Appendix S1 Evaluating the predicted extinction risk of living amphibian species with the fossil record

Supplementary figures, tables and modelling output

Melanie Tietje (Museum für Naturkunde, Leibniz Institute for Evolution and Biodiversity Science, Berlin, Germany)
Mark-Oliver Rödel (Museum für Naturkunde, Leibniz Institute for Evolution and Biodiversity Science, Germany; Berlin-Brandenburg Institute of Advanced Biodiversity Research (BBIB), Germany)

Contents

Intro

This document contains supplementary figures S1-20, tables S1-6, model output files and results from Kruskal-Wallis rank sum tests as well as Pairwise Wilcoxon rank sum tests. It is writen with the knitr package (Xie 2014, 2015, 2016) for the R statistical environment (R Core Team 2017). The entire analysis can be recreated by running knitr::knit() on the Markdown file *ELEtietjeSA1.Rmd*. The necessary files are available in the Git repository: https://github.com/Eryops1/supplement_amphibian_extinction_risk.

We used R version 3.4.3 (2017-11-30) and the following packages:

```
loadedNamespaces()
```

```
##  [1] "maps"         "ddalpha"      "tidyr"        "sfsmisc"
##  [5] "splines"      "foreach"      "gsubfn"       "prodlim"
##  [9] "dotCall64"    "assertthat"   "stats4"       "sp"
## [13] "grDevices"    "DRR"          "yaml"         "robustbase"
## [17] "ipred"        "pillar"       "backports"    "lattice"
## [21] "glue"         "base"         "digest"       "randomForest"
## [25] "colorspace"   "recipes"      "gbm"          "captioner"
## [29] "htmltools"    "Matrix"       "plyr"         "psych"
## [33] "timeDate"     "pkgconfig"    "CVST"         "broom"
## [37] "caret"        "purrr"        "scales"       "gower"
## [41] "lava"         "tibble"       "mgcv"         "datasets"
## [45] "ggplot2"      "withr"        "nnet"         "lazyeval"
## [49] "mnormt"       "proto"        "survival"     "magrittr"
## [53] "evaluate"     "methods"      "mice"         "nlme"
## [57] "MASS"         "dimRed"       "foreign"      "utils"
## [61] "class"        "tools"        "stringr"      "kernlab"
## [65] "munsell"      "bindrcpp"     "stats"        "compiler"
## [69] "RcppRoll"     "rlang"        "grid"         "simpleboot"
## [73] "iterators"    "graphics"     "spam"         "tcltk"
## [77] "rmarkdown"    "boot"         "gtable"       "ModelMetrics"
## [81] "codetools"    "reshape"      "reshape2"     "R6"
## [85] "lubridate"    "gridExtra"    "knitr"        "dplyr"
## [89] "rgdal"        "bindr"        "rprojroot"    "stringi"
## [93] "parallel"     "Rcpp"         "fields"       "rpart"
## [97] "tidyselect"   "DEoptimR"
```

Data collection

Extinct species data

Table S1: Variable description for extinct species.

Variable	Description
Duration	Duration of a species in the fossil record in million years. Time between the midranges of the oldest and youngest chronostratigraphic stage the species was observed, rounded to the next million years.
Abundance	Four categories for abundance were build based on the minimum number of individuals (MNI) and specimen counts for each species. MNI and specimen count data were obtained from the PBDB or literature. Maximum values were calculated per locality and stage. The final values for each species represent the maximum value ever shown by one species over space and time. These maximum values per species were clustered via k-means clustering into four numeric, ordinal categories (1, 2, 3 and 4). K-means clustering is a 1-n dimensional approach which tries to minimize the sum of squares within each cluster, that is the distance between the points and the center of each cluster. This is accomplished by repeatedly and randomly setting the position of the centers in the 2-dimensional space, searching for the optimal solution (Harting & Wong 1979). MNI and specimen counts are the two dimensions used in this clustering.
Geographic range	Geographic range size is calculatd as maximum great circle distance (shortest connection between two points on the surface of a sphere). Great circle distances were calculated for each species in each stage. The maximum values ever achieved by a species is used.
Latitudinal range	Maximum latitudinal range of a species (maximum difference between paleolatitudes). Ranges were calculated for each species in each stage. The maximum values ever achieved is used.
Mean latitude	Mean latitude calculated from all fossil occurrence coordinates of each species.
Minimum latitude	Minimum latitude calculated from all fossil occurrence coordinates of each species.
Body size	Maximum snout-vent-length (SVL) of the species. We collected SVL whenever possible, if not available we collected total length (TL) and absolute skull length (ASL). To estimate the SVL from TL and ASL, we created linear regressions that connect SVL, TL and ASL for species where all or at least two measurements were available. These linear models were used to calculate the SVL for species from TL or ASL. When measures on body size where not available, we used the bodysize of a congeneric species.

Stratigraphic resolution

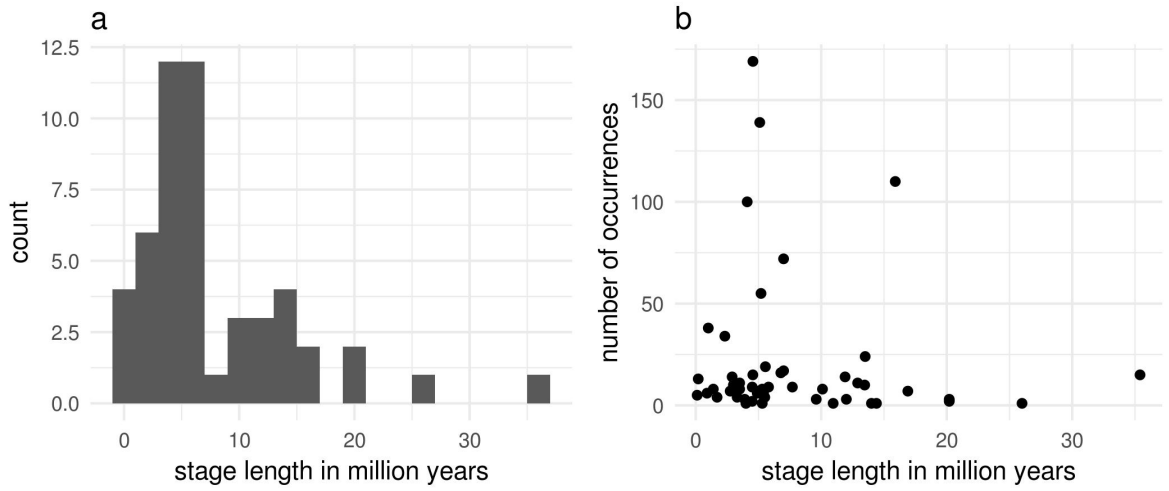

Figure S1: Stratigraphic resolution of the fossil data. Panel a) shows a histogram of the stage lengths (n=51). Panel b) shows the number of occurrences in each stage.

Extinct species data summary

Table S2: Summary statistics for the number of occurrences per species.

Variable	Value
Average number of occurrences per species	2.44
Minimum number of occurrences per species	1
Maximum number of occurrences per species	116

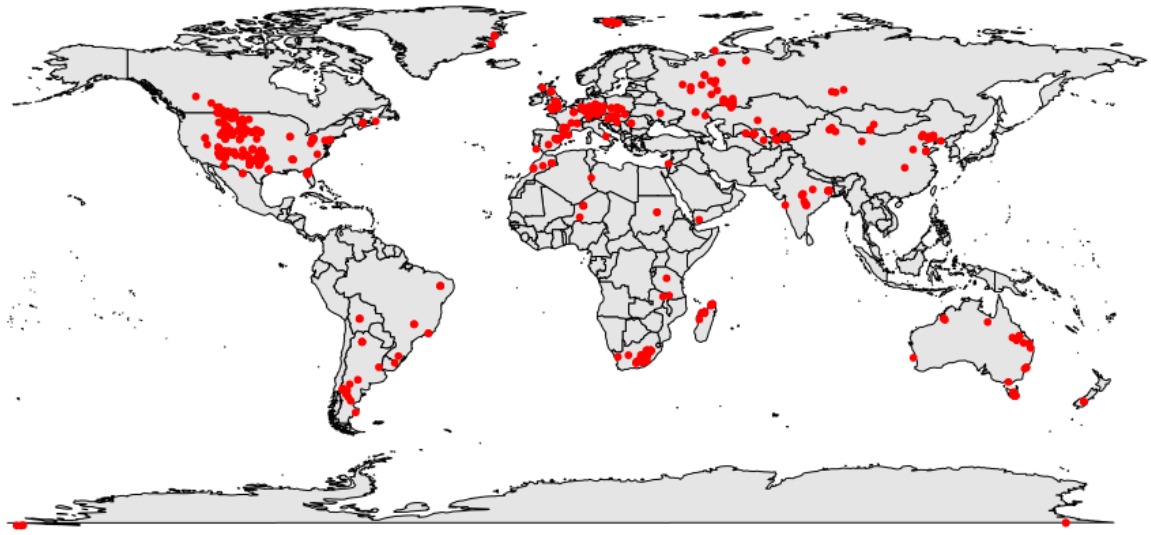

Figure S2: A map of all amphibian fossil occurrences used in the analysis.

