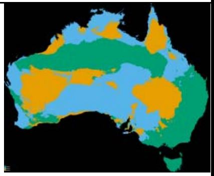
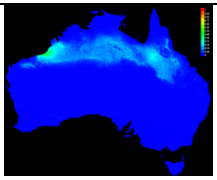
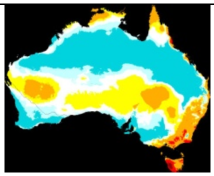
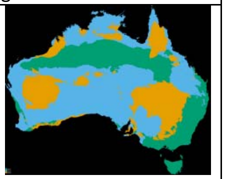
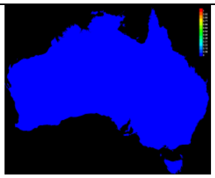
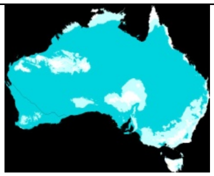
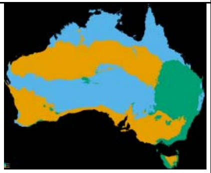
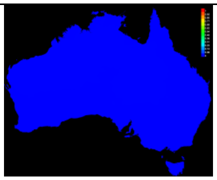
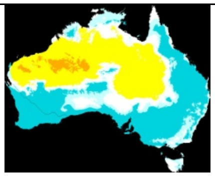
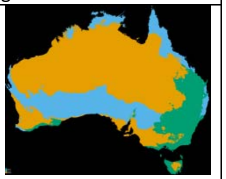
### Table S5.3: Model results comparing use of all feature types with use of only hinge features

| # | Background, features | Predicted current distribution in inhabited area | Fitted functions for TEMPWARMQ | Predicted future distribution in inhabited area |
|---|---|---|---|---|
| 4 | Atlas, all |  |  |  |
| 1 | Atlas, hinge |  |  |  |

## Appendix 6: Case study 1- predictions across Australia for current and future environments

The figures below compare information from clamping (the process by which features are constrained to remain within the range of values in the training data) with that from MESS (multivariate environmental similarity surface) maps. Legends for clamping are defaults in MaxEnt (see Appendix 5), with blue lowest and red highest. The legend to the right is for the MESS maps, where increasing negative numbers show increasing degree of extrapolation on at least one variable. MESS maps are giving different information to that from the clamping, and are most useful for assessing the novelty of the projection space. Note that the legends for limiting factors change with the model. The limiting factors also vary with the model, because each model has a different task given the different backgrounds.

| | |
|---|---|
| 🟫 | -500 - -300 |
| 🟥 | -300 - -100 |
| 🟧 | -100 - -10 |
| 🟨 | -10 - -2 |
| ⬜ | -2 - 2 |
| 🟦 | 2 - 10 |
| 🟦 | 10 - 100 |
| ⬛ | No Data |

| | Current clamping | Current MESS | Current limiting factors | Future clamping | Future MESS | Future limiting factors |
|---|---|---|---|---|---|---|
| **Model 1 (atlas)**<br>MESS maps based on clim17,12,10. |  |  | <br>orange- clim 10<br>blue- clim12<br>green-clim17 |  |  | <br>orange- clim 10<br>blue- clim12<br>green-clim17 |
| **Model 2 (swfra)**<br>MESS maps based on clim17,12,03 |  |  | <br>orange- clim 3<br>blue- clim12<br>green-clim17 |  |  | <br>orange- clim 3<br>blue- clim12<br>green-clim17 |
| **Model 3 (aus)**<br>MESS maps based on clim08, 17,12 |  |  | <br>orange- clim 8<br>blue- clim12<br>green-clim17 |  |  | <br>orange- clim 8<br>blue- clim12<br>green-clim17 |

# Appendix 7: Species data and predictors for case study 2.

The species data were obtained from the Victorian Department of Sustainability and Environment (DSE) (Aquatic Fauna Database, supplied 1 June 2007), the Murray-Darling Basin Commission (SRA Asset SRA524, supplied 4 May 2007) and Koster *et al*. (2006). They were processed to remove conspicuous errors, and exhaustively cross-checked against auxillary information and spatial datasets to validate positional accuracy. These data will be described in detail in a forthcoming manuscript (Chee & Elith unpubl. ms). For this case study we only extracted presence records, though the full data set could be analysed as presence-absence data if assumptions are made about detection rates for survey methods. A comprehensive set (159) of environmental predictors linked to a river network were available (Chee & Elith, unpubl. ms), but not all were likely relevant to blackfish distribution, and many were highly correlated with others. Following the view that a relevant, proximal set of predictors are preferable to a large collection of available but potentially irrelevant covariates (Elith & Leathwick, 2009), we reduced the set to those presented below, based on pairwise correlations and published information on the species .

