HABITUS

Habitat Analysis and Biodiversity Integrated Toolkit for Unified Species Distribution Modelling

A Hands-On Tutorial

Ömer K. Örücü · Seda Örücü
Department of Landscape Architecture, Faculty of Architecture,
Süleyman Demirel University, 32260 Isparta, Türkiye
omerorucu@sdu.edu.tr · ORCID 0000-0002-2162-7553
sedaorucu@sdu.edu.tr · ORCID 0000-0003-1592-5180

INTRODUCTION

This tutorial gives a hands-on introduction to HABITUS, a standalone desktop application for species distribution modelling (SDM). It assumes no programming experience and no prior installation of R, QGIS or Python. Every step below is performed through the graphical interface, and the screenshots are taken from a real analysis rather than mock-ups.

HABITUS brings the whole SDM workflow into one window: data preparation, multicollinearity screening, model fitting with thirteen algorithms, projection to future climate, range change analysis, evaluation, independent validation and automated reporting. The distinguishing feature is that these stages are *sequential*: each tab stays locked until the previous one has completed successfully. You cannot accidentally fit a model before deciding which variables to use, or validate a map that does not yet exist.

The worked example throughout is *Pinus brutia* Ten. (Turkish red pine), using 108 occurrence records and 23 environmental layers covering Türkiye, with CNRM-ESM2-1 climate projections for the SSP5-8.5 scenario and the EUFORGEN chorological map as an independent reference. The dataset is distributed with the software, so you can reproduce every figure in this document.

For the theoretical background to the methods used here, see Guisan & Thuiller (2005) and Elith & Leathwick (2009); for reporting standards, Zurell et al. (2020). Full references are listed at the end.

GETTING STARTED

Downloading and installing

HABITUS is distributed as a self-contained installer. Nothing else needs to be installed — the Python runtime and all scientific libraries are bundled. Download the installer for your platform from:

https://github.com/omerorucu/habitus/releases/latest

On Windows, run `HABITUS_Setup_v1.0.0.exe` and follow the prompts. The installer and the application are code-signed, so Windows SmartScreen will show the publisher name rather than an unknown-publisher warning. On macOS, open the `.dmg` and drag HABITUS to Applications; on Linux, make the AppImage executable and run it.

Firing up

Start HABITUS from the Start menu, Applications folder or the AppImage. The following window appears:

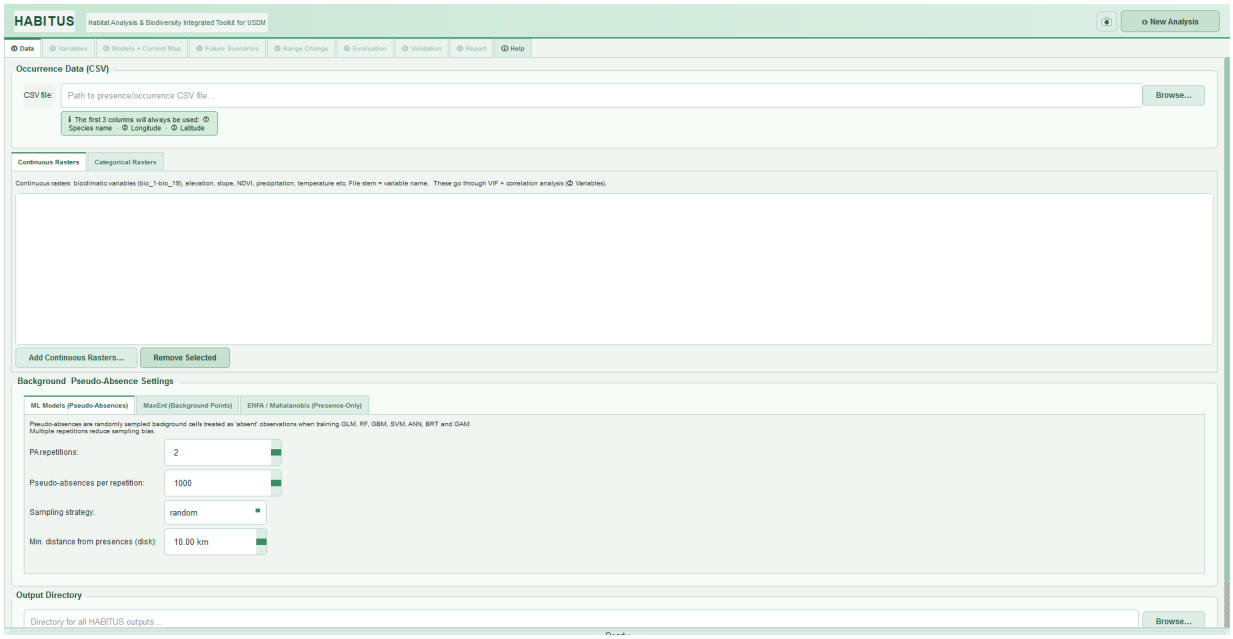


Figure 1. The HABITUS window at startup. Note that only Data and the Help tab are available — the remaining seven stages are greyed out until their prerequisites are met.

The eight numbered tabs across the top are the workflow. They unlock in order as you complete each stage. The Help tab is always available and contains the full user guide, the recommended minimum sample sizes and citation details. The panel at the bottom is the log: everything HABITUS does is written there with a timestamp, and the same log is saved to your output folder.

> **Note.** If you reconfigure an earlier stage — for example, load a different CSV — everything downstream is cleared automatically. This is deliberate: it prevents a map produced from one dataset being reported alongside statistics computed from another.

THE DATA YOU NEED

Two things are required. First, a comma-separated file of occurrence records. HABITUS always uses the first three columns, in this order: species name, longitude, latitude. Column headings may be anything; only the order matters. Our file begins:

```
presence,long,lat
Pinus brutia,35.35278,37.59167
Pinus brutia,35.38333,37.54167
Pinus brutia,30.40833,40.38667
```

Second, environmental raster layers. HABITUS reads ESRI ASCII grids (`.asc`) and GeoTIFF. All layers must share the same geographic extent and cell size. The file name becomes the variable name, so `bio_ (15).asc` is read as `bio_15`; keep the names consistent between your current and future layer sets, because HABITUS matches them by name.

Layers are declared as either continuous or categorical. Categorical variables — vegetation class, soil type, land cover — must be coded as integers, not text, and are placed on the second tab so that the software knows not to treat them as numeric gradients.

> **Note.** Sample size matters more than any modelling choice. With fewer than about 30 records the models become unstable; Pearson et al. (2007) discuss what can and cannot be done with small samples. The Help tab carries a table of recommended minima per algorithm.

LOADING DATA

Open the Data tab and press **Browse…** next to “CSV file”. Select your occurrence file. HABITUS immediately reads the header and reports what it found:

Figure 2. After selecting the CSV, a green banner confirms which column was interpreted as species, longitude and latitude. If your columns are in a different order, you will see it here rather than discovering it later from an implausible map.

Next, add the environmental layers. Press **Add Continuous Rasters…** and select all 23 files at once. If you have categorical layers, switch to the “Categorical Rasters” tab and add them there.

Finally, set the output directory. HABITUS writes every map, table, figure and log to this folder. If you leave it untouched, a folder named after the species with a timestamp is created under your Documents directory automatically — which is usually what you want, since it keeps successive analyses separate.

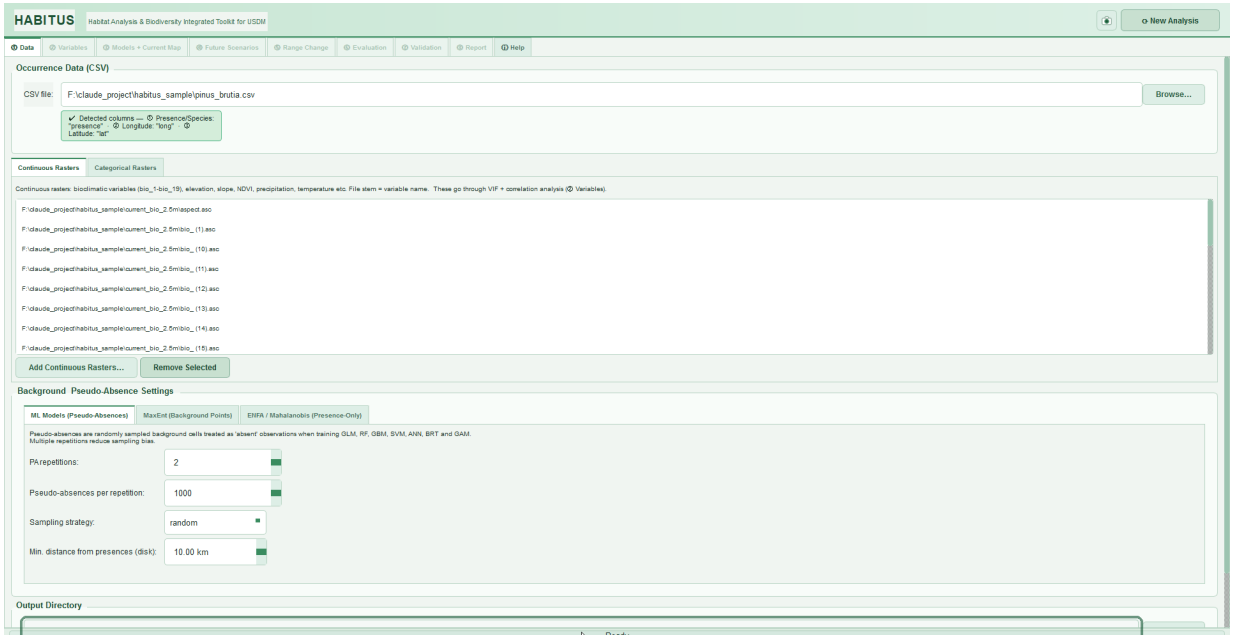


Figure 3. The Data tab fully configured: occurrence file, 23 continuous rasters, and the background sampling settings. The three sub-tabs at the bottom hold independent settings for the three model families.

