

The introduction of the European *Caucasotachea vindobonensis* (Gastropoda: Helicidae) in North America, its origin and its potential range

Bernhard Hausdorf · Matt Parr · Laura J. Shappell · Jens Oldeland · David G. Robinson

Biological Invasions

Supplementary material 2 Details of Environmental Niche Modelling

Methods

We compiled European records of *C. vindobonensis* from Kajtoch et al. (2017), Korábek et al. (2020), [iNaturalist \(https://www.inaturalist.org/observations?place_id=any&taxon_id=922180, accessed on 11.11.2020\)](https://www.inaturalist.org/observations?place_id=any&taxon_id=922180), and the material in the Zoological Museum Hamburg for environmental niche modelling. The compiled occurrence data without duplicates (n = 1032 locations) was strongly clumped in geographic space, which could lead to biased results during the modelling procedure by induced spatial autocorrelation. Hence, we reduced the dataset by conducting a spatial thinning procedure. We used the spThin algorithm (Aiello-Lammens et al. 2015) to compare 100 repetitions of spatial subsamples and set the distance parameter to 10 km ensure large enough distances between occurrences. The algorithm tries to optimize the number of occurrences by comparing multiple reduced subsets of the dataset and keeping the distance threshold. The final solution contained 633 occurrence points (Electronic supplementary material 1) which is about 61% of the original data.

We used Maxent version 3.4.1 (Philips et al. 2004, 2020) for ecological niche modelling as implemented in the ‘kuenm’ workflow (Cobos et al. 2019). This workflow includes model calibration, model selection, and model evaluation of a pre-defined set of Maxent models. First, we split the occurrence dataset into calibration (n=445) and validation (n=188) data by a 70:30 ratio and added a random background sample of 10,000 pseudo absences. We defined 45 models based on the combination of 9 different values for the regularization parameter (0.1, 0.25, 0.5, 0.75, 1, 2, 3, 4, 5) and 5 different combinations of feature classes which Maxent uses to fit a model to the data, lqp, lq, q, qp, lp, with l = linear, q = quadratic, and p = product features. The 45 models were compared according to three independent evaluation criteria: (1) statistical significance according to partial ROC curves (Peterson et al. 2008), (2) omission rates at the 5% level (Anderson et al. 2003); and (3) model complexity compared with AICc (Burnham and Anderson 2002). The best model should have a statistically significant pROC, omission rates below 5% and models with fewer parameters were favoured over more complex models. For the best model, we repeatedly extrapolated 50 model outcomes using the Maxent internal bootstrap algorithm. Output was predicted on the logistic scale. The final model was then the median of the 50 replicates.

During the modelling workflow, we kept all bioclimatic variables as Maxent is robust against multicollinearity (Feng et al. 2019). We kept the correlation matrix for the interpretation of the best model to identify the interplay between correlated environmental drivers. We projected the best model describing the ecological niche of *C. vindobonensis* in Europe to North America. Model projections outside of the range from where the model was trained can be biased when the novel environment strongly differs from the calibration area. Hence, we calculated the mobility-oriented parity metric (Owens et al. 2013) for comparing the environmental conditions between the Europe and North America with the kuenm-package

(Cobos et al. 2019). The output map is scaled between 0 and 1 and represents the degree of similarity between conditions in Europe and North America. Values of zero equal to areas of strict extrapolation and should be avoided for interpretation. Occurrence and climatic raster data were handled in QGIS version 3.12 (QGIS Development Team 2019) and R version 3.6.1 (R Core Team 2019) using the packages raster (Hijmans 2020) and sf (Pebesma 2018).

Results

45 different models were built from which one best model was identified. The mean AUC ratio was 1.503, while the average training AUC out of 30 repeated models was 0.883. Both values indicating a satisfying model outcome. The omission rate of the best model was 6% which was sufficiently close to the envisaged 5% threshold. The regularization parameter of the best model was 2 while the selected features consisted of quadratic and product features only.