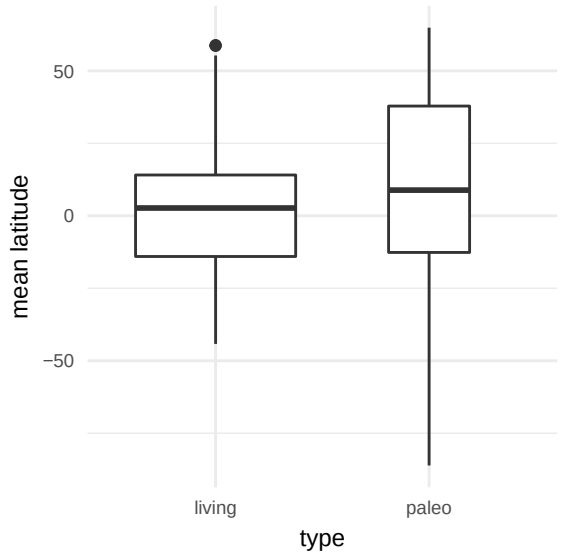

Figure S3: Boxplot showing the mean latitude of each living and fossil species. Boxplot width is scaled to sample size of each group (living: 1382, paleo: 354).

Living species data

Abundance data

Table S3: Keywords for text mining the abundance information of the IUCN Red List webpage for living amphibian species.

Category	Keywords	Conditioning
1	Not common, fewer than, rare, uncommon, small, small population, not abundant	
2	Fairly abundant, fairly common, moderately abundant	
3	Abundant, common, large, large populations	Has to be assigned first to allow for replacement in cases of "not common" or similar combinations
4	Very abundant, very common	

IUCN red list population description get scraped for keywords. All keywords were saved for each species. Keywords were categorized and species assigned according to their keywords. In cases of more than one possible assignment, the species gets assigned to the highest category to account for descriptions of species which are uncommon in xyz, but have large popuations in suitable habitats.

Data imputation

Multivariate imputation by chained equations creates multiple imputed datasets depending on different sampling sets of the other available variables, and takes the mean of these imputed datasets as finally imputed value. In multiple steps, plausible values are being drawn from a conditional density distribution that is modelled for each incomplete variable. Imputed variables subsequently enter the next imputation step. For full details on the algorithm behaviour see Van Buuren & Groothuis-Oudshoorn (2011). We chose this data imputation method because it allows for chosing different imputation models for each variable types, as the body size is a continuous numeric variable, the abundance however is implemented as a factor, coded as integer. The mice package allows for this distinction.

The goal of our data imputation was to be able to use the maximum number of occurrences while adding a minimum of noise in the data. The potential influences of data imputation is difficult to analyse, as the missing data cannot be recovered. Hunt (2017) however show, by randomly deleting variables from a complete dataset and completing it again with imputation, that using various multiple data imputation methods achieve around 80% or higher correct classifications in datasets with two or three different classes. This percentage was independant of the proportion of missing values in the data (10 to 50%). We thus assume our imputation method to be comparably effective.

Technical details

The R Code for the data imputation can be found in *Tietje_Roedel_2017_model_building_and_prediction.R*.

We used the mice::mice() function with settings maxit=20, m=50 and methods "rf" for body size and "polr" for abundance categories. For comparison with an alternative data imputation method, we are also imputed data using the rf::rfImpute() function. Following Van Buuren & Groothuis-Oudshoorn (2011) we ran diagnostic checks on the imputed data by checking for convergence of the imputing algorithms, plausibility of imputed data, and occurrences of impossible data. The results are shown in the following figures.

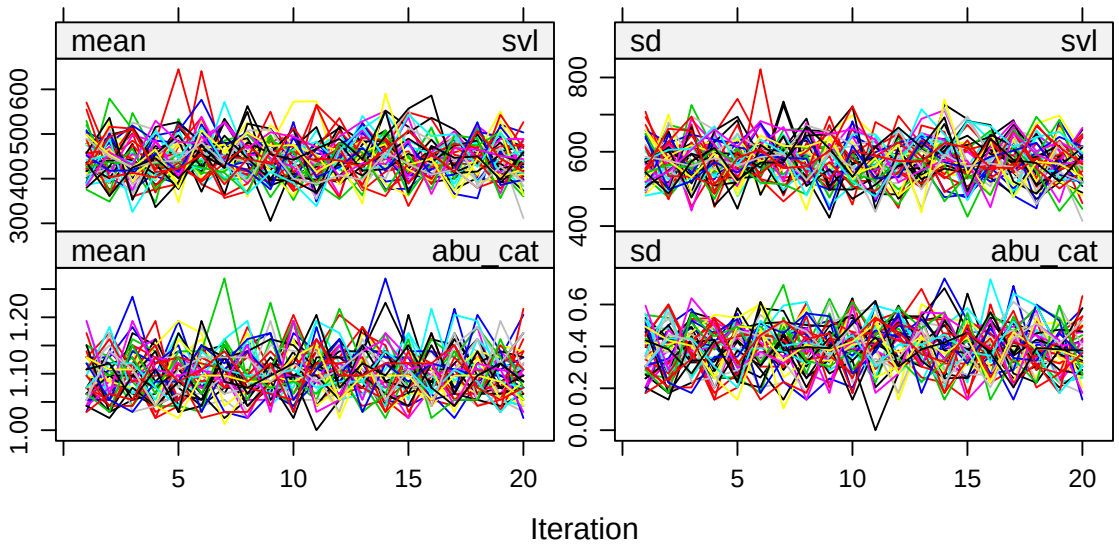

Figure S4: Convergence of the MICE algorithm for body size and abundance. Plotted are the mean and standard deviations of the imputed values per iteration. svl = body size, abu_cat = abundance. For variable explanation see Tab. S1.

Plausibility of the imputed values was checked by comparing density plots of imputed and observed values.

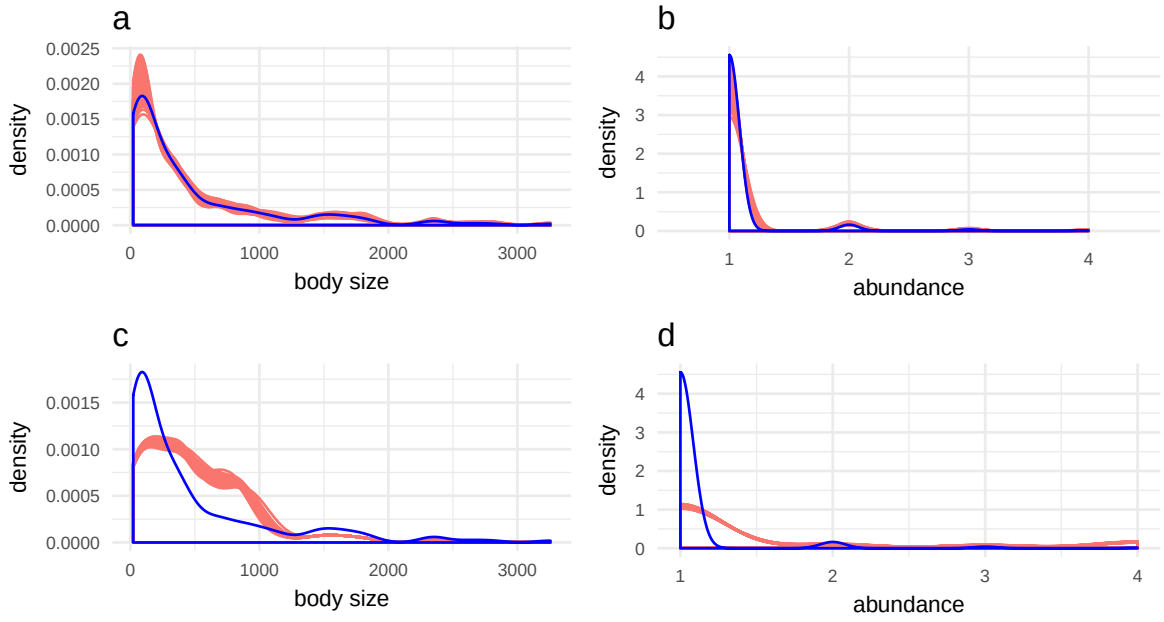

Figure S5: Kernel density estimates of the observed data (blue) and the m=50 densities per variable calculated from the imputed data (red lines) using the MICE algorithm (panel a) and b) and the randomForest algorithm (c and d).

Imputed and observed values for body size and abundance in each group are shown in the following figures.