BACKGROUND AND PSEUDO-ABSENCES

Occurrence records tell you where the species *is*, never where it is *not*. Most algorithms nevertheless need something to contrast presences against, and that contrast set has to be constructed. This is one of the most consequential decisions in the whole workflow (Barbet-Massin et al., 2012), which is why HABITUS gives it a panel of its own rather than a hidden default.

Three sub-tabs appear because the three model families want different things:

ML Models (Pseudo-Absences) — GLM, RF, GBM, SVM, ANN, BRT, GAM and the boosting algorithms treat sampled background cells as absences. Set how many replicates to draw (*PA repetitions*) and how many points per replicate. Drawing two or more replicates lets you see whether your result depends on one particular random draw.

MaxEnt (Background Points) — MaxEnt does not use absences at all; it compares presences against the environmental conditions available across the region, typically characterised by 10,000 background points (Phillips et al., 2006).

ENFA / Mahalanobis (Presence-Only) — these describe the environmental envelope occupied by the presences themselves.

Three sampling strategies are available. *Random* draws anywhere in the study region. *Disk* excludes a buffer around known presences, which is sensible when your records are spatially clustered. *SRE* (surface range envelope) draws only from environmental conditions outside the observed envelope, producing a more conservative contrast.

When everything is set, press **▶ Load Data & Generate Background/Pseudo-Absence Points**. HABITUS extracts environmental values at each record, drops points that fall outside the rasters, on a NoData cell or duplicate an existing coordinate, and reports exactly how many records survived:

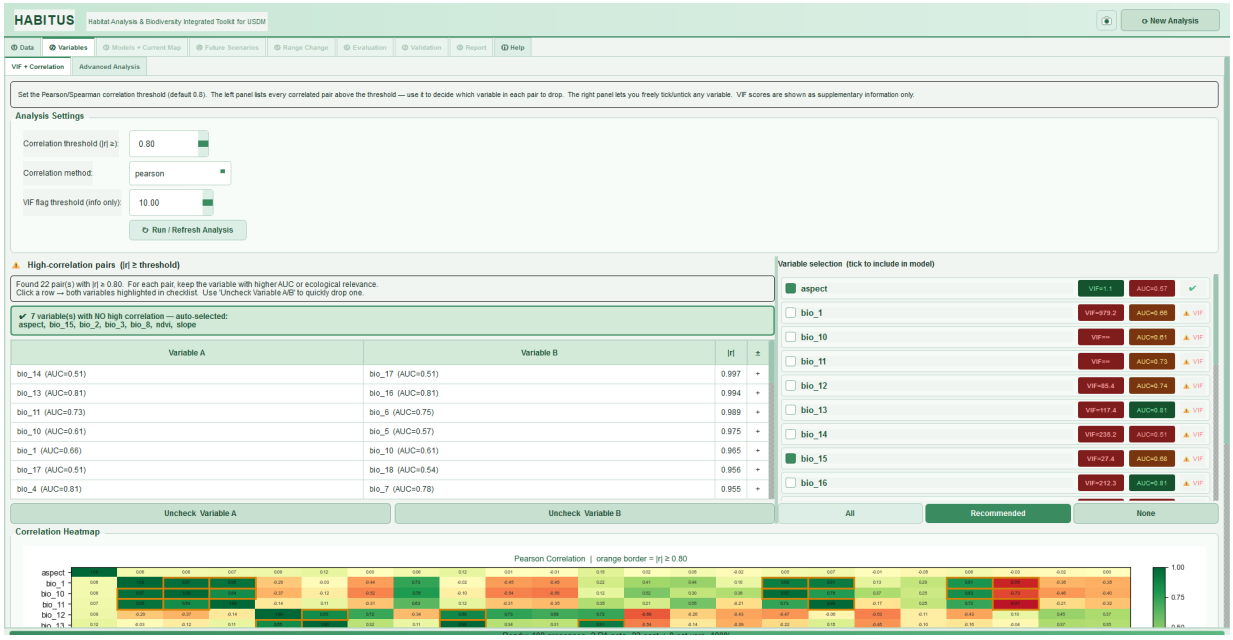


Figure 4. Data loading complete. All 108 records entered the analysis, two pseudo-absence replicates of 1,000 points were drawn and 10,000 MaxEnt background points generated. The Variables tab has unlocked and its analysis has already run.

SELECTING VARIABLES

Bioclimatic variables are derived from the same monthly temperature and precipitation surfaces, so they are strongly intercorrelated by construction: the mean temperature of the coldest quarter and the annual mean temperature carry nearly the same information. Feeding all of them to a model destabilises its coefficients and makes variable importance misleading (Dormann et al., 2013). The Variables tab exists to prevent this.

Two thresholds control the screening. The **correlation threshold** (default |r| ≥ 0.80) decides which pairs are flagged, and the **VIF flag threshold** (default 10) marks variables whose variance is inflated by collinearity (O'Brien, 2007). Press **↻ Run / Refresh Analysis** to compute them.

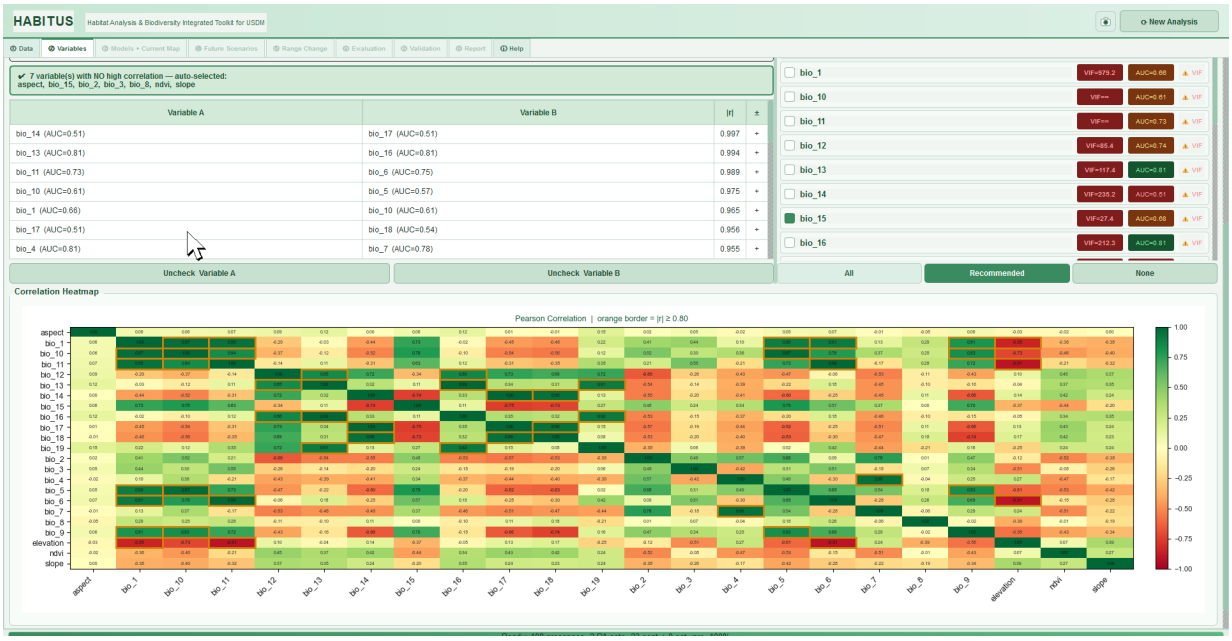


Figure 5. The correlation heatmap. Blocks of mutually correlated bioclimatic variables are normal and expected — this is the pattern the screening step is designed to handle.

Below the settings, HABITUS lists every pair exceeding the threshold together with the univariate AUC of each member — that is, how well each variable discriminates presence from background on its own. This gives you an objective basis for deciding which of the two to keep.