**Table S7.1: Predictor variables for the distribution of *Gadopsis bispinosus***

| Variable | Description | Mean & range |
|---|---|---|
| ***Segment and local watershed*** | | |
| SLOPE_PERC | link slope calculated as rise/run * 100 (percent) | 2.59, 0.00-47.44 |
| CV_TEMP | mean coefficient of variation of mean annual temperature in link watershed | 1.7, 1.4-1.9 |
| MAXWARMP_TEMP | mean maximum temperature of warmest week in link watershed | 27.5, 16.9-31.3 |
| MEAN_MRVBF | mean multi-resolution valley bottom flatness (MrVBF) index value in link watershed. MrVBF is an expression of local relief in terms of valley confinement and floodplain extent with values typically ranging from 2.5 in narrow confined valleys to ≥ 8 in broad floodplains. | 2.37, 0.00-8.99 |
| NONATVEG | mean proportion of link watershed without any native vegetation cover | 0.29, 0.0-1.00 |
| TRECOV | mean proportion of link watershed covered by any type/combination of tree cover | 0.59, 0.00-1.00 |
| ***Entire upstream path or catchment*** | | |
| DRAINAGE_DENS | drainage density in link UCA (km/km2) | 1.85, 0.05-15.0 |
| TOTLENGTH_UCA | sum of length of all links within link UCA (km) | 913, 0.06-18620 |
| UC_RIP_TRECOV | mean proportion of riparian zone in link UCA covered by any type/combination of tree cover | $7.52\text{x}10^{-1}$, $4.80\text{x}10^{-3}$-1.00 |
| UC_SOLDEPTH | mean solum depth in link UCA (m) | 1.09, 0.50-1.50 |
| UC_CV_PPT | mean coefficient of variation of mean annual precipitation in link UCA | 28.6, 14.0-37.0 |
| UCWARMQT_PPT | mean precipitation of warmest quarter (any 13 consecutive weeks) in link UCA | 181, 67-341 |
| UC_TWI | mean topographic wetness index (TWI) value in link UCA | 9.7, 7.6-19.1 |
| UC_ROADDENS | road density in link UCA (km/km$^2$) | 1.40, 0.00-7.56 |
| US_MAXSLOPE | maximum slope encountered along link upstream flow path (percent) | 50.70, 0.00-310.37 |
| ***Downstream*** | | |
| DS_AVGRIPTRECOV | mean riparian tree cover along downstream flow path | 0.49, 0.00-0.94 |
| DS_AVGSLOPE | average slope encountered along link downstream flow path (percent) | 0.16, 0.00-3.02 |
| DS_MINCOLDP_TEMP | mean minimum temperature of the coldest week in watersheds along link downstream flow path | 2.4, 0.8-3.3 |
| ***Geographic position*** | | |
| RIV | site membership within one of four major catchment-based regions in study area : 1 = Goulburn-Broken; 2 = Avoca-Wimmera; 3 = Campaspe-Loddon; 4 = Ovens-Kiewa-Mitta-Upper Murray. | - NA - |

# Appendix 8: Case study 2.

For running models like those in case study 2, we provide data in a zip file as part of the online supplementary information. The instructions below include a method for linking the predictions to a shape file of rivers, using the free statistical software, R (http://www.r-project.org/ ). The result can be viewed in any GIS program, including the free software DIVA-GIS (http://www.diva-gis.org/ ). You may have other methods for these last steps.

Our instructions use a batch file, rather than MaxEnt's point-and-click interface. The same thing could be done via the interface. Batch files are easy, though, so first we explain them.

## Batch files

See http://en.wikipedia.org/wiki/Batch_file for a general explanation of batch files. The batch file that comes with MaxEnt, MaxEnt.bat, is just a text file with instructions in it. You can open it and look at the contents by opening Notepad (Start/programs/accessories/notepad) or some other text editor and then browsing to the file (if using Notepad you will need to select Files of type: All files) and opening it. Alternatively, right-click on the .bat file, and select "Edit". Note that double-clicking on it (or right-clicking and selecting "Open") will not open it for inspection; it will make it do its work. You need to open it from within a text editor / edit it to see its contents. In its original form with no alterations (i.e. as downloaded with MaxEnt), it will contain this:

```
java -mx512m -jar MaxEnt.jar
```

This tells the computer to start Java, to allow 512MB memory (RAM) for the program, then to execute the jar file, MaxEnt.jar, which is the MaxEnt program.