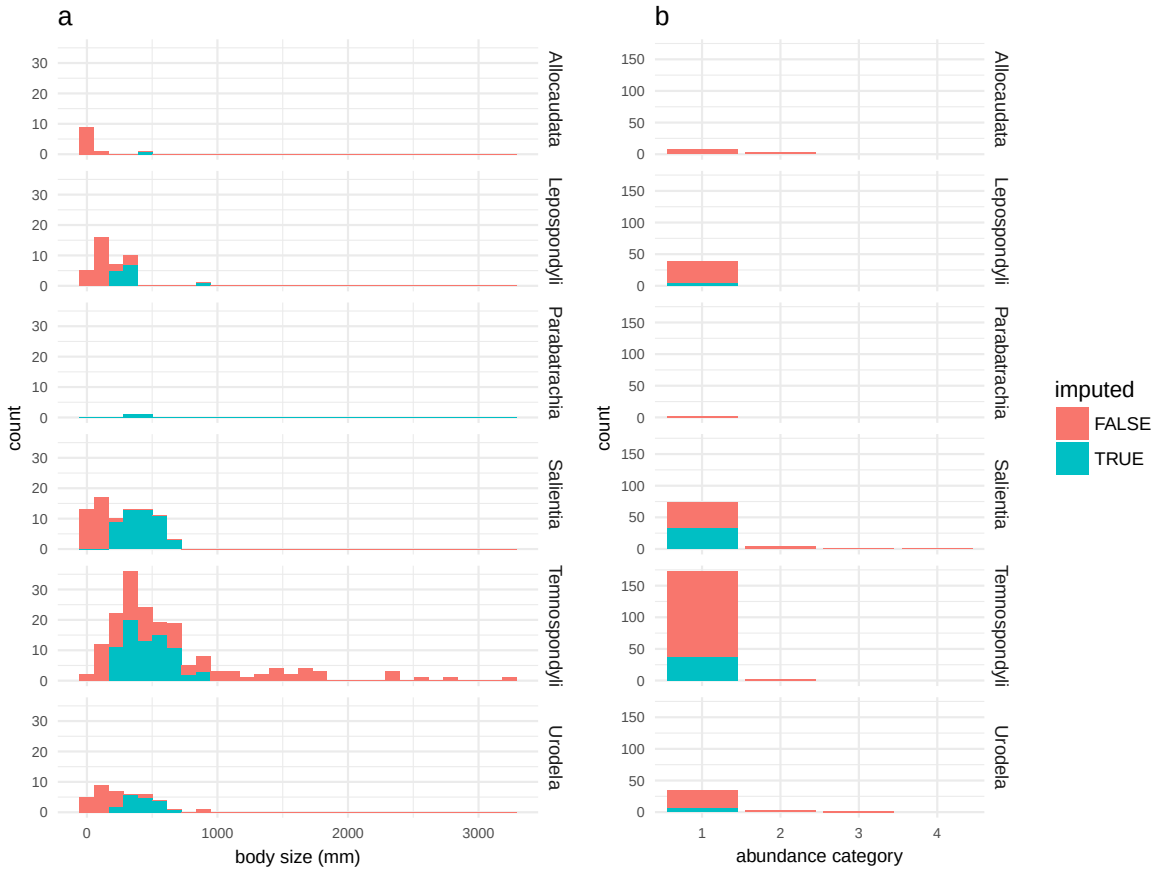

Figure S6: Imputated values in each taxonomic group. Blue color indicates the imputed values. Panel a) shows histograms of the body size in mm, panel b) shows the abundance category data.

Table S4: Percent imputed values in each taxonomic group and variable.

	body size	abundance
Allocaudata	0.09	0.09
Lepospondyli	0.33	0.13
Parabatrachia	1.00	0.00
Salientia	0.61	0.42
Temnospondyli	0.43	0.22
Urodela	0.46	0.18
Total	0.47	0.26

Imputation summary

Of all occurrences, 53% had some missing data, while only 19% of occurrences were missing both body size and abundance measure. Imputed values for the two incomplete variables body size and abundance made up 26% and 47% of those two variables, respectively. Those proportions varied between the taxonomic groups, with Allocaudata showing the highest and Salientia the lowest proportions of imputed data (Fig. S6, Tab. S4). The quality control showed that density plots of observed and imputed values were highly comparable (Fig.

S5 a and b). Further, we did not find any impossible values (like negative body size) in the imputed data. Therefore, the quality check suggests that the imputed data is likely reasonable (Van Buuren & Groothuis-Oudshoorn 2011).

The imputation increased the sample size (number of occurrences we were able to ues in the model) by 62%, which added to the stability of the model. Additionally, the imputed values were restricted to the variables abundance and body size, which were of lower importance for the final model. Therefore the influence of uncertainties introduced by data imputation on the outcome of the model should have been rather small. This was also reflected by the minor differences between the GBM fitted on imputed and unimputed data (Fig. S19, S20, Tab. S5).

Model selection

Due to the nature of the dataset (high skewness and kurtosis of the variables) we applied three different models to our data to connect traits with the duration of species: Generalized additive model (GAM), randomForest (rF), and Generalized boosted model (GBM).

The process of selecting the best parameter settings for the models as done in the caret package (Kuhn *et al.* 2017) is illustrated in the following Fig. S7:

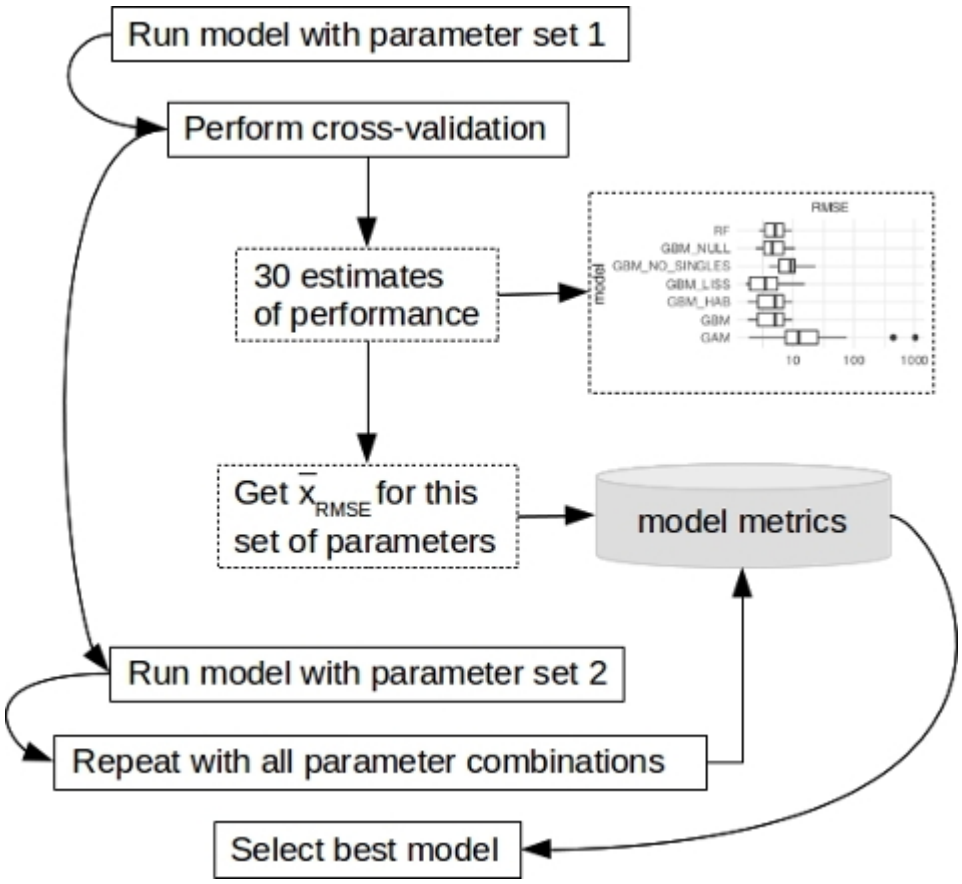

Figure S7: Model selection process, depicted as flowchart. Each model type (GAM, rF and GBM) was fitted using cross-validation. The model performance metrics were collected for each parameter set, finding the best parameter setting for the model.

Generalized additive model

Before adjusting the model parameters, we checked if logging the skewed variables in the extinct species data results in a better model performance (using default settings). While logging positively influenced the GAM fitted to the original dataset including missing values (which simply removes all cases with missing data), it only slighty influenced the GAM fitted on the imputed dataset. We therefore decided to stick to the unlogged, imputed dataset for further modelling.

A generalized additive model was fitted to the extinct species dataset using the caret::train() function. We used 3 separate 10-fold cross validations, meaning the extinct species-dataset was randomly split into 10 equal sized subsamples, from which 9 subsamples were used as

training data and one was retained as validation data for testing the model. This process was repeated until each of the subsamples was used once as validation set. The procedure was repeated 3 times.

Output 1: Console output fitting the gam using caret::train() function.

```
## Generalized Additive Model using Splines
##
## 354 samples
##   6 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 3 times)
## Summary of sample sizes: 319, 318, 319, 319, 319, 318, ...
## Resampling results across tuning parameters:
##
##   select  RMSE      Rsquared   MAE
##   FALSE   62.42695  0.2980428  12.817146
##    TRUE   21.95516  0.2718881   5.682457
##
## Tuning parameter 'method' was held constant at a value of GCV.Cp
## RMSE was used to select the optimal model using the smallest value.
## The final values used for the model were select = TRUE and method = GCV.Cp.
```

Output 2: Console output for the final gam.

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## .outcome ~ s(lat_range) + s(gcd) + s(min_lat) + s(mean_lat) +
##     s(svl)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.7994     0.1917   9.384   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                 edf Ref.df      F  p-value
## s(lat_range) 7.9139      9  8.412 6.88e-13 ***
## s(gcd)       7.7328      9 12.141  < 2e-16 ***
## s(min_lat)   7.7675      9 41.521  < 2e-16 ***
## s(mean_lat)  7.6842      9 41.645  < 2e-16 ***
## s(svl)       0.8119      9  0.124     0.24
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.662   Deviance explained = 69.3%
## GCV =  14.35  Scale est. = 13.016    n = 354
```

Random Forest

A Random Forest model was fitted to the extinct species dataset using the caret::train() function. We used 3 separate 10-fold cross validations, meaning the extinct species-dataset was randomly split into 10 equal sized subsamples, from which 9 subsamples were used as training data and one was retained as validation data for testing the model. This process was repeated until each of the subsamples was used once as validation set. The procedure was repeated 3 times.

Output 3: Console output fitting the rF using caret::train() function.

```
## Random Forest
##
## 354 samples
##   6 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 3 times)
## Summary of sample sizes: 319, 318, 319, 319, 319, 318, ...
## Resampling results across tuning parameters:
##
##   mtry  RMSE      Rsquared   MAE
##   2     5.198038  0.2492967  2.402612
##   3     5.260152  0.2498780  2.348284
##   4     5.315078  0.2470060  2.322943
##   5     5.366857  0.2398260  2.332688
##   6     5.393832  0.2365635  2.321819
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was mtry = 2.
```

Output 4: Console output for the final rF.

