# Lithobates pipiens
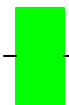
Habitat Suitability Model (HSM) assessment metrics and metadata
Common name: Northern Leopard Frog
Grank: G5 - Secure
Date: 03 Dec 2021; Code: lithpipi (EGT_ID: 101454)

good
TSS=0.99
validation success

This model was developed for the Arizona Game and Fish Department. This HSM was modeled using the algorithms in Table 1 using R[4], incorporating the number of known and background locations indicated in Table 2. We validated the model by jackknifing[5,6,7] by spatial grouping and testing 50 groups. Table 3 reports the validation statistics for these jackknifing runs.

Expert reviewers indicated this algorithm performed the best: xgb. The reported TSS (above and Table 3b) is based on the reviewer-defined threshold for this model. Some data are provided for all models run, but the focus of this metadata report will be on this model chosen by expert review.

Table 1. These algorithms were created for review. If multiple passed review, those were ensembled. The specific implementations are listed in the package column.

| Name | Code | R package |
|------|------|-----------|
| Random Forest | rf | randomForest |
| Extreme Gradient Boosting | xgb | xgboost |

Table 2. Input statistics. Presence locations may be based on point observations buffered by their spatial accuracy and polygons with nearby locations grouped together. Environmental conditions are randomly subsampled within each of these groups as noted to generate the number of presence inputs (by tree for rf; for entire model for others). Background points are placed throughout model except for presence locations.

| Sample | Count |
|--------|-------|
| Presence locations (groups) | 187 |
| Subsamples within groups | 4.5 |
| Total presence inputs | 841 |
| Background inputs - rf | 841 |
| Background inputs - xgb | 841 |

Table 3. Mean validation statistics for jackknife trials (+/- standard deviation). AUC = area under the ROC curve, Sens = sensitivity, Spec = specificity, TSS = True Skill Statistic[8,9,6].

| alg | AUC | Sens | Spec | TSS |
|-----|-----|------|------|-----|
| rf | 0.98(0.04) | 0.96(0.17) | 0.88(0.07) | 0.84(0.18) |
| xgb | 0.94(0.07) | 0.89(0.16) | 0.86(0.16) | 0.75(NA) |

Table 3b. Validation statistics for expert-chosen algorithm and threshold. TSS = True Skill Statistic[8].

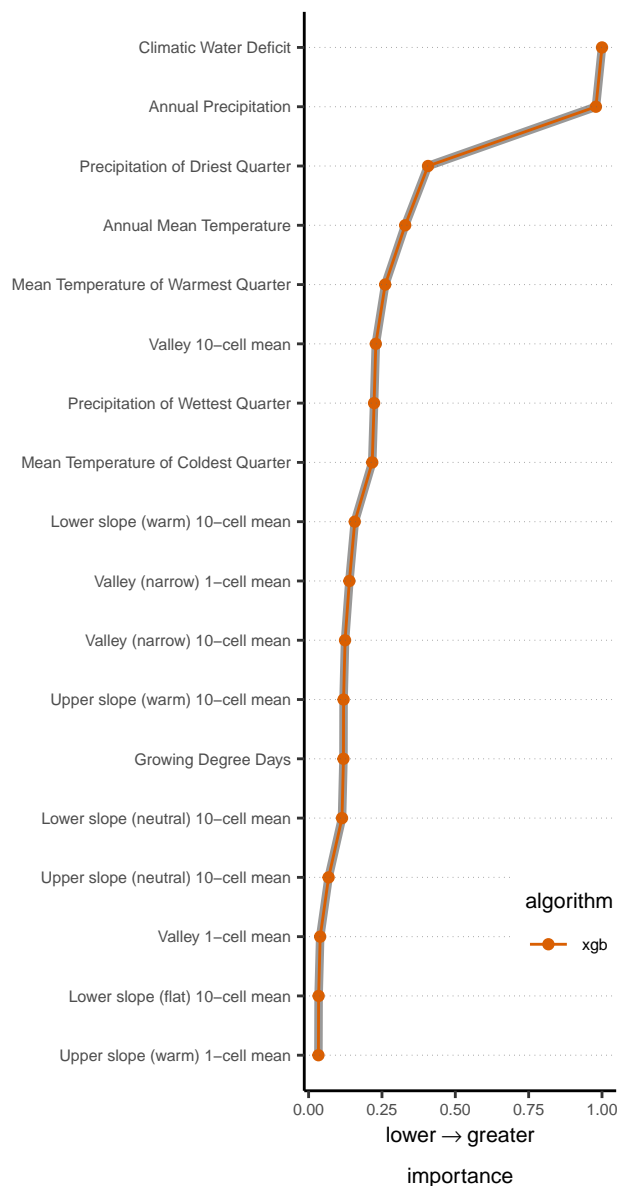| alg | thresh | TSS |
|-----|--------|-----|
| xgb | 0.47 | 0.99 |