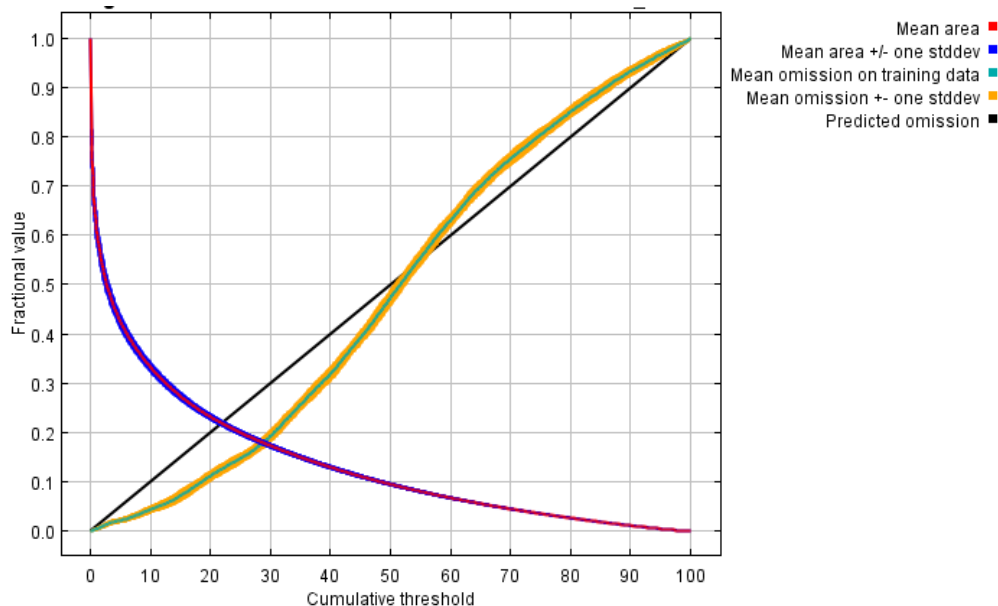
Replicated maxent model for *Caucasotachaea vindobonensis*

This page summarizes the results of 50 bootstrap models for *C. vindobonensis*, created Fri Nov 06 12:50:03 CET 2020 using Maxent version 3.4.1.

Analysis of omission/commission

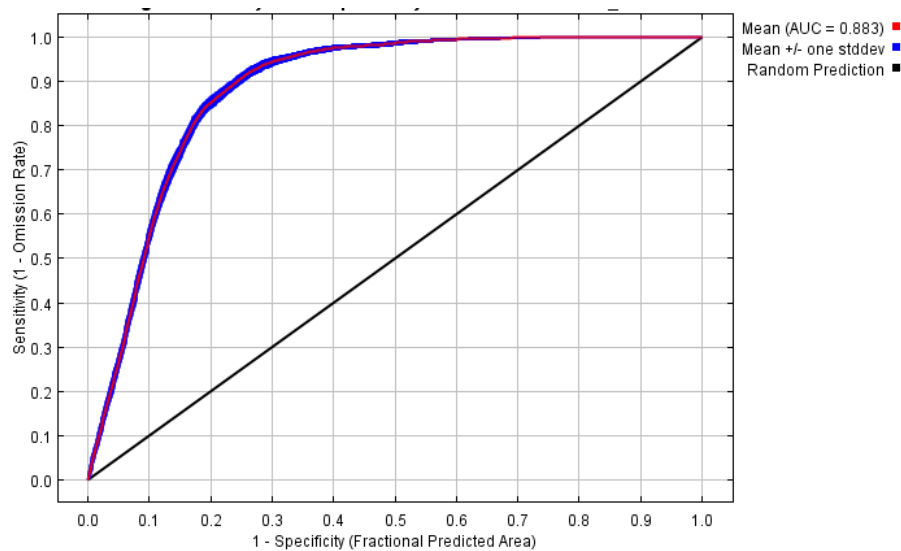
The following picture shows the training omission rate and predicted area as a function of the cumulative threshold, averaged over the replicate runs.