```
##
## Call:
##  randomForest(x = x, y = y, mtry = param$mtry, importance = TRUE,      verbose = FALSE)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 2
##
##           Mean of squared residuals: 33.52467
##                     % Var explained: 12.73
```

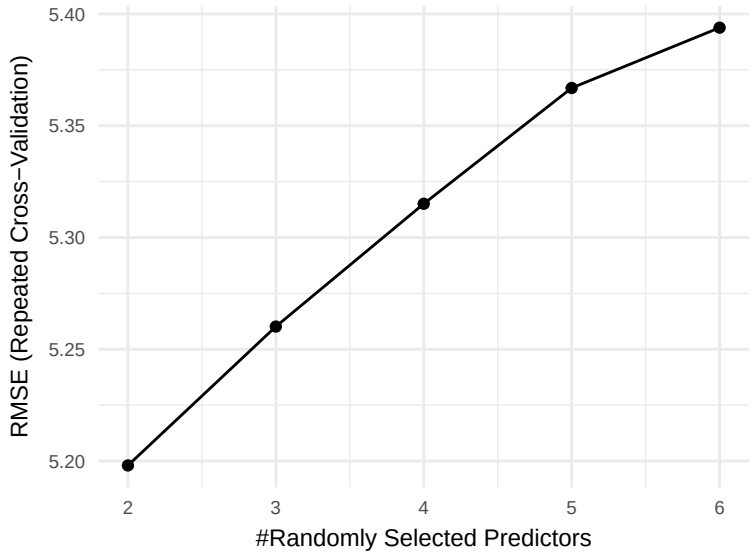

Figure S8: Root mean squared error (RMSE) for each cross validation set of different parameters mtry (number of predictor to chose from at each split in the random Forest).

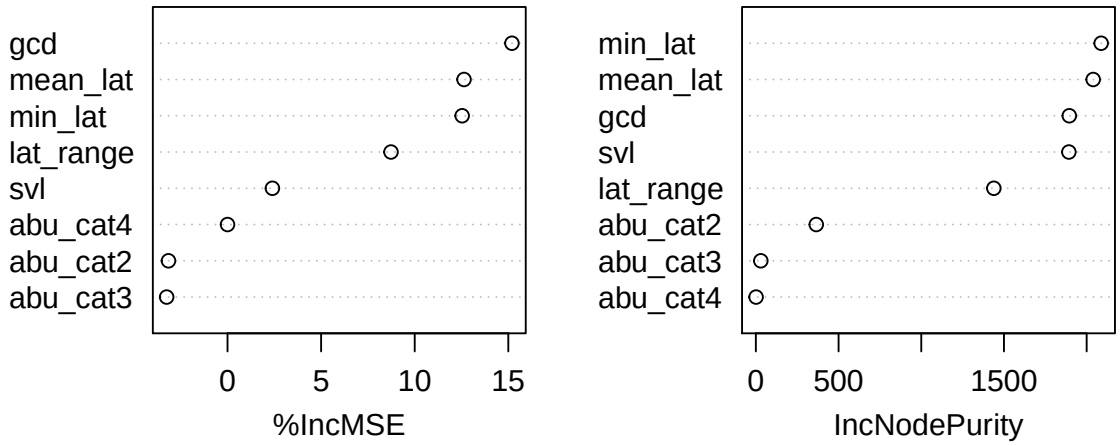

Figure S9: Variable importance measures for the final randomForest. Left plot: Difference in percent between the prediction error (mean squared error (MSE)) of the model and the prediction error after permuting each variable while holding all other data constant (%IncMSE). Differences are normalized by the standard deviation of the differences. Right plot: The total decrease in node impurity from splitting on the variable, averaged over all trees, measured by residual sum of squares (IncNodePurity). Variable names on the y-axis are the variables as used in the model and explained in Table S1.

Generalized boosted model

A generalized boosted model (GBM) was fitted to the extinct species dataset using the caret::train() function. We used the same cross-validation procedure to obtain the optimal tuning parameters as for the GAM and rF.

Output 5: Output fitting the GBM using caret::train() function. For full output run the R script.

```
## Stochastic Gradient Boosting
##
## 354 samples
##   6 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 3 times)
## Summary of sample sizes: 318, 319, 319, 319, 318, 319, ...
## Resampling results across tuning parameters:
##
##   shrinkage  interaction.depth  n.minobsinnode  n.trees  RMSE      Rsquared
##   0.001      1                   5               50      5.516714  0.2717127
##   0.001      1                   5              100      5.472123  0.2832056
##   0.001      1                   5              150      5.428993  0.2835112
##   0.001      1                   5              200      5.391019  0.2823545
##   0.001      1                   5              250      5.356412  0.2785768
##   0.001      1                   5              300      5.325467  0.2776956
##   0.001      1                   5              350      5.296146  0.2785732
##   MAE
##   3.052148
##   3.021109
##   2.990499
##   2.962987
##   2.936768
##   2.911829
##   2.887404
##  [ reached getOption("max.print") -- omitted 263 rows ]
##
## RMSE was used to select the optimal model using the smallest value.
## The final values used for the model were n.trees = 200, interaction.depth =
##  1, shrinkage = 0.01 and n.minobsinnode = 5.
```

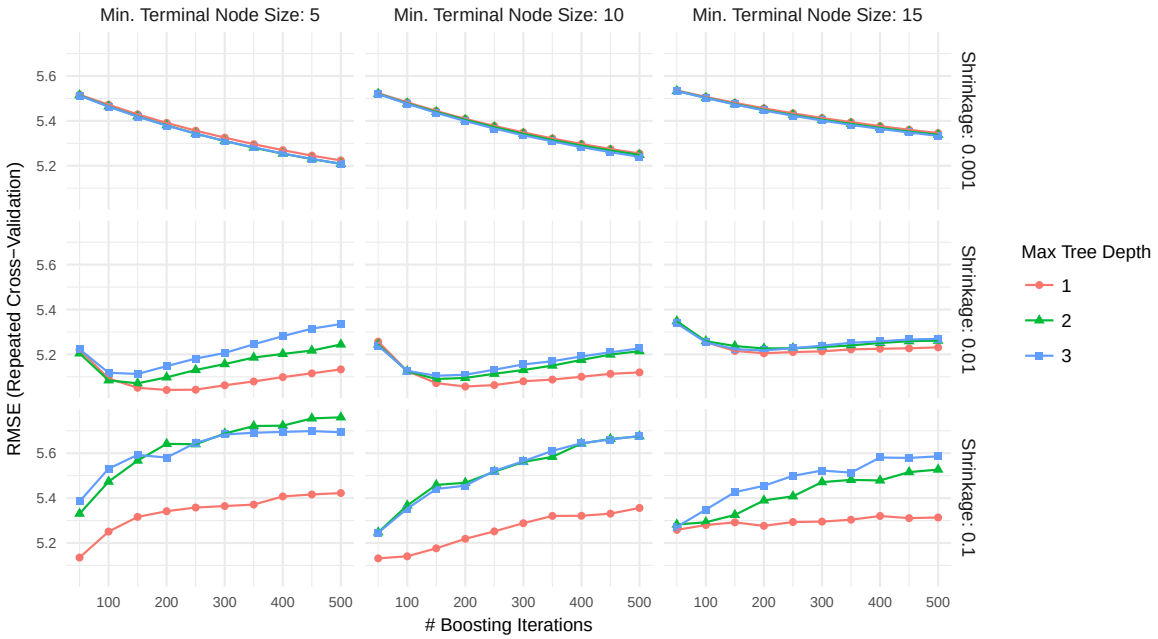

Figure S10: Relationship between the Root-mean-squared error (RMSE) of the GBM as estimate of performance and the tuning parameters: minimum terminal node size (vertical panels), shrinkage (horizontial panels) and maximum tree depth (legend). Minimum terminal node size defines the minimum number of observations in the trees terminal nodes, shrinkage is the learning rate of the model, maximum tree depth is the interaction depth for variables, with 1 being an additive model.

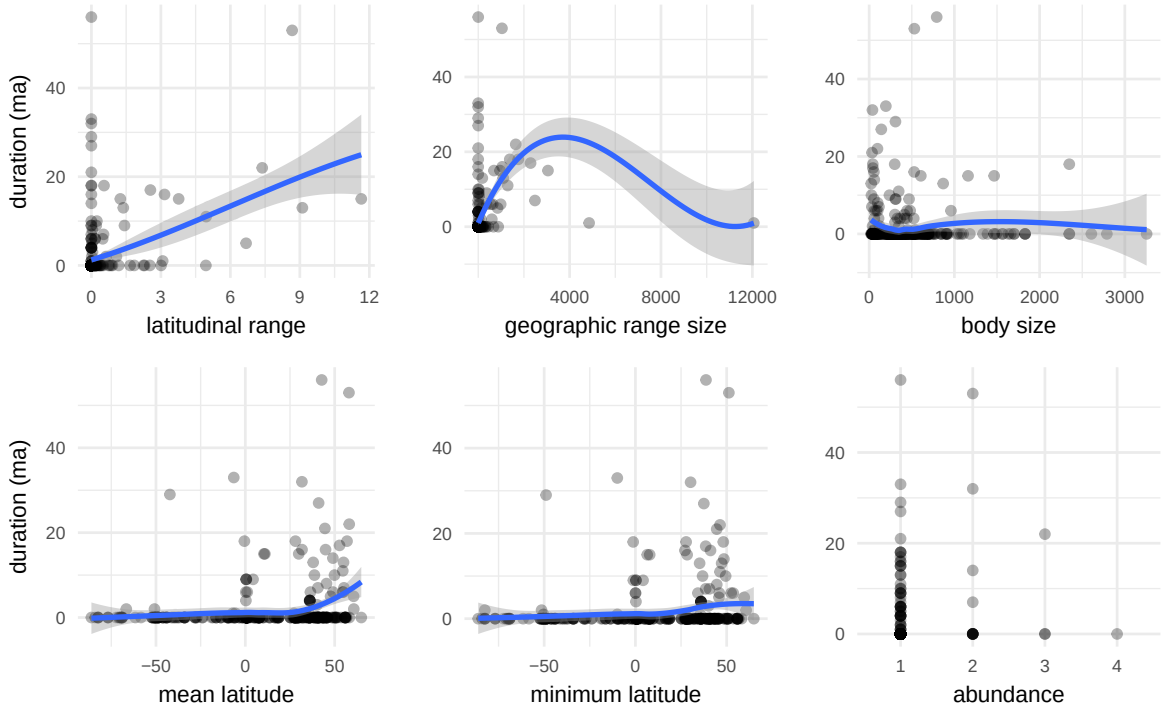

Figure S12: Scatterplots of each prediction variable with the response variable for the fossil species data. Variables are described in detail in Table S1.

Bias

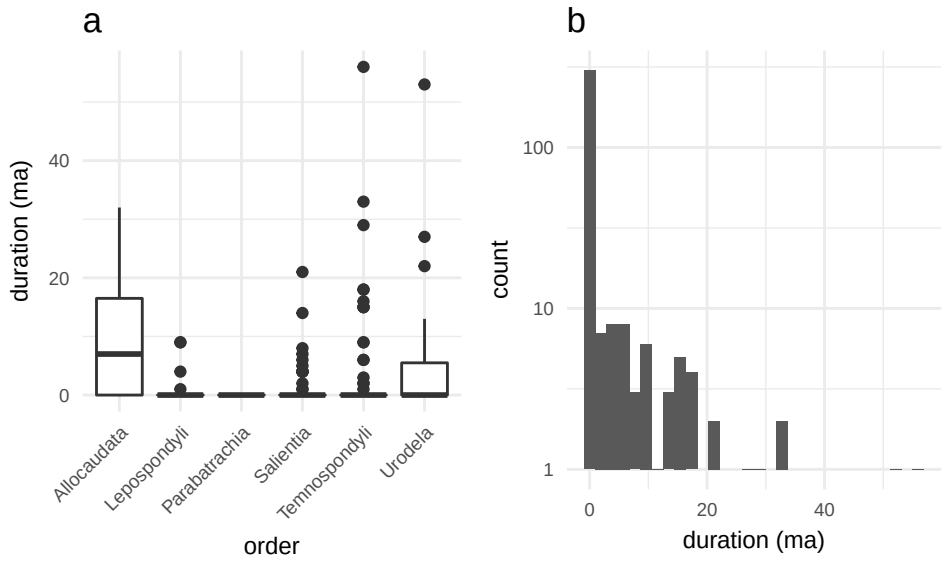

Figure S13: Durations of fossil species as histogram in the main taxonomic groups (a) and as histogram (b). Number of species per group are: Allocaudata (11), Lepospondyli (39), Parabatrachia (2), Salientia (80), Tenmospondyli (175), Urodela (39).

There is a slight taxonomic bias on the duration with Allocaudata having longer durations than Lepospondlyi, Salientia and Temnospondyli. Temnospondyli show shorter duations than Urodela. Allocaudata and Urodela seem to have slightly longer durations on average.

Output 6: Pairwise comparison output from comparing durations between different taxonomic groups.