You can alter details (such as the amount of memory allocated to the program) and add code to run MaxEnt from the batch file rather than the interface. This saves a lot of clicking of buttons and allows you to leave many models running. It's not hard to do.

## Running a model – general explanations

Extract all the files from the .zip file into one directory. You will have a batch file (MaxEnt.bat), the data needed to run the model and predict to sites , and a shape file for later viewing of predictions. You need to get the MaxEnt program (the MaxEnt.jar file) from MaxEnt's download site (http://www.cs.princeton.edu/~schapire/MaxEnt/). The data files MaxEnt will use are:

gadbis.csv – the presence records  
background.csv – all background sites  
rivers.csv – all river segments we want to predict to.

Note that these are in "SWD" (samples with data) format, which is explained in the tutorial that comes with MaxEnt. They are comma delimited. Columns 1 to 3 are always the same types: column 1 is an identifier, the same for every record in the file (unless multiple species are being used, in which case it's the same for every record for one species). Columns 2 and 3 are the x and y coordinates, respectively. You'll see we just have dummy data in those, since MaxEnt isn't using this information for this example (we're not sampling grids). Then the following columns are the predictors, one column per variable. These must have exactly consistent naming (in row 1) for all files – i.e., the occurrence records, the background records and the sites for prediction. As an aside, if you were using SWD format for the presence records combined with grids for the environmental variables, the grids would have to have identical names to those in the header of the SWD file (eg clim1 in the SWD would match clim1.asc in the grid).

Open the batch file (remember to use a text editor, don't double-click it), and look at the contents. There is code for 2 models. The first is:

```
java -mx1500m -jar MaxEnt.jar  -s c:\fishtute\gadbis.csv -e c:\fishtute\background.csv  -j c:\fishtute\rivers.csv -o c:\fishtute\out1 -P -J -a -t RIV  nopictures
```

First, notice that the filepaths in the code assume you've extracted the data to a folder called fishtute, which is directly on c drive.  You can change the filepath to suit your setup. To understand the code above and to learn other possibilities, look in the MaxEnt help file (click "Help" on the main interface), and find  the section on "Batch mode" and the table that lists the "flags". Flags are command line arguments that indicate the status of a setting.  They might be binary, where the setting can only have 2 states, and you can toggle between them. Or they might require further information, such as a file path or a variable name. Everything you can do (and more) at the interface can be done with code. If you tick a check-box or enter a file path, there will be a way to do it with code. In the code above the first bits, in orange, are the standard bits, though we're using 1500MB of RAM. Then the flags used are:

| | |
|---|---|
| -s | followed by the location of the species file, shows where it is |
| -e | same as above, but for environmental data directory |
| -j | showing that you want to use the projection facility – this is a way to predict to sites that you specify (it's also used to predict to new regions or times). Again, the string following the flag shows where the data are. If you wanted to predict to several data sets you can list them all, separated by commas. |
| -o | followed by a directory location, shows where the output is to go, You will need to create the directory you specify here (i.e. make a folder called "out1") |
| -P | draw the response curves; using it here toggles the value from the default, false, to true. |
| -J | use jackknife to measure variable importance; as above, toggles from the default value, false, to true. |
| -a | start immediately, without waiting for Run button to be pushed |
| -t | followed by the name of an environmental variable in your environmental data directory, toggles the data type from the default (continuous) to categorical. |

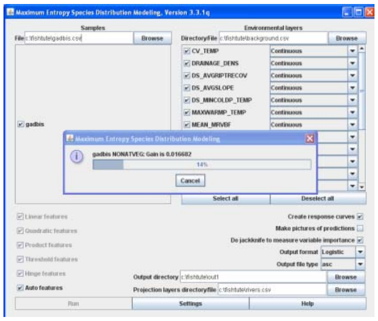
Finally, "nopictures" means we don't want pictures (maps) of the predictions. This is just a different way to toggle something; picture-making has no abbreviation. Another way to say the same thing would be "dontpictures".