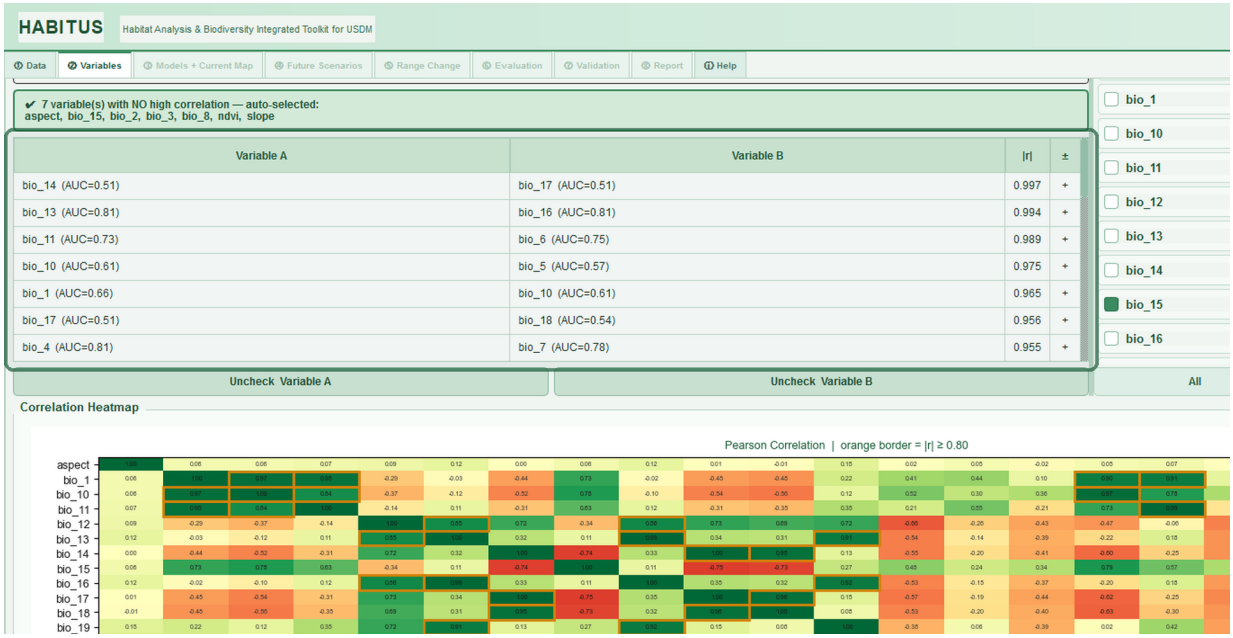


Figure 6. High-correlation pairs. In this dataset 22 pairs exceeded |r| ≥ 0.80. Clicking a row highlights both variables in the checklist on the right; "Uncheck Variable A/B" drops one of them immediately.

The right-hand panel is the actual selection. Three buttons operate on it: **All**, **None**, and **Recommended**. The last selects the variables that appear in no high-correlation pair — a defensible starting point, though not a substitute for ecological judgement. If you know that a particular variable is physiologically meaningful for your species, keep it and drop its correlate instead.

> **Note.** A variable flagged VIF=999 is not merely collinear — it is almost exactly a linear combination of the others. HABITUS caps the value rather than printing a number like 1.9×10^{13}, but such a variable should essentially always be removed.

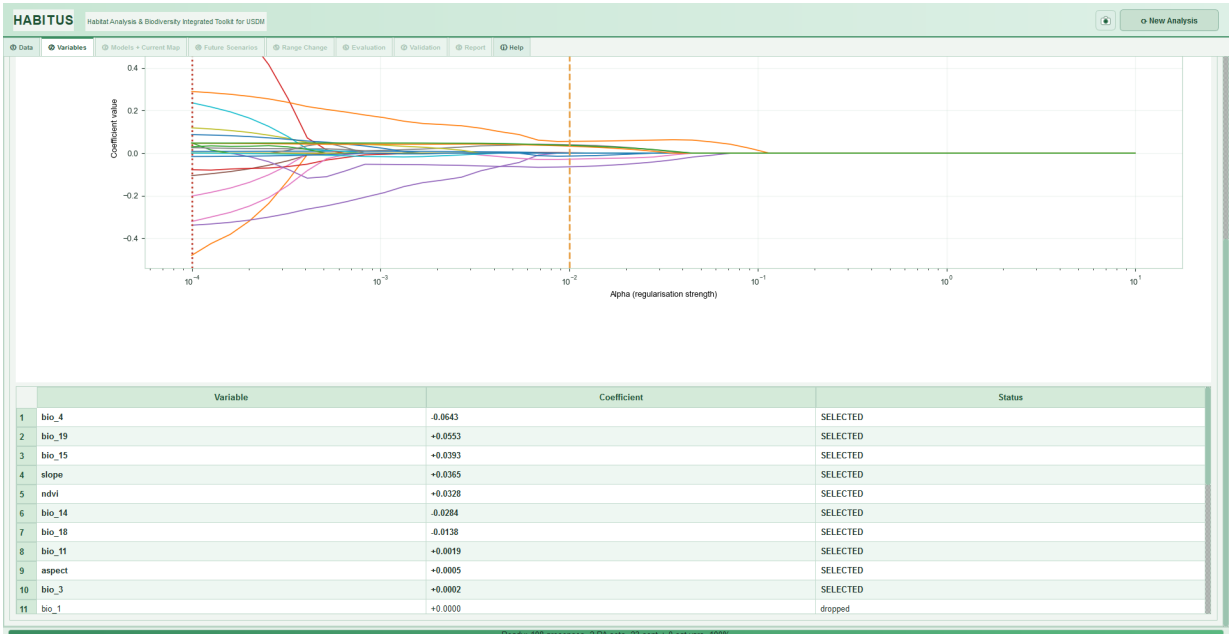
ADVANCED SCREENING

VIF and correlation are pairwise diagnostics; they can miss structure that involves several variables at once. Switch to the **Advanced Analysis** sub-tab and press **Run All Analyses** for four further checks.

Condition Number — a single figure describing how ill-conditioned the whole predictor matrix is (Belsley et al., 1980). Values above 30 indicate serious collinearity regardless of what the pairwise correlations show.

PCA — how many dimensions your variables actually occupy. If the first two or three components explain most of the variance, your 23 layers are measuring perhaps three independent things.

LASSO (L1) — shrinks coefficients toward zero and sets redundant ones exactly to zero, so it performs selection by itself (Tibshirani, 1996). The alpha value is chosen by cross-validation.



	Variable	Coefficient	Status
1	bio_4	-0.0643	SELECTED
2	bio_19	+0.0553	SELECTED
3	bio_15	+0.0393	SELECTED
4	slope	+0.0365	SELECTED
5	ndvi	+0.0328	SELECTED
6	bio_14	-0.0284	SELECTED
7	bio_18	-0.0138	SELECTED
8	bio_11	+0.0019	SELECTED
9	aspect	+0.0005	SELECTED
10	bio_3	+0.0002	SELECTED
11	bio_1	+0.0000	dropped

Figure 7. The LASSO regularisation path. Each line is one variable's coefficient as the penalty increases; the dashed vertical line marks the cross-validated alpha. The table below reports which variables survive (SELECTED) and which are driven to zero (dropped).

Ridge (L2) — shrinks coefficients but never eliminates them (Hoerl & Kennard, 1970). Because it keeps correlated variables stable rather than arbitrarily picking one of a pair, reading Ridge alongside LASSO is more informative than either alone: variables that LASSO drops but Ridge retains with a substantial coefficient deserve a second look.

When the four diagnostics point the same way, you can be reasonably confident in the selection. Return to the first sub-tab and press **Confirm Variable Selection → Proceed to Models**. In our example eight variables were retained: aspect, bio_2, bio_3, bio_8, bio_9, bio_15, NDVI and slope.

FITTING MODELS

The Models tab offers thirteen algorithms in three families. Tick the ones you want; each has its own settings sub-tab where hyperparameters (number of trees, learning rate, depth, kernel and so on) can be adjusted. The defaults follow values commonly used in the literature and are a reasonable starting point.