Average Omission and Predicted Area for *C. vindobonensis*



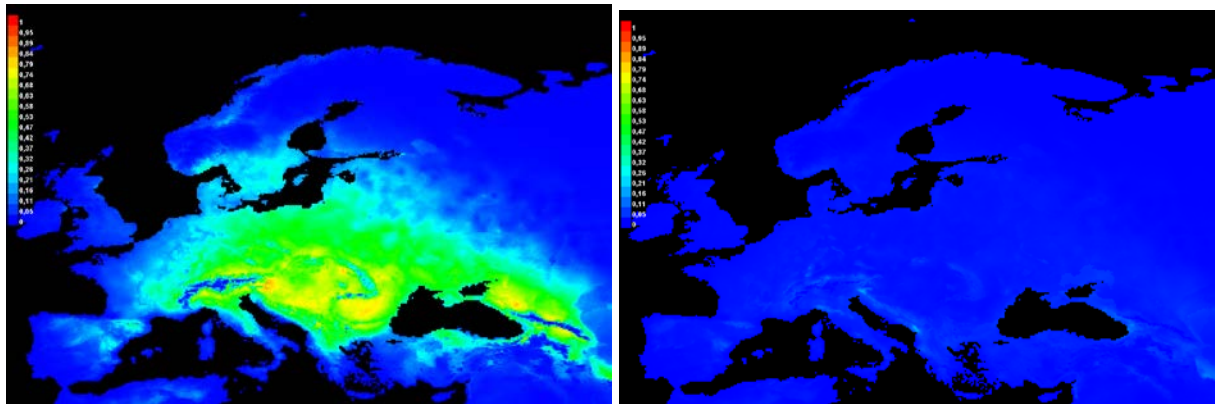
The next picture is the receiver operating characteristic (ROC) curve for the same data, again averaged over the replicate runs. Note that the specificity is defined using predicted area, rather than true commission (see the paper by Phillips, Anderson and Schapire cited on the help page for discussion of what this means). The average training AUC for the replicate runs is 0.883, and the standard deviation is 0.003.

Average Sensitivity vs 1- Specificity for *C. vindobonensis*

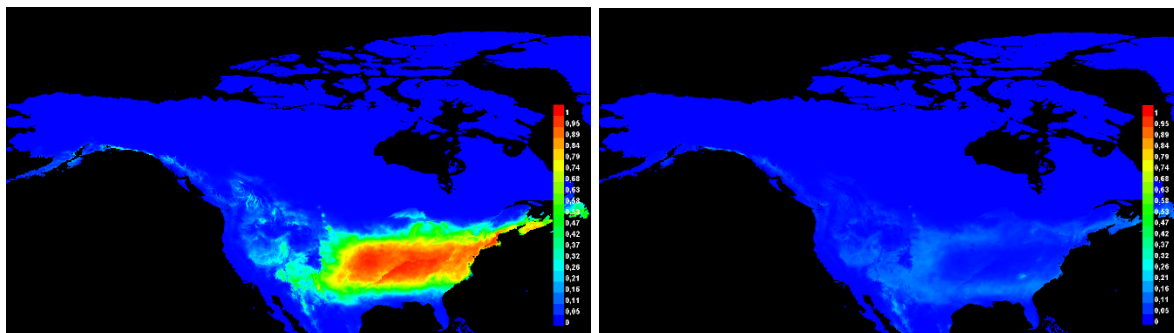


Pictures of the model

The following two pictures show the point-wise mean and standard deviation of the 50 output grids. Other available summary grids are min, max, median and 95% confidence level (lowerci).