You can now probably work out what the second model, below, is doing. It's the same as the above, except it's writing results to a different directory (out2), and using 10-fold cross-validation. The "replicates=10" is the 10-fold bit, and "cross-validate" tells it to take the samples organized into folds necessary for cross-validation, rather than random samples.  We run this model so that we get estimates of uncertainty for the response curves, the predictions, and the statistics that MaxEnt calculates (e.g., AUC).  We explain cross-validation Appendix 4.

```
java -mx1500m -jar MaxEnt.jar  -s c:\fishtute\gadbis.csv -e c:\fishtute\background.csv  -j c:\fishtute\rivers.csv -o c:\fishtute\out2  -P -J -a -t RIV nopictures replicates=10 crossvalidate
```

## Running a model – steps

1. We assume you've extracted the files, into a directory c:\fishtute.  If your filepath is different, change the code in the batch file, as explained earlier
2. Make 2 new directories within c:\fishtute, called "out1" and "out2". Again, if you want to use different names, change the code appropriately in the batch file.
3. Double-click on MaxEnt.bat.  This will run both models in the batch file.  It will bring up the interface (example right), and you will see that file paths exist for the samples (the species data), environmental layers (the background data), output, and projection (the file you're predicting to, rivers.csv). Check boxes for response curve and jackknife measures of importance are ticked, because


Maximum Entropy Species Distribution Modeling, Version 3.3.1q

Samples

File c:\fishtute\gadbis.csv

☑ gadbis

Environmental layers

Directory/File c:\fishtute\background.csv

| | |
|---|---|
| ☑ CV_TEMP | Continuous |
| ☑ DRAINAGE_DENS | Continuous |
| ☑ DS_AVGRIPTRECOV | Continuous |
| ☑ DS_AVGSLOPE | Continuous |
| ☑ DS_MINCOLDP_TEMP | Continuous |
| ☑ MAXWARMP_TEMP | Continuous |
| ☑ MEAN_MRVBF | Continuous |

Maximum Entropy Species Distribution Modeling

gadbis NONATVEG: Gain is 0.016682

14%

☑ Linear features
☑ Quadratic features
☑ Product features
☑ Threshold features
☑ Hinge features
☑ Auto features

Create response curves ☑
Make pictures of predictions □
Do jackknife to measure variable importance ☑
Output format Logistic
Output file type asc
Output directory c:\fishtute\out1
Projection layers directory/file c:\fishtute\rivers.csv

you toggled them to "true", and picture-making is not checked, because you commanded "nopictures". Amongst the environmental predictors one (RIV) is listed as "categorical", consistent with the toggle commands in the batch file. In other words, you could run the same model from the interface if you wanted to, by clicking or entering file paths, as appropriate.

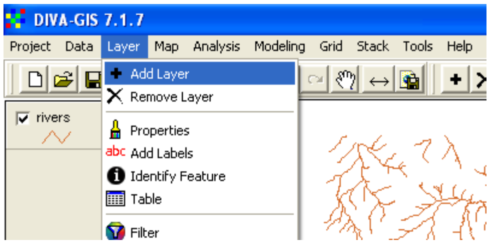
Don't click anything – just leave it to run. The models will take some minutes to finish. The slowest parts are the jackknife measures of variable importance, especially when combined with cross-validation (which will repeat everything 10 times).

## Exploring the output

Once finished, explore the contents of "out1" and "out2". The MaxEnt tutorial and help files (provided with the program) explain them.

## Mapping predictions

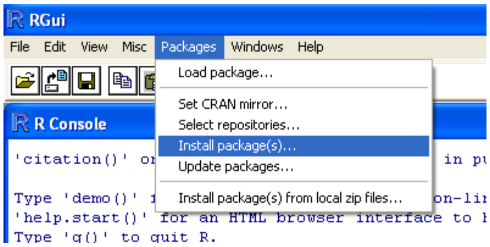
Our rivers are provided as a shape file that can be viewed within a GIS. For instance, DIVA-GIS (http://www.diva-gis.org/download) is free and handles shape files (we did this with version 7). In DIVA, to view the layer go to Layer/Add layer and browse to the rivers shape file.



Once the shape file is loaded, you'll see the river segments we've provided, and if you go to Layer/Table you'll see the fields with data for the various environmental variables linked to the segments. Details about the construction of this river network and the environmental attributes estimated will be available in Chee and Elith (unpubl. ms). To the very right of the table you'll see a column "GADBIS" full of zeros – this is provided by us, for adding the predictions. The rivers.csv file you predicted to, in MaxEnt, has exactly the same row order as this file, so the predictions can be directly added .

Remove the layer now, so you can add the predictions to it (Layer/Remove layer). Here we hit a snag because it's not straightforward. We want to add data to the .dbf part of the shape file. Read the Diva-GIS blog on some issues associated with this, and some fixes (http://www.diva-gis.org/dbf_trouble). Here we do it in R (http://www.r-project.org/) , but you could follow one of the other methods mentioned on that website.