```
##
##  Pairwise comparisons using Wilcoxon rank sum test
##
## data:  extinct.raw$ma_range and extinct.raw$order
##
##               Allocaudata Lepospondyli Parabatrachia Salientia Temnospondyli
## Lepospondyli  0.0029      -            -             -         -
## Parabatrachia 0.3965      0.7294       -             -         -
## Salientia     0.0037      0.4766       0.6697        -         -
## Temnospondyli 0.0001      0.9889       0.7294        0.2120    -
## Urodela       0.1839      0.0500       0.5270        0.1145    0.0029
##
## P value adjustment method: fdr
```

Adjusting the model for taxonomic differences

To account for differences in model performance between the taxonomic groups, that might be caused by their slightly differing mean stratigraphic ranges, we analysed model performance within each group seperately.

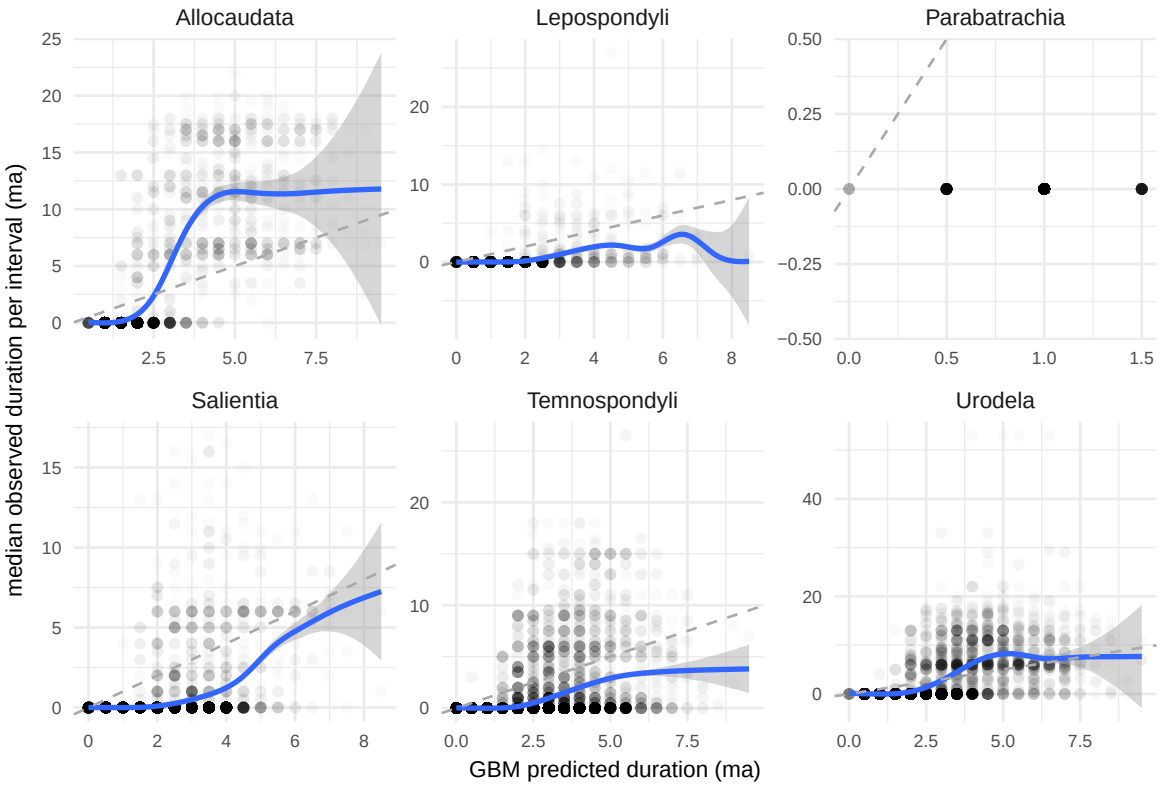

Figure S14: Predicted species durations plotted against observed durations for each taxonomic group. Results are for n=500 repetitions in which the data was randomly split into calibration and test dataset. The blue curve is a generalized additive model with default settings used in geom_smooth() of the ggplot2 package, the grey areas are 95% confidence intervals for the gam. The dashed line is a line through the origin with a slope of 1.

Within Lissamphibia, Salientia seem to be mostly underpredicted in the model, whereas Urodela durations are switching from being underpredicted for short durations to being overpredicted for longer durations. These differences between observed and predicted duration were corrected for in the GBM-CORR predictions (Fig. S20).

Lissamphibia model

To control for potential further influence of these minor differences, we fitted a GBM to the extinct species dataset, using lissamphibian species only (Salientia, Urodela and Parabatrachia). Results are comparable to the GBM on all data in terms of variable importance (Fig. S15) and error estimates (Fig. S19).

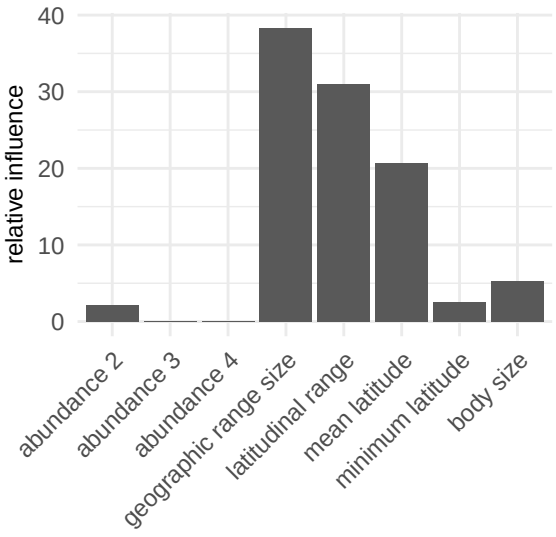

Figure S15: Relative influence of each variable for the final Lissamphibia GBM. The relative influence is an importance measure for the influence of each predictor variable in the model. Values are scaled to match 100%. Variables are described in detail in Table S1.

No singleton model

Removing all taxa who appear in one single chronostratigraphic stage is a common praxis in paleontological quantitative data analysis and supposed to reduce the inclusion of false single-interval species due to bad conversation. Although we doubt that the removal of this potential bias towards short durations outweighs the bias introduced by dramatic dataset reduction, accompanied by massive diversity loss, we fitted a model to a subset of our data. This subset only includes species which have a duration length greater than 0, meaning they were found in at least two chronostratigraphic stages.

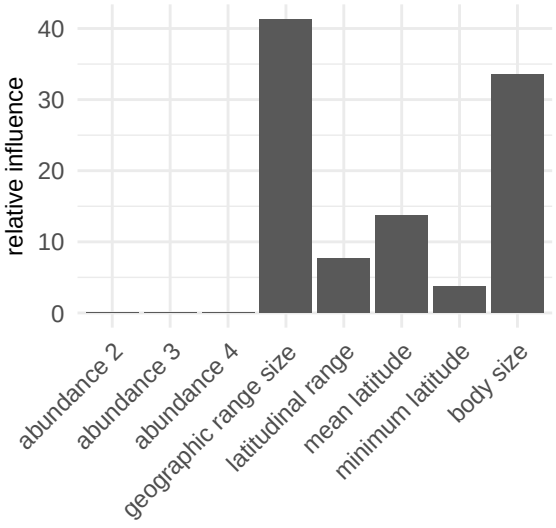

Figure S16: Relative influence of each variable for the no-single-interval species GBM. The relative influence is an importance measure for the influence of each predictor variable in the model. Values are scaled to match 100%. Variables are described in detail in Table S1.

Null model

To test the Null hypothesis that there is no connection between traits and the survival length, and therefore extinction risk, of species, we created a null model by fitting the GBM to a randomized dataset. The only variable which was randomized was the duration, therefore the other trait combinations stayed as they were to avoid having biological unmeaningful combinations.

Output 7: Output of the null GBM.