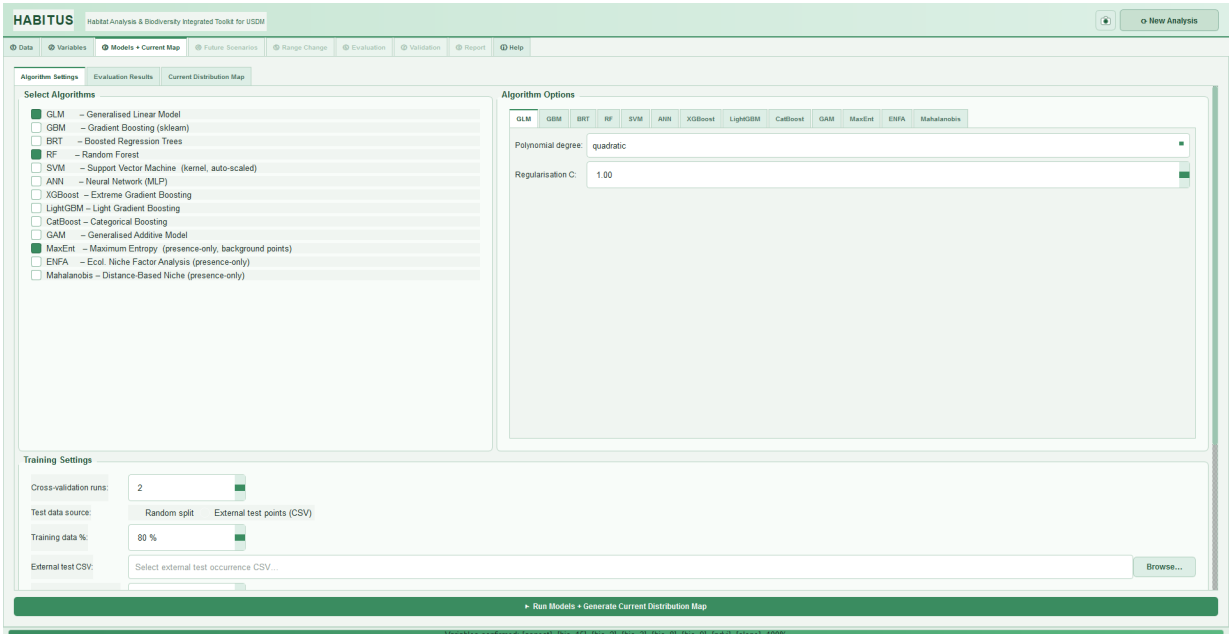
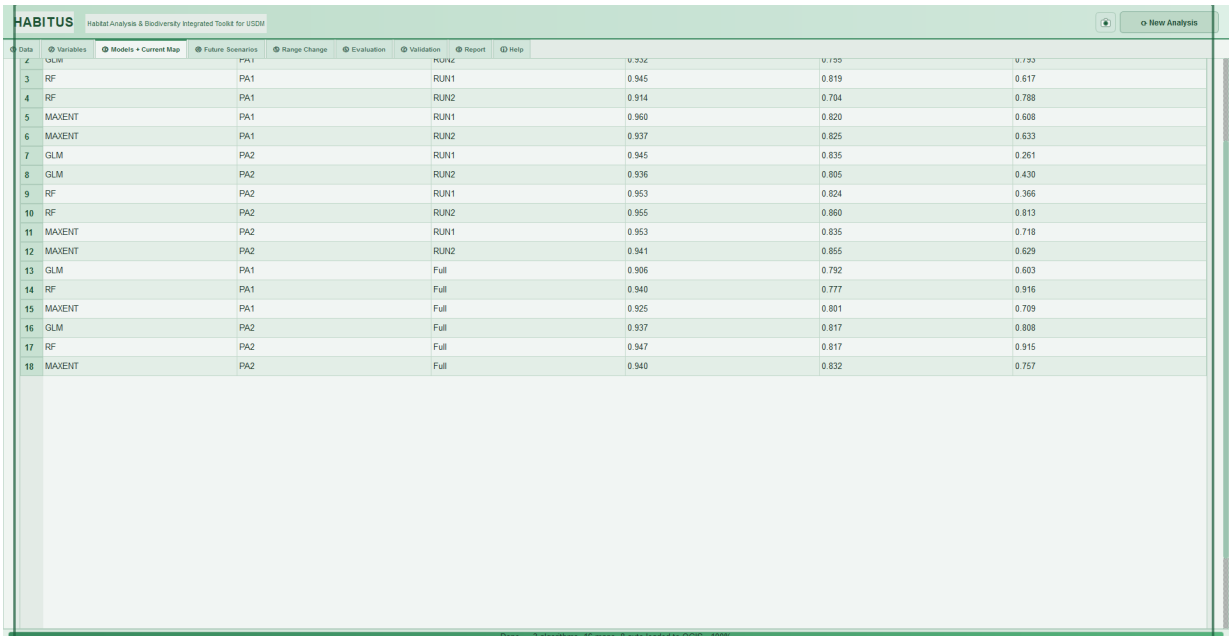


Figure 8. Algorithm selection. For this tutorial we use GLM (classical regression), RF (machine learning) and MaxEnt (presence-only) — one representative from each family.

Two further settings matter. The **train/test split** is stratified: presences and background points are divided separately so the held-out set keeps the same prevalence as the training set. The **ensemble** combines the fitted models, either as a weighted mean in which each model contributes in proportion to its evaluation score, or by committee averaging of the binary predictions. Ensembles are generally more stable than any single model (Araújo & New, 2007).

Press **▶ Run Models + Generate Current Distribution Map**. Each algorithm is fitted on each pseudo-absence replicate, cross-validated, scored, and finally projected onto the current climate layers. Depending on the number of algorithms and the size of your rasters this takes from under a minute to some tens of minutes; progress is reported in the status bar and the log.



3	RF	PA1	RUN1	0.945	0.819	0.617
4	RF	PA1	RUN2	0.914	0.704	0.788
5	MAXENT	PA1	RUN1	0.960	0.820	0.608
6	MAXENT	PA1	RUN2	0.937	0.825	0.633
7	GLM	PA2	RUN1	0.945	0.835	0.261
8	GLM	PA2	RUN2	0.936	0.805	0.430
9	RF	PA2	RUN1	0.953	0.824	0.366
10	RF	PA2	RUN2	0.955	0.860	0.813
11	MAXENT	PA2	RUN1	0.953	0.835	0.718
12	MAXENT	PA2	RUN2	0.941	0.855	0.629
13	GLM	PA1	Full	0.906	0.792	0.603
14	RF	PA1	Full	0.940	0.777	0.916
15	MAXENT	PA1	Full	0.925	0.801	0.709
16	GLM	PA2	Full	0.937	0.817	0.808
17	RF	PA2	Full	0.947	0.817	0.915
18	MAXENT	PA2	Full	0.940	0.832	0.757

Figure 9. The evaluation results table. Every algorithm × pseudo-absence replicate × cross-validation run appears as a separate row with its AUC, TSS and Boyce index, followed by the ensemble rows.

LOOKING AT A PREDICTION

Switch to the **Current Distribution Map** sub-tab. The ensemble prediction is displayed automatically, with the occurrence points overlaid.

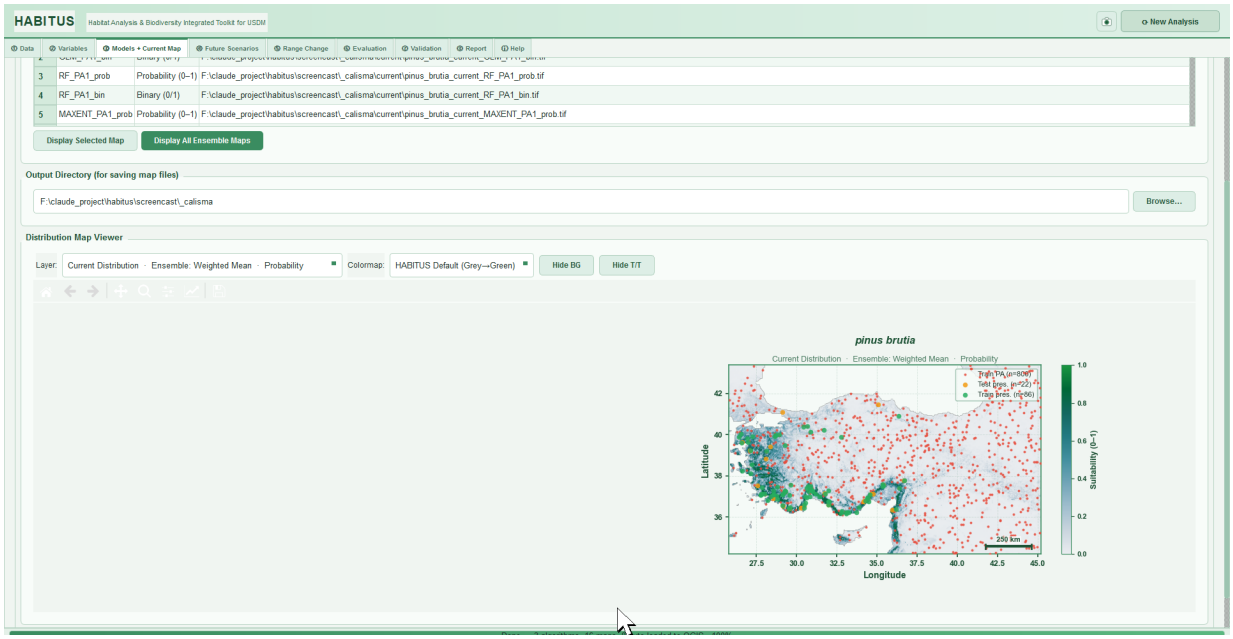


Figure 10. The embedded map viewer. The layer selector switches between individual algorithms, pseudo-absence replicates, ensembles, and probability versus binary versions of each.

Several controls are worth knowing. The **Colormap** selector offers eight palettes, including colourblind-safe (Viridis) and ecology-standard (YlGnBu) options. **Hide BG** and **Hide T/T** toggle the background points and the train/test presence points. The save icon in the toolbar writes the current view at 300 dpi, named automatically after the species and layer.

Every map is also written to disk as a GeoTIFF in your output folder, so you can open the same layer in QGIS or ArcGIS for cartographic finishing. Two versions are produced for each model: `_prob` (continuous suitability, 0–1) and `_bin` (presence/absence after thresholding).

PROJECTING TO FUTURE CLIMATE

The Future Scenarios tab applies your fitted models to a different set of climate layers. Give the scenario a name — it will label the output folder and appear in the range change tab — then add the layers with **Browse Folder…** or **Add Files…**.