Figure 1. Relative importance of environmental variables based on the full model. Importance values are extracted from each algorithm and plotted in descending order based on an average of all algorithms (grey line). Note each algorithm has variable removal rules (see Appendix 2); a missing point indicates the variable was not used by that algorithm. See Appendix 1 for variable descriptions.
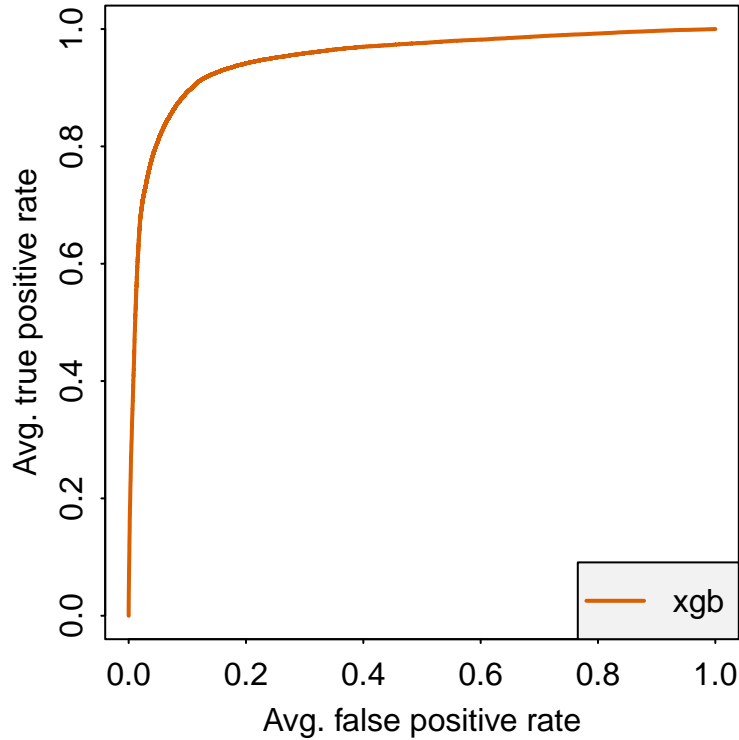
Figure 2. ROC plot for all 50 validation runs, averaged along cutoffs, for each modeling algorithm.

Species distribution model outputs display the probability (0-1) of a location (i.e. stream reach or raster cell) having similar environmental conditions in comparison to known presence locations. No model will ever depict sites where a targeted element will occur with certainty, it can *only* depict locations it interprets as appropriate habitat for the targeted element. The delineation of suitable habitats is made by the selection of a threshold value, where locations with values above the threshold are designated as likely suitable habitat, and those with values below the threshold may be unsuitable. Threshold values are often statistically calculated. SDMs can be used in many ways and the depiction of appropriate habitat should be varied depending on intended use. For targeting field surveys, an SDM may be used to refine the search area; users should always employ additional GIS tools to further direct search efforts. A lower threshold depicting more area may be appropriate to use in this case. For a more conservative depiction of suitable habitat that shows less area, a higher threshold may be more appropriate. Different thresholds for this model (full model) are described in Table 4.

Table 4. Thresholds[11,12] calculated from the final model. The Value column reports the threshold; Groups indicates the percentage (number in brackets) of groups within which at least one point was predicted as suitable habitat; Pts indicates the percentage of PR points predicted having suitable habitat. Total numbers of groups and presence points used in the final model are reported in Table 1.

| Code | Value | Groups | Points |
|---|---|---|---|
| Algorithm = xgb | | | |
| eqSS | 0.722 | 99(186) | 100 |
| maxSSS | 0.722 | 99(186) | 100 |
| MTP | 0.285 | 100(187) | 100 |
| MTPGP | 0.285 | 100(187) | 100 |
| TenPctile | 0.918 | 95(177) | 90 |
| ROC | 0.722 | 99(186) | 100 |