```
## Stochastic Gradient Boosting
##
## 354 samples
##   6 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 3 times)
## Summary of sample sizes: 318, 319, 319, 319, 318, 319, ...
## Resampling results across tuning parameters:
##
##   shrinkage  interaction.depth  n.minobsinnode  n.trees  RMSE      Rsquared
##   0.001      1                   5               50      5.654018  0.021974056
##   0.001      1                   5              100      5.655204  0.020665476
##   0.001      1                   5              150      5.656760  0.017016359
##   0.001      1                   5              200      5.658972  0.017744389
##   0.001      1                   5              250      5.660760  0.017194330
##   0.001      1                   5              300      5.662761  0.016984220
##   0.001      1                   5              350      5.665063  0.017423250
##   MAE
##   3.084217
##   3.084275
##   3.084732
##   3.085677
##   3.086048
##   3.086237
##   3.086703
##  [ reached getOption("max.print") -- omitted 263 rows ]
##
## RMSE was used to select the optimal model using the smallest value.
## The final values used for the model were n.trees = 50, interaction.depth =
##  1, shrinkage = 0.001 and n.minobsinnode = 15.
```

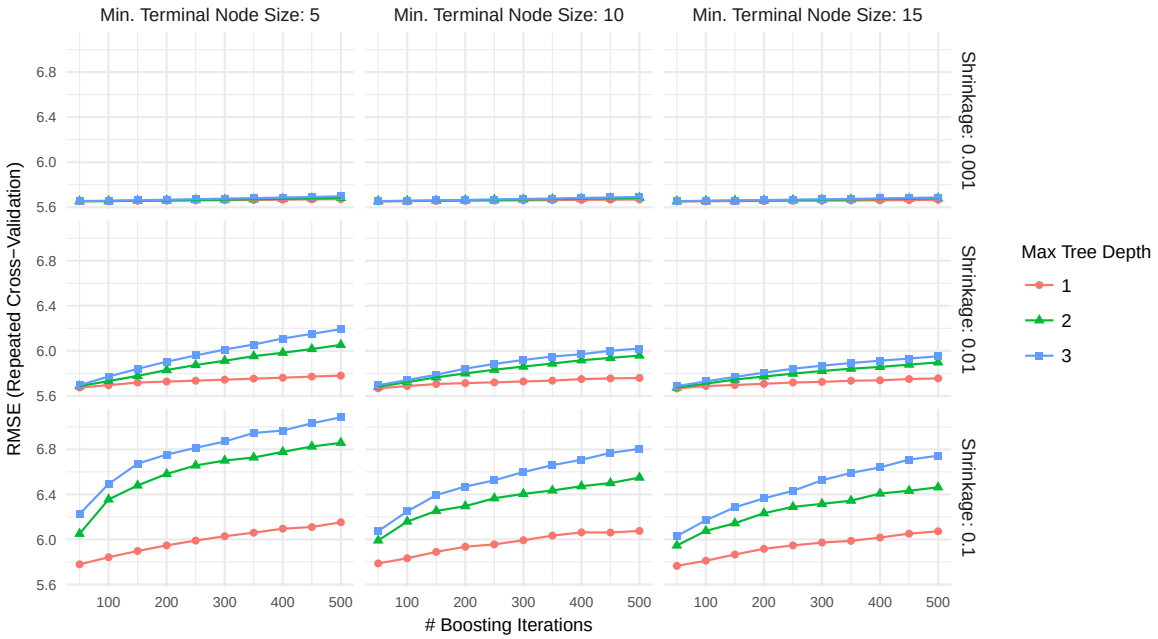

Figure S17: Relationship between the Root-mean-squared error (RMSE) of the GBM as estimate of performance and the tuning parameters: minimum terminal node size (vertical panels), shrinkage (horizontial panels) and maximum tree depth (legend). Minimum terminal node size defines the minimum number of observations in the trees terminal nodes, shrinkage is the learning rate of the model, maximum tree depth is the interaction depth for variables, with 1 being an additive model.

Null model bootstrap

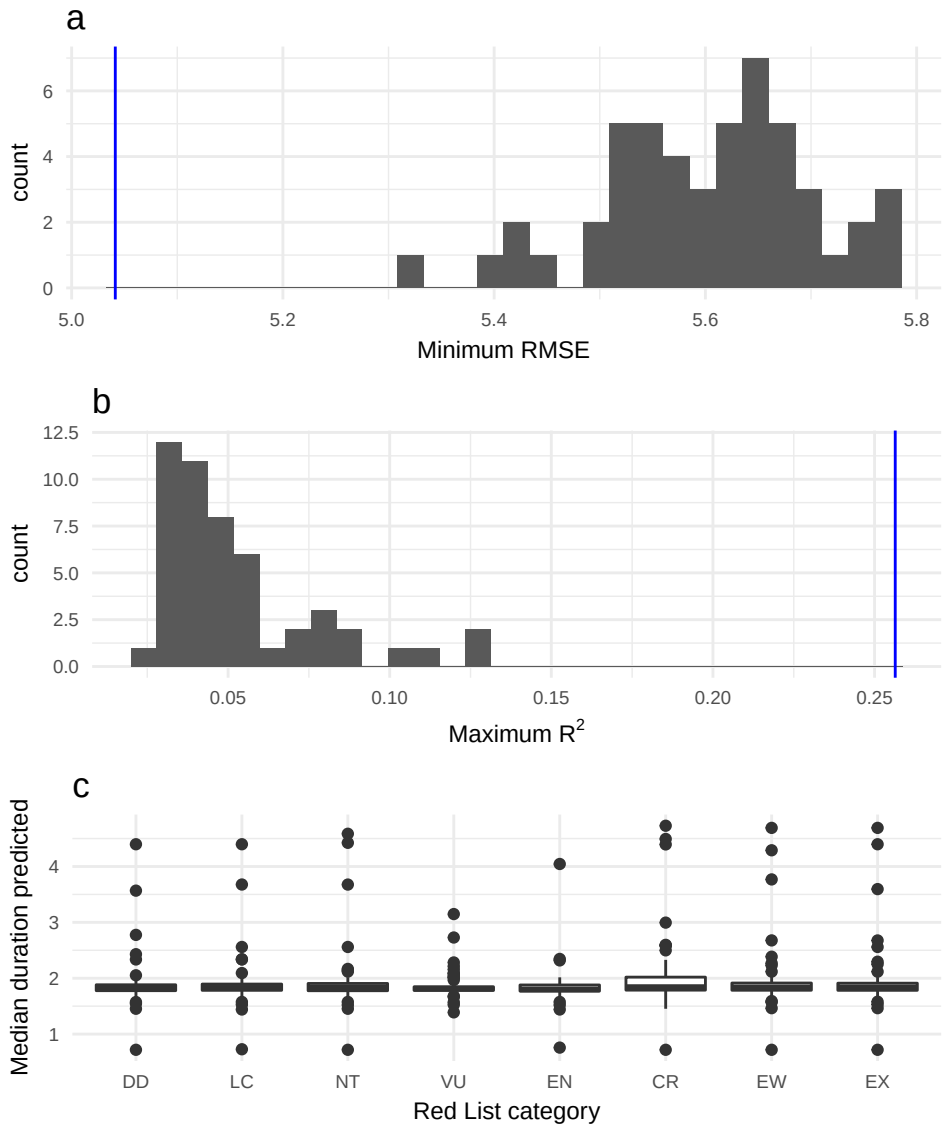

Figure S18: Null GBM bootstrap summaries. A) shows the RMSE values of all 50 null models, b) shows the maximum R^2 in each null model fitting, c) shows the predicted medium durations for the IUCN Red List categories of all 50 null models. The blue lines depict the corresponding values for RMSE and R^2 from the full GBM.

Model comparison

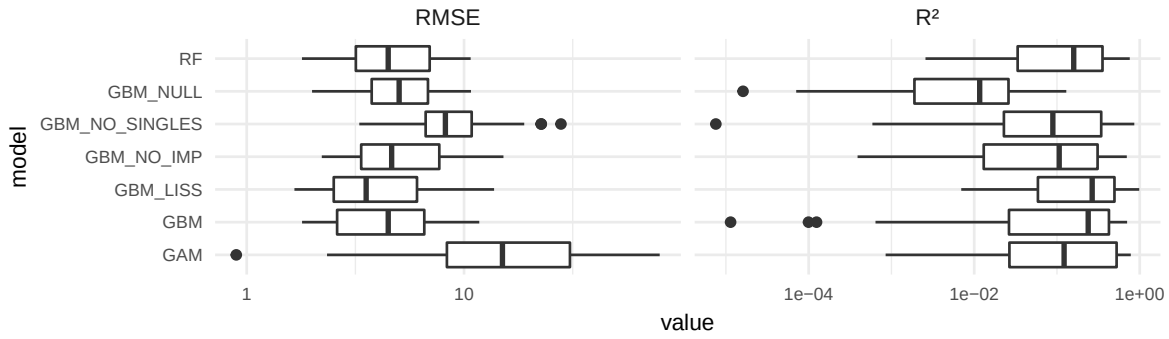

Figure S19: Model performance distributions for the final model cross validation. Each model configuration is tested in a 3 times 10-fold cross validation procedure, the results are displayed here. Sample size is n=30 for each distribution, in accordance with the 3 times 10-fold cross validation procedure. The final model is chosen based on the mean RMSE values of these distribution. RF = randomForest, GBM_NULL = GBM Null model, GBM_NO_SINGLES = GBM without single interval species, GBM_NO_IMP = GBM on data without imputation, GBM_LISS = GBM on Lissamphibia only, GBM = Generalized boosted model, GAM = Generalized additive model.

Table S5: Model comparison. The average RMSE (Root mean squared error) and R^2 including their standard deviations (SD) associated with the optimal tuning parameters across the resamples from cross validation. N=30 for each model, like in Fig. S19.

	RMSE	R^2	RMSE SD	R^2 SD
GBM	5.04	2.79	0.26	0.23
GBM no imputation	5.93	3.65	0.22	0.24
GBM Lissamphibia	4.86	3.48	0.32	0.30
GBM Null model	5.65	2.61	0.02	0.03
random Forest	5.20	2.58	0.25	0.25
GBM no single interval species	10.02	5.99	0.22	0.27
GAM	21.96	20.20	0.27	0.27

In terms of RMSE, the only model being significantly different from the others was the model excluding all single interval taxa, which was not surprising given the much lower sample size. The R^2 value of the null model differed significantly from all other models.

Output 8: Model comparison t-Test.