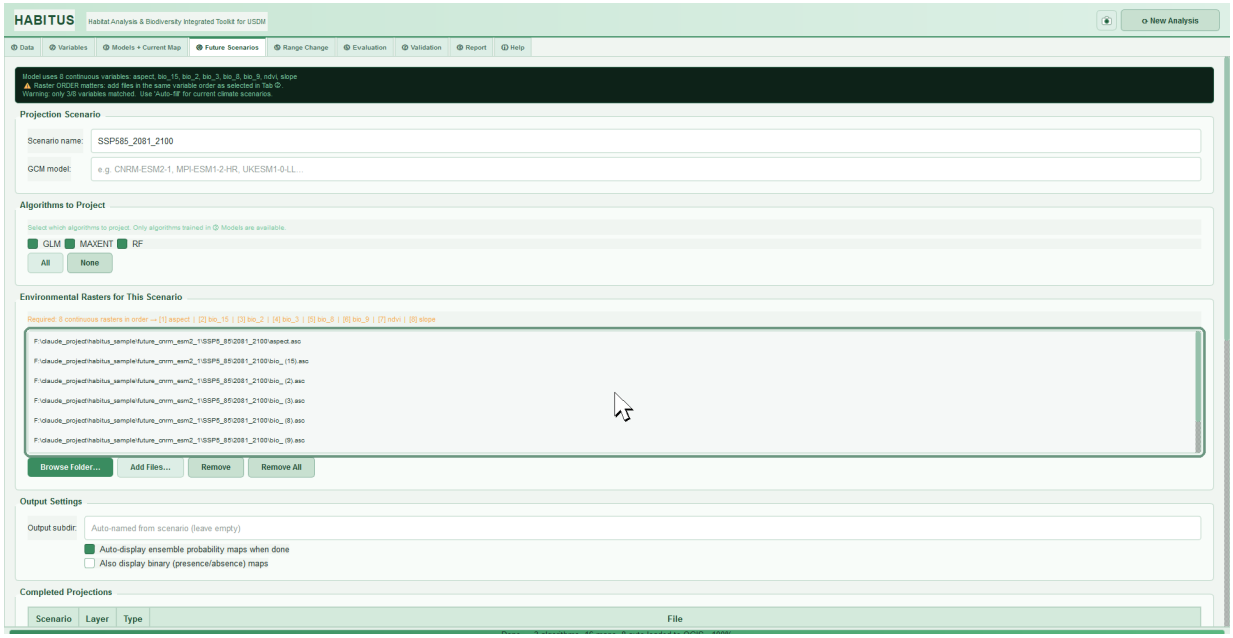


Figure 11. Adding the layers for the SSP5-8.5, 2081–2100 scenario. Only the variables actually used by the model are added, in the same order.

> **Note.** This is the step where mistakes are easiest to make. The layer set must correspond *exactly* — in composition and in order — to the variables the model was trained on. HABITUS refuses to run otherwise and tells you which variables it expected. Categorical variables have no meaningful future counterpart and are reused from the training data automatically, so do not add them here.

Choose which algorithms to project (**All** is usual) and press ▶ **Run Projection**. A map is produced for each algorithm plus the ensembles, all written as GeoTIFFs under a sub-folder named after the scenario.

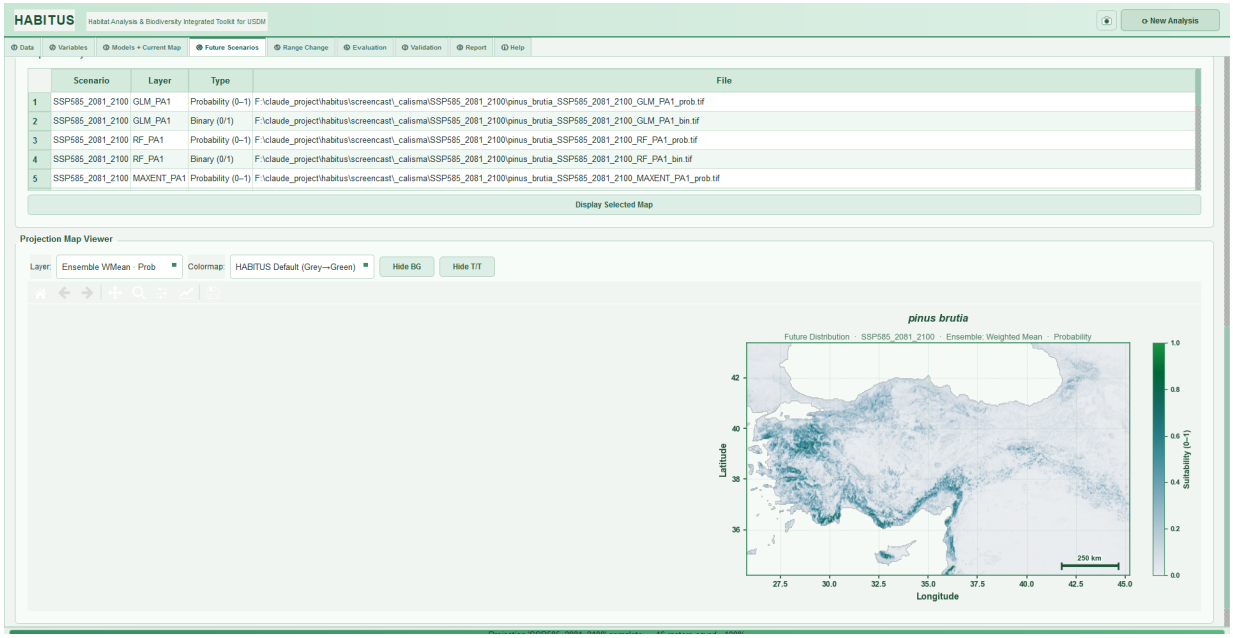


Figure 12. The projected distribution under SSP5-8.5 for 2081–2100. Comparing this with Figure 10 shows suitability contracting along the western lowlands and expanding inland to the north.

You may add as many scenario–period combinations as you like. Each is stored separately, and any two of them can be compared in the next step.

RANGE CHANGE

The Range Change tab compares two binary maps cell by cell. Select the current and future layers from the two drop-down lists — only binary (`_bin`) layers appear, because range change is a statement about presence and absence, not about continuous suitability — then press **Fill Paths** and **▶ Compute Range Change**.

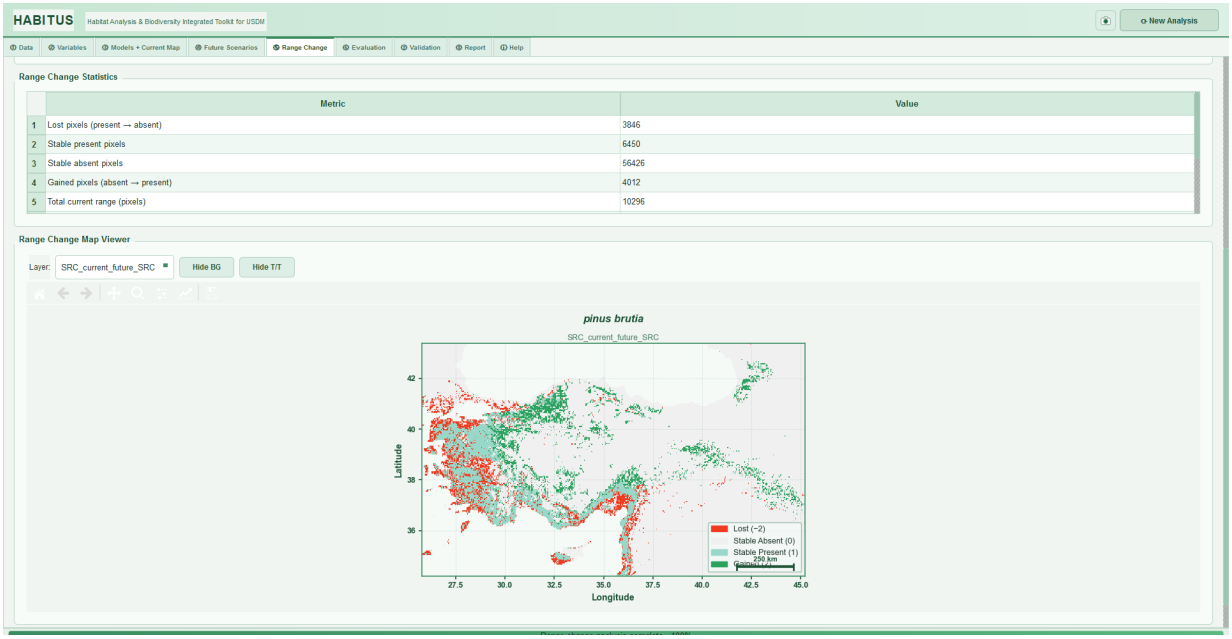


Figure 13. Range change between current climate and SSP5-8.5 (2081–2100). Red marks habitat lost, dark green habitat gained, and pale turquoise areas suitable in both periods.

The statistics table reports the pixel count of each class. For our example: 10,296 cells currently suitable, 10,462 projected, 6,450 stable, with 37.4% lost and 39.0% gained.