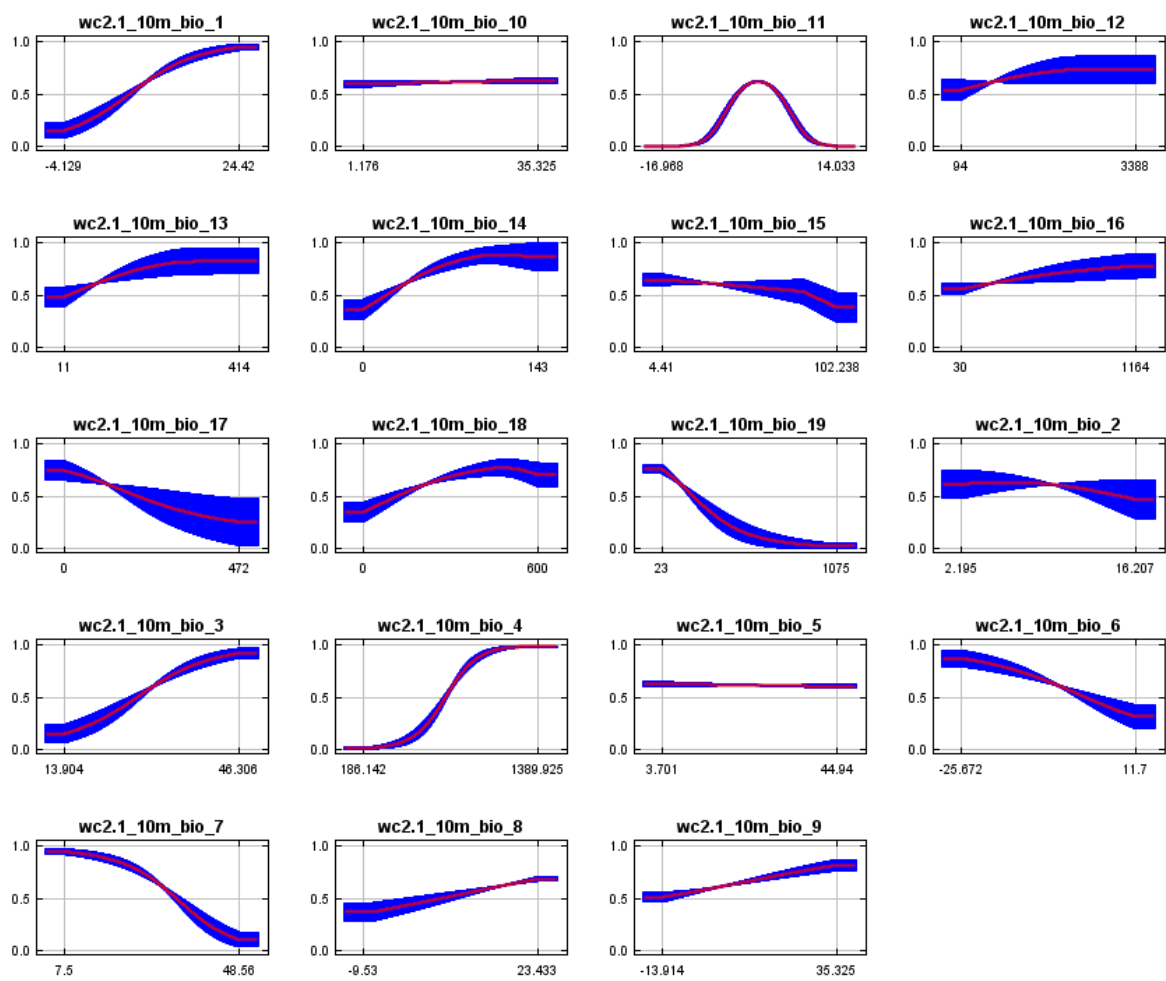


The following two pictures show the point-wise mean and standard deviation of the 50 models applied to the environmental layers in current. Other available summary grids are min, max, median and 95% confidence level (lowerci).