Start R. You will need to have the package "foreign", so install it if you don't have it:



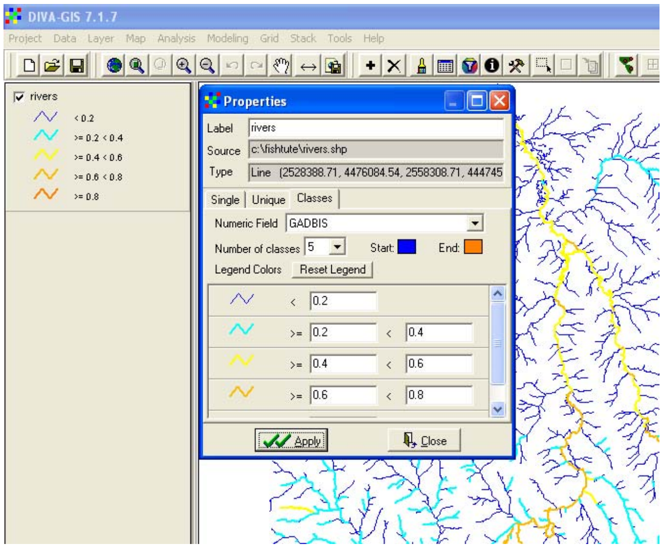
```
'citation()' o                          in pu

Type 'demo()' f                         on-lin
'help.start()' for an HTML browser interface to h
Type 'q()' to quit R.
```

Then these lines of code will work, if your file paths are the same as ours:

```
library(foreign)
shape <- read.dbf("c:/fishtute/rivers.dbf")
pred.tute <- read.csv("c:/fishtute/out1/gadbis_rivers.csv")
```

```
shape$GADBIS <- pred.tute[,3]
write.dbf(shape, "c:/fishtute/rivers.dbf")
```

Now add the layer again to your GIS. To view the predictions you'll need to set it to be viewing the correct field, and you might want to set up a suitable legend. For instance, in DIVA-GIS go to Layers/Properties. Select the "classes" tab, chose the correct field ("GADBIS"), and set up a legend by selecting number of classes, and altering the breaks, and the colours and line thicknesses. Now you can view the predictions, which should be comparable to those in our paper.



Label: rivers
Source: c:\fishtute\rivers.shp
Type: Line (2528388.71, 4476084.54, 2558308.71, 444745
Numeric Field: GADBIS
Number of classes: 5
Start: 🟦
End: 🟧
Legend Colors

| | | | | |
|---|---|---|---|---|
| | < | 0.2 | | |
| | >= | 0.2 | < | 0.4 |
| | >= | 0.4 | < | 0.6 |
| | >= | 0.6 | < | 0.8 |

# References

ANU (2009) ANUCLIM http://fennerschool.anu.edu.au/publications/software/anuclim.php#contacts. In:

Chee, Y. & Elith, J. (unpubl. ms) Predicting fish species distributions in freshwater ecosystems using a purpose-built GIS stream network and boosted regression trees.

Dudík, M. & Phillips, S.J. (2009) Generative and discriminative learning with unknown labeling bias. *Advances in Neural Information Processing Systems*, **21**: 401-408

Elith, J. & Leathwick, J.R. (2009) Conservation prioritization using species distribution models. *Spatial conservation prioritization: quantitative methods and computational tools* (ed. by A. Moilanen, K.A. Wilson & H.P. Possingham), pp. 70-93. Oxford University Press.

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

Hastie, T., Tibshirani, R. & Friedman, J.H. (2009) *The elements of statistical learning: data mining, inference, and prediction, second edition*, 2nd edn. Springer-Verlag, New York.

Keating, K.A. & Cherry, S. (2004) Use and interpretation of logistic regression in habitat selection studies. *Journal of Wildlife Management*, **68**, 774-789.

Kohavi, R. (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence* pp. 1137-1143. San Mateo, CA.

Koster, W., Crook, D., O'Mahony, D. & Fairbrother, P. (2006) Surveys of Fish Communities in the lower Goulburn River. Final Report to the Goulburn Valley Association of Angling Clubs. 53pp. Arthur Rylah Institute for Environmental Research, Department of Sustainability and Environment, Heidelberg, Victoria

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

Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K.D., Tignor, M. & Miller, H.L. (eds) (2007) *Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*.

Ward, G. (2007) *Statistics in ecological modeling; presence-only data and boosted mars*. Stanford University, Palo Alto.

Western, A. & McKenzie, N. (2006) Soil Hydrological Properties of Australia - User Guide. CRC for Catchment Hydrology

Yates, C., McNeill, A., Elith, J. & Midgley, G. (2010) Assessing the impacts of climate change and land transformation on Banksia in the South West Australian Floristic Region. *Diversity and Distributions*, **16**, 187-201.