> **Note.** Read those two numbers together. The *net* change is +1.6%, which by itself suggests the species is barely affected. The *turnover* is nearly 40% — more than a third of the suitable area moves. Reporting only the net figure would be misleading, and the spatial pattern in Figure 13 is what carries the conservation message: losses in the lowlands, gains inland and upslope.

To convert pixel counts to area, multiply by the cell area of your layers. At 2.5 arc-minutes near 38° N a cell is roughly 21 km².

STATISTICAL ANALYSIS

The Evaluation tab holds seven sub-tabs of diagnostics. Press **↻ Refresh All Charts** once; everything is computed from the held-out test data.

Model Scores compares AUC, TSS and the Boyce index across all algorithms in one chart. **ROC Curves** shows discrimination in detail:

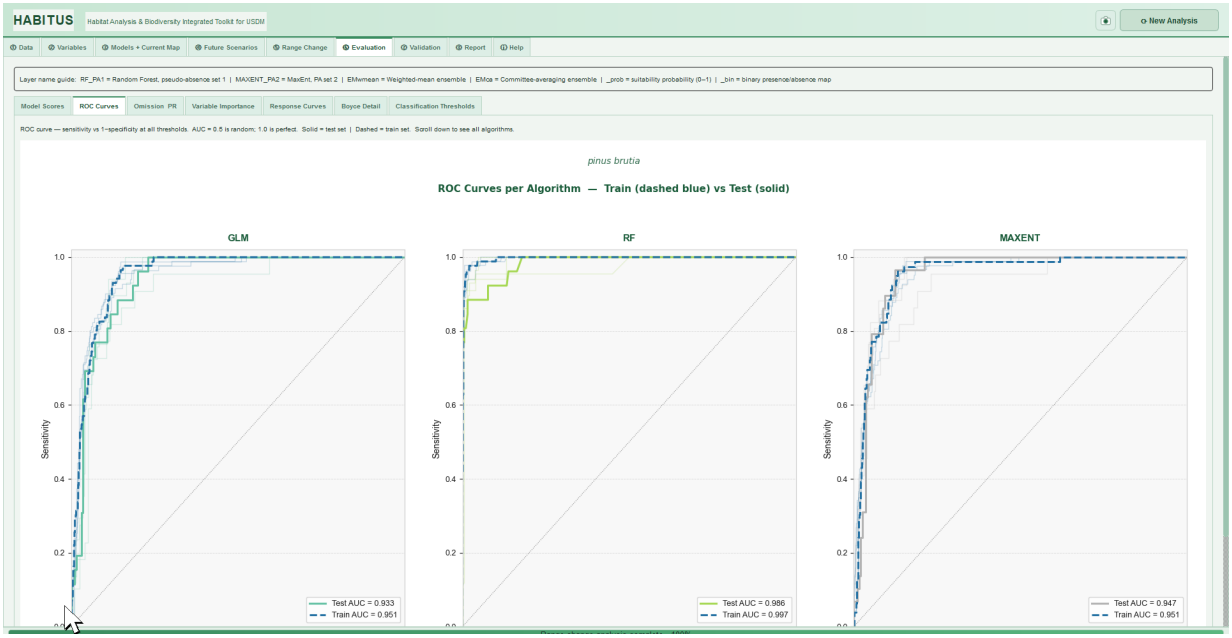


Figure 14. ROC curves per algorithm, with training (dashed) and test (solid) curves overlaid. The closer the solid curve runs to the top-left corner the better; the diagonal is random prediction. A large gap between dashed and solid indicates overfitting.

Three measures are reported, and they answer different questions. **AUC** is the probability that a randomly chosen presence is ranked above a randomly chosen background point — a measure of ranking, insensitive to threshold. **TSS** (sensitivity + specificity − 1) is threshold-dependent but, unlike kappa, independent of prevalence (Allouche et al., 2006). The **continuous Boyce index** needs only presence data and is therefore the appropriate measure for presence-only models, where AUC is hard to interpret (Boyce et al., 2002; Hirzel et al., 2006).

> **Note.** Do not report a single Boyce value from one cross-validation fold as if it were a property of the model. In our example AUC varied by 0.054 across the eighteen model realisations while the Boyce index ranged from 0.261 to 0.916 — a spread wider than the difference between a poor model and an excellent one. HABITUS shows every realisation precisely so that this variability is visible; report the range, not one number.

WHICH VARIABLES MATTER

The **Variable Importance** sub-tab estimates importance by permutation: the values of one variable are shuffled and the drop in model performance is measured. A variable the model relies on heavily produces a large drop.

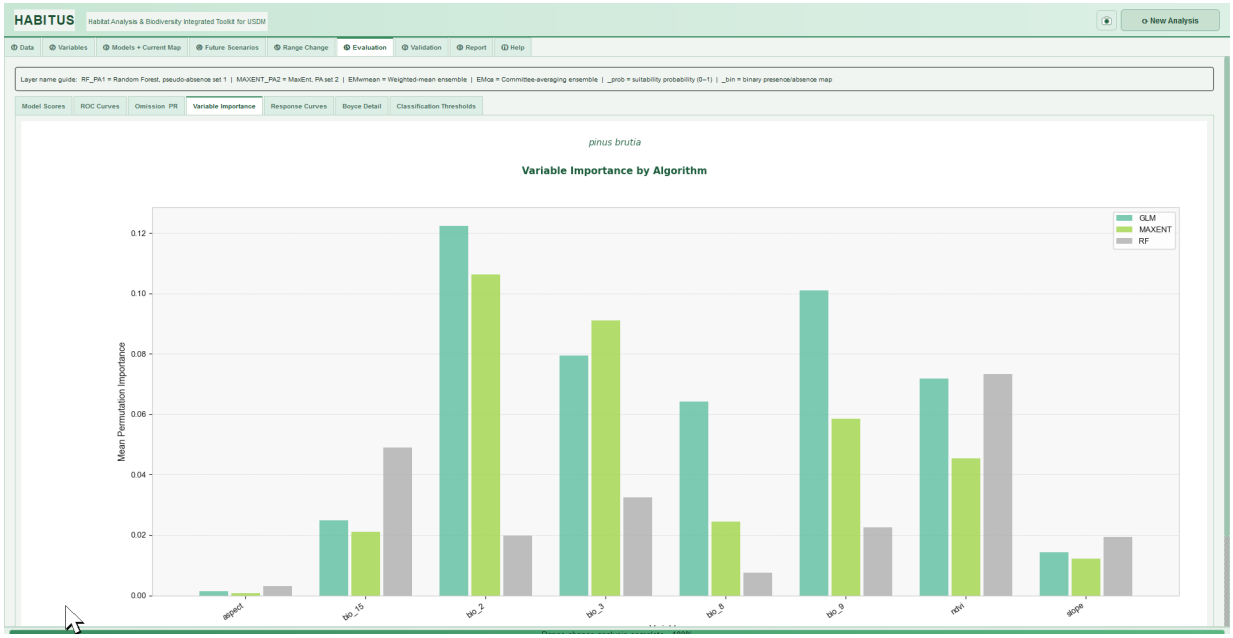


Figure 15. Permutation importance, shown separately for each algorithm. Agreement across algorithms — as seen here for bio_2 and bio_9 — is more convincing than a high score in any single model.

Treat importance rankings with care when predictors are correlated. If two variables carry the same information, permuting one of them may barely affect performance because the other compensates, and both will appear unimportant. This is exactly why the screening in step comes first.

HOW THE PREDICTION DEPENDS ON THE VARIABLES

Importance tells you *how much* a variable matters; response curves tell you *in which direction*. The **Response Curves** sub-tab offers three methods, selectable from the drop-down:

Marginal varies one variable while holding the others at their median — fast, and the easiest to read. **Partial dependence** averages over the observed distribution of the other variables (Friedman, 2001), which is more rigorous when variables interact. **ICE** draws one line per observation rather than an average (Goldstein et al., 2015), revealing cases where the mean response hides opposite behaviour in different parts of the data.