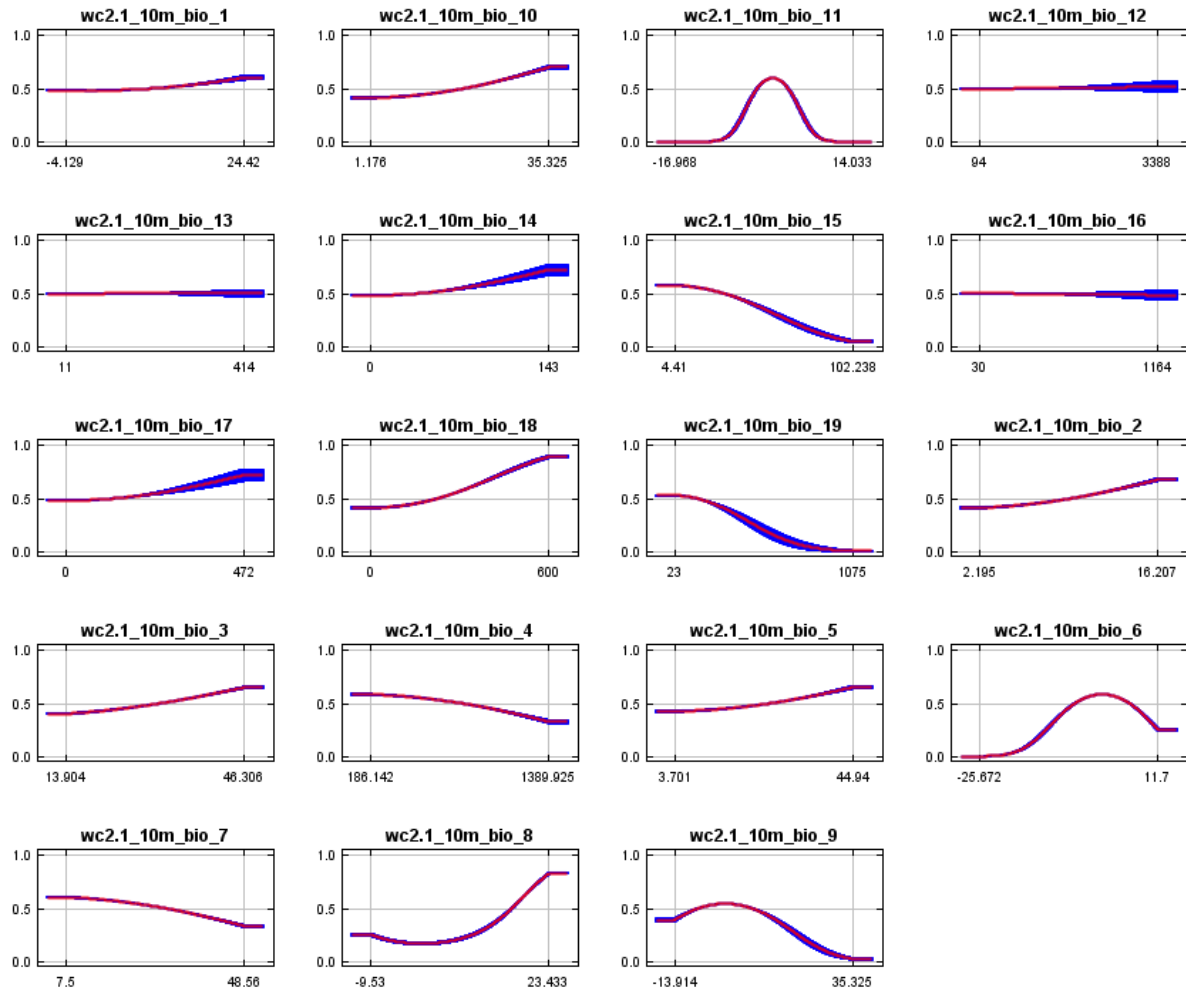


Response curves

These curves show how each environmental variable affects the Maxent prediction. The curves show how the predicted probability of presence changes as each environmental variable is varied, keeping all other environmental variables at their average sample value. Click on a response curve to see a larger version. Note that the curves can be hard to interpret if you have strongly correlated variables, as the model may depend on the correlations in ways that are not evident in the curves. In other words, the curves show the marginal effect of changing exactly one variable, whereas the model may take advantage of sets of variables changing together. The curves show the mean response of the 50 replicate Maxent runs (red) and the mean $\pm$ one standard deviation (blue, two shades for categorical variables).



In contrast to the above marginal response curves, each of the following curves represents a different model, namely, a Maxent model created using only the corresponding variable. These plots reflect the dependence of predicted suitability both on the selected variable and on dependencies induced by correlations between the selected variable and other variables. They may be easier to interpret if there are strong correlations between variables.