```
##
## Call:
## summary.diff.resamples(object = difValues, round = 2)
##
## p-value adjustment: fdr
## Upper diagonal: estimates of the difference
## Lower diagonal: p-value for H0: difference = 0
##
## MAE
##                GBM       GBM_NO_IMP GBM_LISS  RF        GBM_NULL  GAM
## GBM                      -1.0920    -0.2444    0.1701   -0.5105   -3.1098
## GBM_NO_IMP     0.001857              0.8476    1.2621    0.5815   -2.0178
## GBM_LISS       0.426580  0.012074              0.4144   -0.2661   -2.8654
## RF             0.412882  7.525e-05  0.145852           -0.6806   -3.2798
## GBM_NULL       0.011968  0.071834   0.334138  0.003172           -2.5993
## GAM            0.001511  0.020294   0.001857  0.000273  0.004350
## GBM_NO_SINGLES 1.389e-06 1.001e-05  1.623e-06 6.617e-07 1.924e-06 0.052298
##                GBM_NO_SINGLES
## GBM            -5.2262
## GBM_NO_IMP     -4.1341
## GBM_LISS       -4.9818
## RF             -5.3962
## GBM_NULL       -4.7157
## GAM            -2.1164
## GBM_NO_SINGLES
##
## RMSE
##                GBM       GBM_NO_IMP GBM_LISS  RF        GBM_NULL  GAM
## GBM                       -0.8913     0.1849   -0.1567   -0.6110  -16.9138
## GBM_NO_IMP     0.4704483              1.0762    0.7346    0.2803  -16.0225
## GBM_LISS       0.8471417 0.4486286             -0.3416   -0.7960  -17.0988
## RF             0.8471417 0.4881829  0.7882626           -0.4543  -16.7571
## GBM_NULL       0.5185537 0.8363928  0.4620276 0.6192691           -16.3028
## GAM            0.0007052 0.0007052  0.0007052 0.0007052 0.0007052
## GBM_NO_SINGLES 0.0013219 0.0030928  0.0009961 0.0007052 0.0010864 0.0059675
##                GBM_NO_SINGLES
## GBM             -4.9740
## GBM_NO_IMP      -4.0827
## GBM_LISS        -5.1589
## RF              -4.8172
## GBM_NULL        -4.3629
## GAM             11.9399
## GBM_NO_SINGLES
##
## Rsquared
##                GBM       GBM_NO_IMP GBM_LISS  RF        GBM_NULL  GAM
## GBM                       0.039199  -0.065800  0.012810  0.233964 -0.009782
## GBM_NO_IMP     0.8273996            -0.104999 -0.027066  0.194765 -0.049657
```

```
## GBM_LISS       0.7666365 0.4836255            0.064931  0.299764  0.042340
## RF             0.9030371 0.8273996  0.8273996           0.226103 -0.022591
## GBM_NULL       7.634e-05 0.0008001  7.634e-05 0.0001940           -0.248694
## GAM            0.9260584 0.8273996  0.8273996 0.8273996 0.0001940
## GBM_NO_SINGLES 0.8273996 0.9260584  0.5143430 0.9030371 0.0014623 0.8273996
##                GBM_NO_SINGLES
## GBM             0.032601
## GBM_NO_IMP     -0.006599
## GBM_LISS        0.098401
## RF              0.019881
## GBM_NULL       -0.201364
## GAM             0.042473
## GBM_NO_SINGLES
```

Predictions

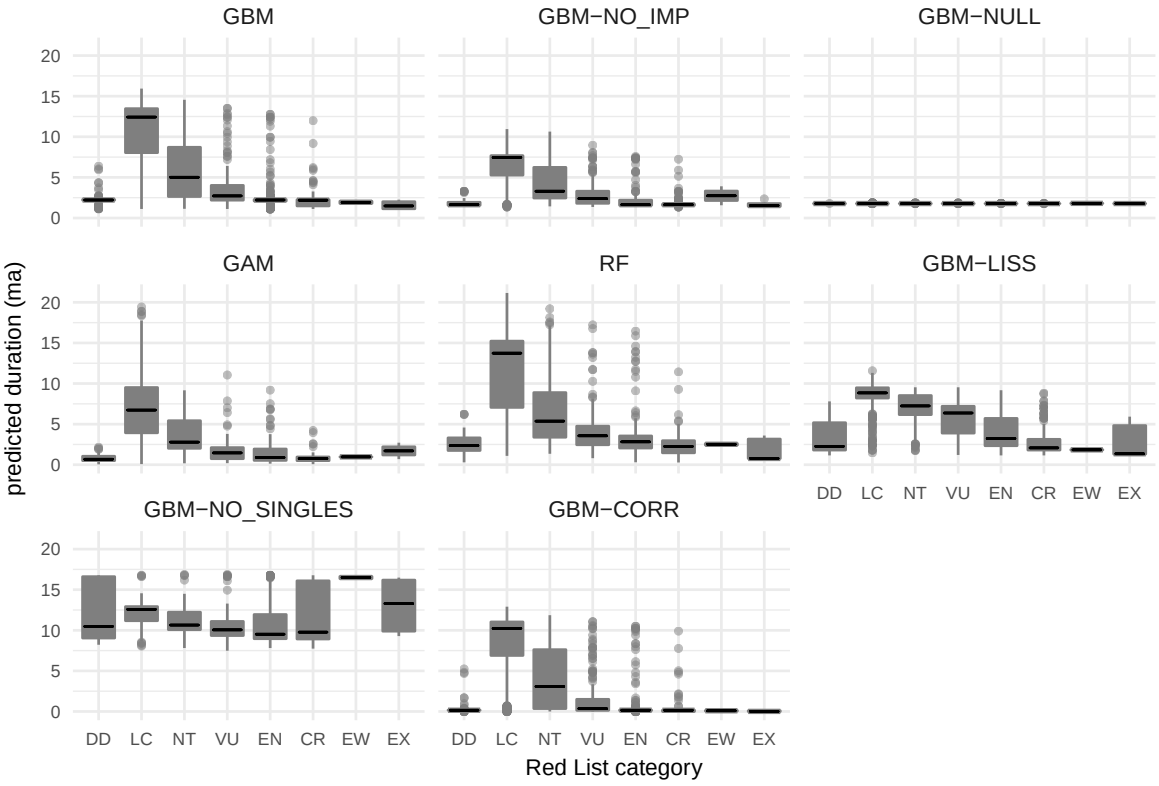

Figure S20: Predicted durations of living species based on all build models. Note that the predicted durations of the GAM have been square-root transformed for better visibility, as the range of durations was much larger than for the other models. GBM = Generalized boosted model, GBM-NO_IMP = GBM on data without imputation, GBM_NULL = GBM Null model, GAM = Generalized additive model, RF = randomForest, GBM_LISS = GBM on Lissamphibia only, GBM_NO_SINGLES = GBM without single interval species, GBM-CORR = GBM with taxonomic bias correction (Fig. S14).

Statistical analysis of the prediction results

The pairwise comparisons of predicted values per group of IUCN Red List extinction risk category are depicted in the following output files.

Output 9: Kruskal-Wallis rank sum test and Pairwise Wilcoxon rank sum test console outputs for the gam.