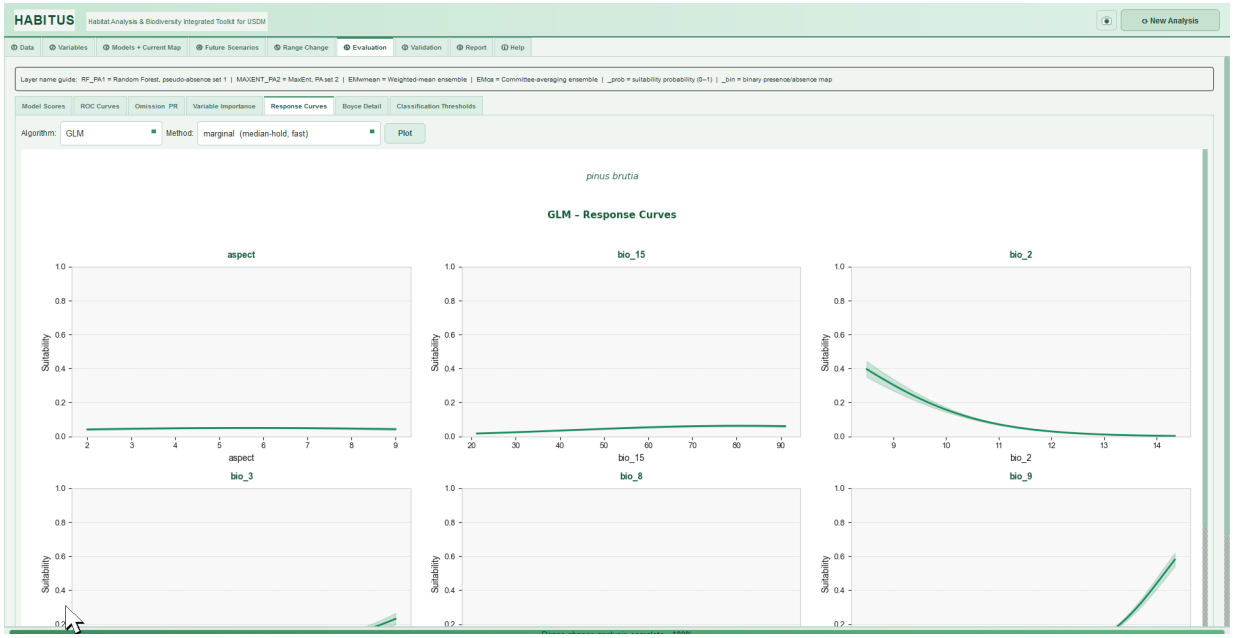
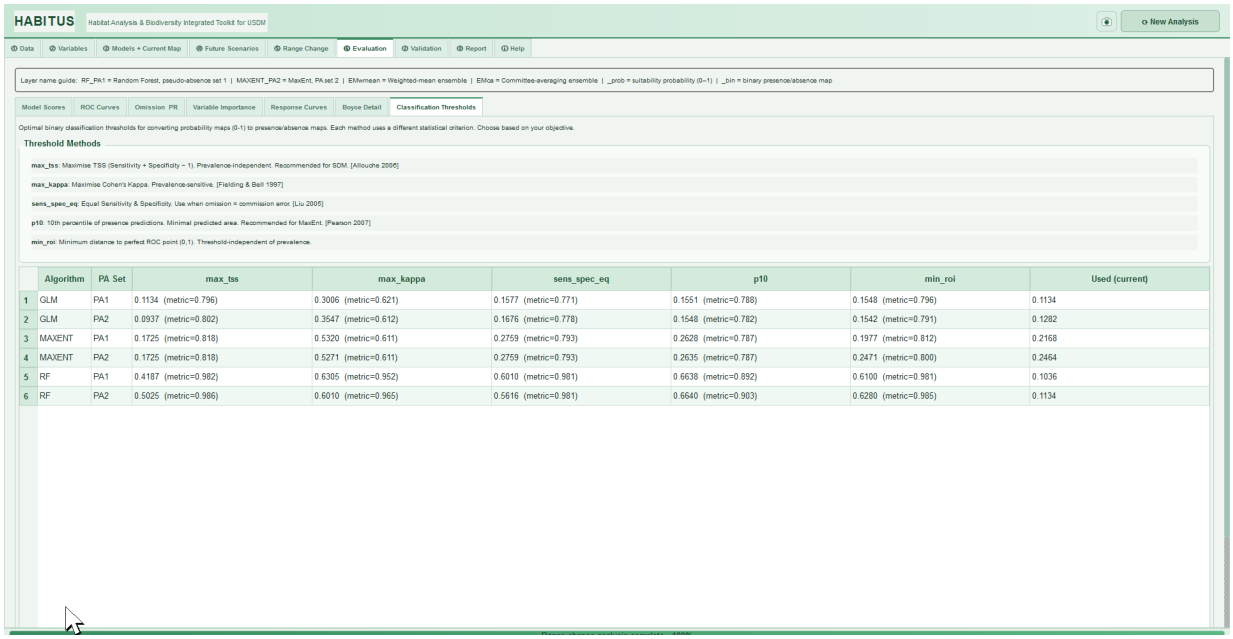


Figure 16. Response curves. Ecologically implausible shapes — for example suitability rising without limit at extreme values — are usually a sign of extrapolation beyond the sampled range rather than a biological finding.

THRESHOLDS Turning a continuous suitability surface into a presence/absence map requires a threshold, and the choice materially changes the mapped area. The **Classification Thresholds** sub-tab computes five rules for every model:



	Algorithm	PA Set	max_tss	max_kappa	sens_spec_eq	p10	min_roi	Used (current)
1	GLM	PA1	0.1134 (metric=0.796)	0.3006 (metric=0.621)	0.1577 (metric=0.771)	0.1551 (metric=0.788)	0.1548 (metric=0.796)	0.1134
2	GLM	PA2	0.0937 (metric=0.802)	0.3547 (metric=0.612)	0.1676 (metric=0.778)	0.1548 (metric=0.782)	0.1542 (metric=0.791)	0.1282
3	MAXENT	PA1	0.1725 (metric=0.818)	0.5320 (metric=0.611)	0.2759 (metric=0.793)	0.2628 (metric=0.787)	0.1977 (metric=0.812)	0.2168
4	MAXENT	PA2	0.1725 (metric=0.818)	0.5271 (metric=0.611)	0.2759 (metric=0.793)	0.2635 (metric=0.787)	0.2471 (metric=0.800)	0.2464
5	RF	PA1	0.4187 (metric=0.982)	0.6305 (metric=0.952)	0.6010 (metric=0.981)	0.6638 (metric=0.892)	0.6100 (metric=0.981)	0.1036
6	RF	PA2	0.5025 (metric=0.986)	0.6010 (metric=0.965)	0.5616 (metric=0.981)	0.6640 (metric=0.903)	0.6280 (metric=0.985)	0.1134

Figure 17. The five threshold rules computed per algorithm, with the resulting sensitivity and specificity.

max_TSS maximises sensitivity + specificity − 1 and is the standard default in SDM. **max_Kappa** maximises Cohen's kappa. **sens_spec_eq** sets sensitivity equal to specificity, appropriate when omission and commission errors are equally costly (Liu et al., 2005). **P10** uses the tenth percentile of predicted values at presence sites, accepting that the 10% of records with the lowest predictions may be erroneous — common practice with MaxEnt and small samples (Pearson et al., 2007). **min_roi** takes the point closest to the perfect ROC corner.

> **Note.** Whichever rule you use, state it in your methods and report the threshold value. Two studies of the same species can disagree on range size purely because one used P10 and the other max_TSS.

INDEPENDENT VALIDATION

Cross-validation measures internal consistency: how well the model predicts records drawn from the same sample it was fitted on. It cannot tell you whether the map agrees with an independent description of the species' range. The Validation tab is for that.

Press **Load HABITUS Output Layers...** to populate the two drop-downs, then set the **Reference Map** to your external source and the **Validation Map** to one of your own outputs. Here the reference is the EUFORGEN chorological map of *Pinus brutia* — expert-delineated, and never seen by the model.

> **Note.** Both drop-downs are filled with the same list, so it is entirely possible to select the same layer twice. If you do, you will get an overall accuracy of 1.000 and a kappa of 1.000, which means nothing at all. Check that the two fields differ before you run.

Because the EUFORGEN map is already a 0/1 surface, tick **Use binary comparison**. For a continuous reference, leave it unticked and set the reclassification thresholds instead. Choose a sampling design — random, stratified, systematic, or a CSV of your own points — and a number of points, then press **Run Validation**.

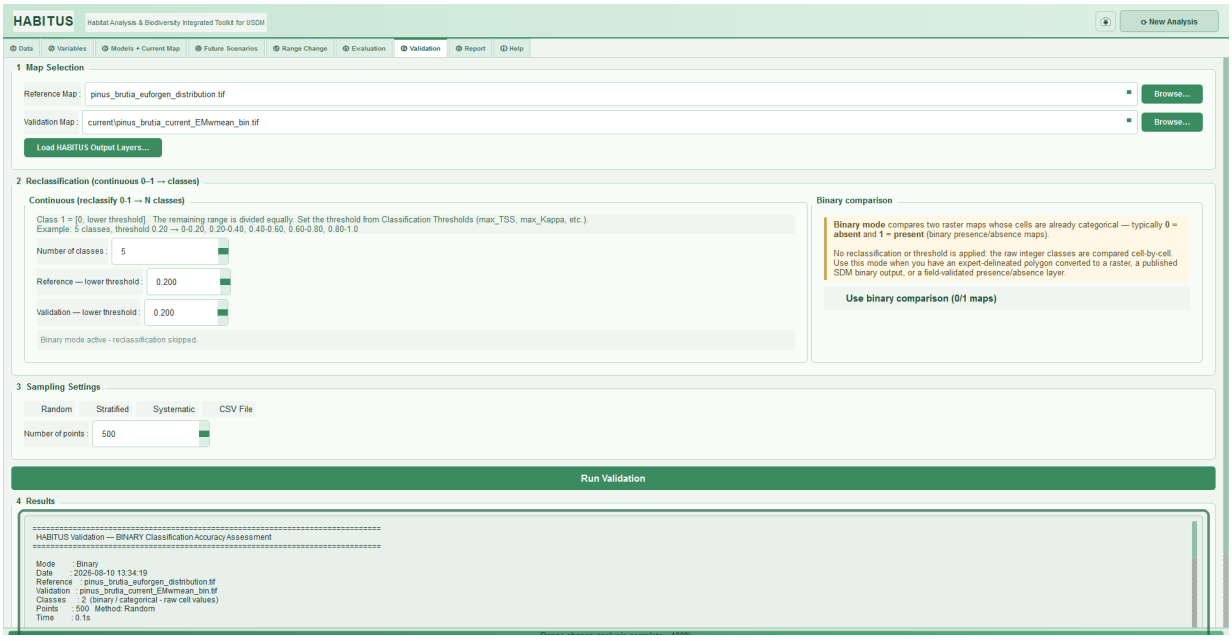


Figure 18. Validation output: a confusion matrix with overall accuracy, the F1 score and Cohen's kappa. Our ensemble achieved OA = 0.898 and κ = 0.648 against EUFORGEN at 500 random points.