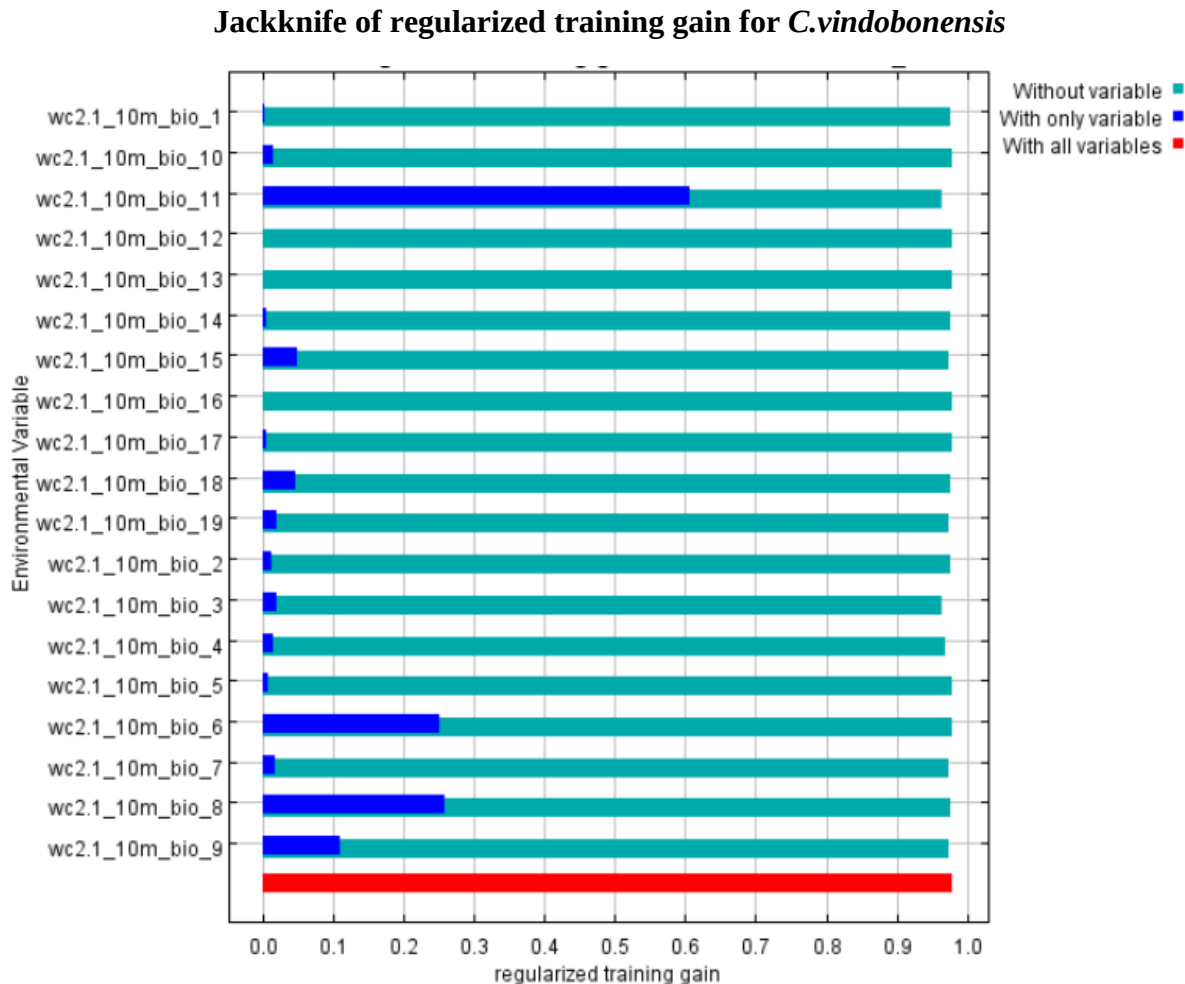


Analysis of variable contributions

The following table gives estimates of relative contributions of the environmental variables to the Maxent model. To determine the first estimate, in each iteration of the training algorithm, the increase in regularized gain is added to the contribution of the corresponding variable, or subtracted from it if the change to the absolute value of lambda is negative. For the second estimate, for each environmental variable in turn, the values of that variable on training presence and background data are randomly permuted. The model is reevaluated on the permuted data, and the resulting drop in training AUC is shown in the table, normalized to percentages. As with the variable jackknife, variable contributions should be interpreted with caution when the predictor variables are correlated. Values shown are averages over 50 replicate runs.