```
##
##  Kruskal-Wallis rank sum test
##
## data:  res$predict.gam and res$Red.List.status
## Kruskal-Wallis chi-squared = 368.34, df = 7, p-value < 2.2e-16

##
##  Pairwise comparisons using Wilcoxon rank sum test
##
## data:  as.numeric(res$predict.gam) and res$Red.List.status
##
##    DD      LC      NT      VU      EN      CR      EW
## LC 1.3e-15 -       -       -       -       -       -
## NT 0.01767 5.1e-12 -       -       -       -       -
## VU 0.55831 < 2e-16 0.00079 -       -       -       -
## EN 0.00679 < 2e-16 0.00014 0.54427 -       -       -
## CR 0.04830 < 2e-16 0.00066 0.85147 0.27342 -       -
## EW 0.96205 0.30050 0.84154 0.85147 0.55831 0.55831 -
## EX 0.87223 0.04830 0.61962 0.88826 0.76062 0.85147 1.00000
##
## P value adjustment method: fdr
```

Output 10: Kruskal-Wallis rank sum test and Pairwise Wilcoxon rank sum test console outputs for the rF.

```
##
##  Kruskal-Wallis rank sum test
##
## data:  res$predict.rf and res$Red.List.status
## Kruskal-Wallis chi-squared = 746.55, df = 7, p-value < 2.2e-16

##
##  Pairwise comparisons using Wilcoxon rank sum test
##
## data:  res$predict.rf and res$Red.List.status
##
##    DD      LC      NT      VU      EN      CR      EW
## LC < 2e-16 -       -       -       -       -       -
## NT 2.6e-14 3.6e-16 -       -       -       -       -
## VU 2.5e-07 < 2e-16 7.4e-07 -       -       -       -
## EN 0.02173 < 2e-16 9.2e-15 4.7e-05 -       -       -
## CR 0.35076 < 2e-16 < 2e-16 3.5e-13 1.6e-05 -       -
## EW 0.87025 0.03271 0.10943 0.31451 0.62282 0.63518 -
## EX 0.31451 0.00048 0.00504 0.02740 0.13439 0.35076 0.87025
```

```
##
## P value adjustment method: fdr
```

Output 11: Kruskal-Wallis rank sum test and Pairwise Wilcoxon rank sum test console outputs for the GBM.

```
##
##  Kruskal-Wallis rank sum test
##
## data:  res$predict.gbm1 and res$Red.List.status
## Kruskal-Wallis chi-squared = 734.79, df = 7, p-value < 2.2e-16

##
##  Pairwise comparisons using Wilcoxon rank sum test
##
## data:  res$predict.gbm1 and res$Red.List.status
##
##    DD      LC      NT      VU      EN      CR      EW
## LC < 2e-16 -       -       -       -       -       -
## NT 2.0e-10 < 2e-16 -       -       -       -       -
## VU 7.9e-05 < 2e-16 9.7e-06 -       -       -       -
## EN 0.34523 < 2e-16 1.4e-13 1.9e-05 -       -       -
## CR 0.17458 < 2e-16 < 2e-16 1.8e-09 0.00752 -       -
## EW 0.32261 0.03365 0.09321 0.18873 0.32063 0.62102 -
## EX 0.05961 0.00043 0.00261 0.01152 0.04602 0.12650 0.59259
##
## P value adjustment method: fdr
```

Output 12: Kruskal-Wallis rank sum test and Pairwise Wilcoxon rank sum test console outputs for the GBM on Lissamphibia.

```
##
##  Kruskal-Wallis rank sum test
##
## data:  res$predict.gbm.liss and res$Red.List.status
## Kruskal-Wallis chi-squared = 789.38, df = 7, p-value < 2.2e-16

##
##  Pairwise comparisons using Wilcoxon rank sum test
##
## data:  res$predict.gbm.liss and res$Red.List.status
##
##    DD      LC      NT      VU      EN      CR      EW
## LC < 2e-16 -       -       -       -       -       -
## NT < 2e-16 < 2e-16 -       -       -       -       -
## VU 7.1e-12 < 2e-16 6.9e-06 -       -       -       -
## EN 0.00077 < 2e-16 < 2e-16 6.1e-11 -       -       -
## CR 0.43395 < 2e-16 < 2e-16 < 2e-16 2.7e-08 -       -
## EW 0.46541 0.02305 0.03351 0.04943 0.08848 0.46541 -
## EX 0.31384 0.00036 0.00205 0.01338 0.17335 0.43395 0.85714
##
## P value adjustment method: fdr
```

Output 13: Kruskal-Wallis rank sum test and Pairwise Wilcoxon rank sum test console outputs for the GBM on the subset without single-interval species.

```
##
##  Kruskal-Wallis rank sum test
##
## data:  res$predict.gbm.nosingles and res$Red.List.status
## Kruskal-Wallis chi-squared = 194.88, df = 7, p-value < 2.2e-16

##
##  Pairwise comparisons using Wilcoxon rank sum test
##
## data:  res$predict.gbm.nosingles and res$Red.List.status
##
##    DD      LC      NT      VU      EN      CR      EW
## LC 0.22917 -       -       -       -       -       -
## NT 0.91852 2.8e-10 -       -       -       -       -
## VU 0.21681 < 2e-16 0.00017 -       -       -       -
## EN 0.04342 < 2e-16 1.2e-05 0.04342 -       -       -
## CR 0.30690 0.00057 0.15248 0.83872 0.22227 -       -
## EW 0.21831 0.04679 0.05977 0.05977 0.09672 0.16662 -
## EX 0.70278 0.49778 0.47507 0.21831 0.17457 0.33949 0.25397
##
## P value adjustment method: fdr
```

Output 14: Kruskal-Wallis rank sum test and Pairwise Wilcoxon rank sum test console outputs for the null model.

```
##
##  Kruskal-Wallis rank sum test
##
## data:  res$predict.gbm.null and res$Red.List.status
## Kruskal-Wallis chi-squared = 72.278, df = 7, p-value = 5.113e-13

##
##  Pairwise comparisons using Wilcoxon rank sum test
##
## data:  res$predict.gbm.null and res$Red.List.status
##
##    DD      LC      NT      VU      EN      CR      EW
## LC 0.01326 -       -       -       -       -       -
## NT 0.17509 0.79977 -       -       -       -       -
## VU 0.88359 0.00034 0.09811 -       -       -       -
## EN 0.90963 3.4e-05 0.08966 0.90963 -       -       -
## CR 0.08966 3.3e-11 0.00068 0.09811 0.04727 -       -
## EW 0.11502 0.21762 0.36991 0.17509 0.13233 0.09086 -
## EX 0.88359 0.74059 0.88359 0.88359 0.88359 0.47332 0.32105
##
## P value adjustment method: fdr
```

Potentially misclassified species

Using our model to identify potential misclassifications in the IUCN Red List assignments could help focusing the limited conservation actions to the right species. However, a classification is proving difficult as our model is a numerical model, not a classfication model. Therefore, a species that is indeed falsely classified within the Red List will likely show up with an unusually long or short predicted duration in this category, as we predict durations, but not cateogries.

While there are more ways to identify potential misclassification, we chose the simplest one and defined a misclassfication as a species whos duration plots outside the whiskers of the predicted duration boxplots in each catergory (Fig. 3); that is durations either larger than the third quartile + 1.5 * IQR, or shorter than the first quartile - 1.5 * IQR. The following table (Tab. S6) summarizes these potential misclassifications for each Red List status.

Table S6: Number of potentially misclassified species in the IUCN Red List. Misclassifications were defined as statistical outliers, as seen in the prediction boxplots (Fig. 3).

	DD	LC	NT	VU	EN	CR	EW	EX
Longer	13	0	0	19	42	10	0	0
Shorter	15	0	0	0	44	0	0	0
Total	66	740	98	151	192	128	2	5
Longer %	20	0	0	13	22	8	0	0
Shorter %	23	0	0	0	23	0	0	0

References

Harting, J. & Wong, M. (1979). Algorithm AS 136: a k-means clustering algorithm. *J. R. Stat. Soc. C*, 28, 100–108.

Hunt, L.A. (2017). Missing data imputation and its effect on the accuracy of classification. In: *Data science. innovative developments in data analysis and clustering* (eds. Palumbo, F., Montanari, A. & Vichi, M.). Springer Nature, Cham, Switzerland, pp. 3–14.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C. & Engelhardt, A. *et al.* (2017). *caret: Classification and Regression Training*.

R Core Team. (2017). R: A Language and Environment for Statistical Computing. Vienna, Austria. https://www.R-project.org/.

Van Buuren, S. & Groothuis-Oudshoorn, K. (2011). Multivariate Imputation by Chained Equations. *J. Stat. Softw.*, 45, 1–67.

Xie, Y. (2014). knitr: A Comprehensive Tool for Reproducible Research in {R}. In: *Implementing reproducible computational research* (eds. Stodden, V., Leisch, F. & Peng, R.D.). Chapman; Hall/CRC.

Xie, Y. (2015). *Dynamic Documents with {R} and knitr*. 2nd edn. Chapman; Hall/CRC, Boca Raton, Florida.

Xie, Y. (2016). *knitr: A General-Purpose Package for Dynamic Report Generation in R*.