Kappa is the figure to emphasise. Overall accuracy is inflated whenever one class dominates — and in most distribution maps absence dominates heavily — whereas kappa discounts the agreement expected by chance (Cohen, 1960; Fielding & Bell, 1997). An internal AUC above 0.9 together with an external kappa of 0.65 is not a contradiction: the two statistics answer different questions.

THE REPORT

The Report tab assembles everything into a single HTML document with ten sections: summary, variable selection, methods, model parameters, evaluation results, current distribution, future projections, range change, validation and the session log. Untick any section you do not want, then press **Generate HTML Report**.

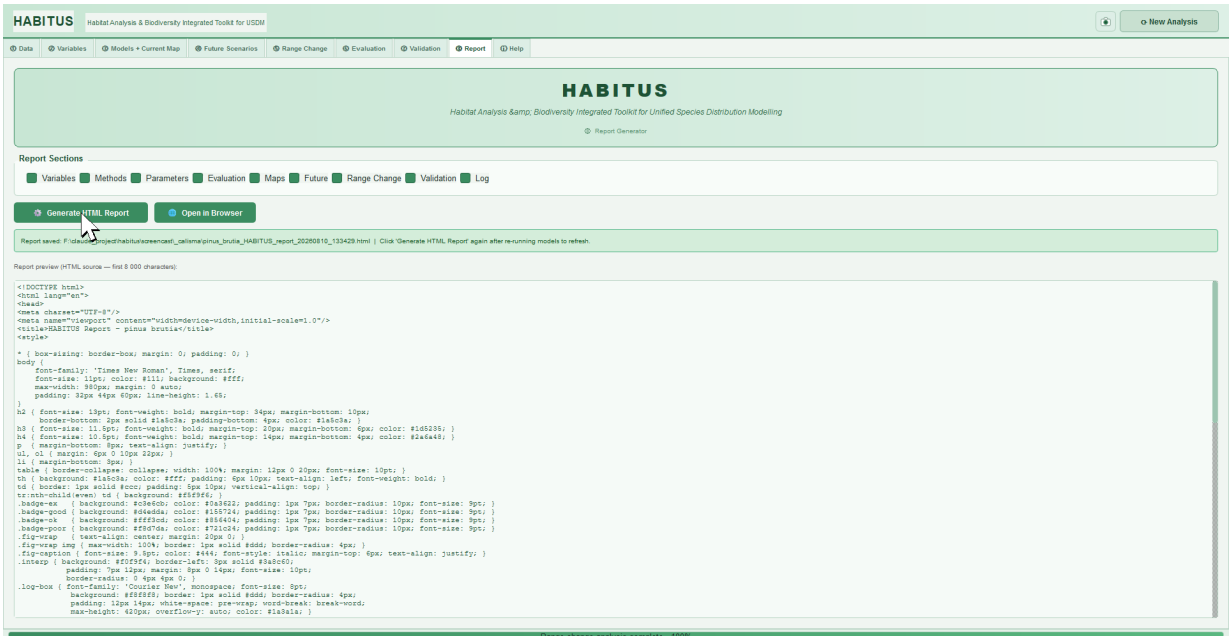


Figure 19. The generated report. Every metric is accompanied by a literature-based interpretation, and all figures are embedded as base64 data so the report is a single self-contained file.

The practical value is reproducibility. The parameter settings, thresholds, sample sizes and diagnostics that a reviewer will ask for under a protocol such as ODMAP (Zurell et al., 2020) are recorded automatically, at the moment the analysis ran, rather than reconstructed from memory months later. Open the file in a browser and print to PDF for a permanent record; the embedded figures are at 300 dpi and can be lifted directly into a manuscript.

OUTPUT FILES

Everything is written to the output folder you chose in step :

```
current/        GeoTIFFs for current climate (_prob and _bin per model)
<scenario>/     one folder per future projection
figures/        evaluation charts at 300 dpi
csv/            scores, thresholds, VIF ranking, correlated pairs
validation/     reclassified reference and validation rasters
occurrence_train.csv, occurrence_test.csv   the exact split used
*.log           timestamped log with per-algorithm timings
*_report.html   the self-contained report
```

The train/test CSV files deserve a mention: they record precisely which records were used for fitting and which were held out, so a reviewer can reproduce your evaluation exactly.

COMMON PITFALLS

Layer mismatch when projecting. The commonest error in applied SDM is projecting a model onto a layer stack that differs in composition or order from the training set. HABITUS refuses to run in that case; if you see the “Raster count mismatch” message, compare the list it prints against the variables you confirmed in step .

Reporting net range change alone. As shown above, a net change near zero can conceal turnover approaching 40%. Report lost, gained and stable separately.

Validating against your own output. Comparing a map with itself gives perfect agreement and no information.

Trusting a single fold. Especially for the Boyce index at small sample sizes, report the spread across replicates rather than one value.

Ignoring what the model cannot know. These are correlative models of the realised niche. They assume niche conservatism and take no account of dispersal ability, biotic interactions, soil or disturbance regimes such as fire. A projection showing newly suitable area 80 years hence does not mean the species will reach it.

TERMS OF USE

HABITUS is free and open source for academic use under the MIT licence. Installers for Windows, macOS and Linux, the source code and the *Pinus brutia* example dataset used throughout this tutorial are available at:

```
https://github.com/omerorucu/habitus
```

If HABITUS contributes to a publication, please cite:

> Örücü, Ö. K., & Örücü, S. (2026). HABITUS: Habitat Analysis and Biodiversity Integrated Toolkit for Unified Species Distribution Modelling. *Ecological Perspective*, [Technical Report]. https://doi.org/10.53463/ecopers.20260435

Questions, bug reports and feature requests are welcome through the issue tracker on the repository.

REFERENCES

Allouche, O., Tsoar, A., & Kadmon, R. (2006). Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). Journal of Applied Ecology, 43(6), 1223–1232.

Araújo, M. B., & New, M. (2007). Ensemble forecasting of species distributions. Trends in Ecology & Evolution, 22(1), 42–47.

Barbet-Massin, M., Jiguet, F., Albert, C. H., & Thuiller, W. (2012). Selecting pseudo-absences for species distribution models: how, where and how many? Methods in Ecology and Evolution, 3(2), 327–338.

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. New York: Wiley.

Boyce, M. S., Vernier, P. R., Nielsen, S. E., & Schmiegelow, F. K. A. (2002). Evaluating resource selection functions. Ecological Modelling, 157(2–3), 281–300.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1), 37–46.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., … Lautenbach, S. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography, 36(1), 27–46.

Elith, J., & Leathwick, J. R. (2009). Species distribution models: ecological explanation and prediction across space and time. Annual Review of Ecology, Evolution, and Systematics, 40, 677–697.

Fielding, A. H., & Bell, J. F. (1997). A review of methods for the assessment of prediction errors in conservation presence/absence models. Environmental Conservation, 24(1), 38–49.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. The Annals of Statistics, 29(5), 1189–1232.

Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation. Journal of Computational and Graphical Statistics, 24(1), 44–65.

Guisan, A., & Thuiller, W. (2005). Predicting species distribution: offering more than simple habitat models. Ecology Letters, 8(9), 993–1009.

Hirzel, A. H., Le Lay, G., Helfer, V., Randin, C., & Guisan, A. (2006). Evaluating the ability of habitat suitability models to predict species presences. Ecological Modelling, 199(2), 142–152.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. Technometrics, 12(1), 55–67.

Liu, C., Berry, P. M., Dawson, T. P., & Pearson, R. G. (2005). Selecting thresholds of occurrence in the prediction of species distributions. Ecography, 28(3), 385–393.

O'Brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. Quality & Quantity, 41(5), 673–690.

Pearson, R. G., Raxworthy, C. J., Nakamura, M., & Peterson, A. T. (2007). Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. Journal of Biogeography, 34(1), 102–117.

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

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society: Series B, 58(1), 267–288.

Zurell, D., Franklin, J., König, C., Bouchet, P. J., Dormann, C. F., Elith, J., ... Merow, C. (2020). A standard protocol for reporting species distribution models. Ecography, 43(9), 1261–1277.