Variable	Percent contribution	Permutation importance
we2.1_10m_bio_11	58.6	27.5
we2.1_10m_bio_8	18.2	1
we2.1_10m_bio_18	7.5	2.7
we2.1_10m_bio_2	4.8	1.1
we2.1_10m_bio_3	1.9	8.8
we2.1_10m_bio_4	1.8	20.6
we2.1_10m_bio_14	1.1	3.8
we2.1_10m_bio_7	1	7.7
we2.1_10m_bio_1	1	8.8
we2.1_10m_bio_10	0.9	0
we2.1_10m_bio_17	0.7	2.2
we2.1_10m_bio_15	0.6	1.1
we2.1_10m_bio_5	0.5	0
we2.1_10m_bio_19	0.5	6.5
we2.1_10m_bio_9	0.3	2.2
we2.1_10m_bio_6	0.3	3.4
we2.1_10m_bio_13	0.2	1.1
we2.1_10m_bio_16	0.1	0.5
we2.1_10m_bio_12	0.1	0.9

The following picture shows the results of the jackknife test of variable importance. The environmental variable with highest gain when used in isolation is wc2.1_10m_bio_11, which therefore appears to have the most useful information by itself. The environmental variable that decreases the gain the most when it is omitted is wc2.1_10m_bio_11, which therefore appears to have the most information that isn't present in the other variables. Values shown are averages over 50 replicate runs.



Command line to repeat this species model:

```
java density.MaxEnt nowarnings noprefixes -E "" -E Caucastachaea_vindobonensis
responsecurves jackknife outputformat=logistic
outputdirectory=D:\kuenm\Caucastachaea\data\caucvini5\batch_final\M_2_F_qp_Set1_EC
projectionlayers=D:\kuenm\Caucastachaea\data\caucvini5\G_variables\Set1\current
samplesfile=D:\kuenm\Caucastachaea\data\caucvini5\occ_joint.csv
environmentallayers=D:\kuenm\Caucastachaea\data\caucvini5\M_variables\Set1
randomseed nowarnings noaskoverwrite betamultiplier=2.0 replicates=50
replicatetype=bootstrap nonlinear nohinge autorun novisible noautofeature threads=10
```

Additional references

- Aiello-Lammens ME, Boria RA, Radosavljevic A, Vilela B, Anderson RP (2015) spThin: An R package for spatial thinning of species occurrence records for use in ecological niche models. *Ecography* 38:541–545
- Anderson RP, Lew D, Peterson AT (2003) Evaluating predictive models of species' distributions: Criteria for selecting optimal models. *Ecol Modell* 162:211–232
- Burnham KP, Anderson DR (2002) Model selection and multimodel inference: A practical information-theoretic approach, 2nd edn. Springer, New York
- Feng X, Park DS, Liang Y, Pandey R, Papeş M (2019) Collinearity in ecological niche modeling: Confusions and challenges. *Ecol Evol* 9:10365–10376
- Hijmans RJ (2020) Package 'raster' version 3.1-5. Geographic Data Analysis and Modeling. <https://CRAN.R-project.org/package=raster>
- Owens HL, Campbell LP, Dornak LL, Saupe EE, Barve N, Soberón J, Ingenloff K, Lira-Noriega A, Hensz CM, Myers CE, Peterson AT (2013) Constraints on interpretation of ecological niche models by limited environmental ranges on calibration areas. *Ecol Modell* 263:10–18
- Pebesma E (2018) Simple Features for R: Standardized Support for Spatial Vector Data. *R Journal* 10:439–446
- Peterson AT, Papeş M, Soberón J (2008) Rethinking receiver operating characteristic analysis applications in ecological niche modeling. *Ecol Modell* 213:63–72
- Phillips SJ, Dudík M, Schapire RE (2004) A maximum entropy approach to species distribution modeling. *Proceedings of the twenty-first International Conference on Machine Learning*, p 83
- QGIS Development Team (2019) QGIS Geographic Information System (version 3.4). Open Source Geospatial Foundation. <http://qgis.osgeo.org>