| Code | Threshold full name | Threshold description |
|---|---|---|
| eqSS | Equal sensitivity and specificity | The probability at which the absolute value of the difference between sensitivity and specificity is minimized. |
| maxSSS | Maximum of sensitivity plus specificity | The probability at which the sum of sensitivity and specificity is maximized. |
| MTP | Minimum Training Presence | The highest probability value at which 100% of input presence points remain classified as suitable habitat. |
| MTPGP | Minimum Training Presence by Group | The highest probability value at which 100% of input groups have at least one presence point classified as suitable habitat. |
| TenPctile | Tenth percentile of training presence | The probability at which 90% of the input presence points are classified as suitable habitat. |
| ROC | ROC plot upper left corner | The point on the ROC curve with the shortest distance to the top-left corner (0,1) of the ROC plot. |
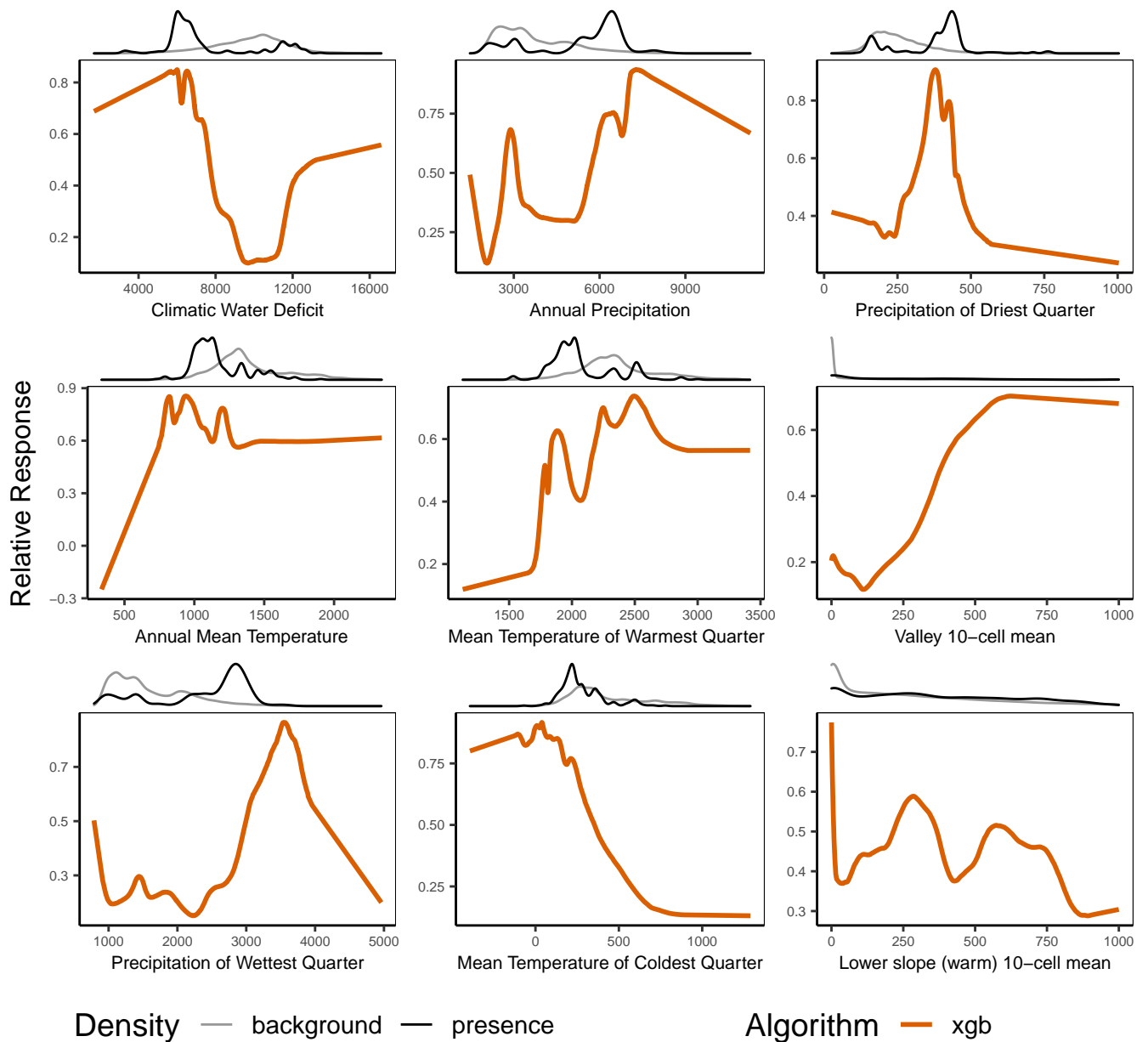
Figure 3. Partial dependence plots for the 9 environmental variables with the most influence on the models (averaged). Each plot shows the effect of the variable on the probability of appropriate habitat with the effects of the other variables removed [3]. The x-axis covers the range of values for the variable assessed; the y-axis represents the effect between the variable and model response. Peaks in the line indicate where this variable had the strongest influence on predicting appropriate habitat. Decreasing lines from left to right show a negative relationship overall; increasing lines, positive. The distribution of each category (thin red = Background points, thick blue = Presence points) is depicted at the top margin. See Appendix 1 for variable descriptions.

## Model Evaluation and Intended Use

All SDMs are sensitive to data inputs and methodological choices. Table 5 presents scoring of modeling factors based on the model evaluation rubric presented in Sofaer et al. 2019[13].

Table 5. Model evaluation results based on Sofaer et al. (2019). Scores can be attributed as ideal, acceptable, or interpret with caution.

| Category | Metric | Score | Notes |
|---|---|---|---|
| Species Data | Presence data quality | Acceptable | Heritage Network data augmented with outside data which may or may not be vetted for accuracy or weighted for spatial representation. |
| | Absence/Background Data | Acceptable | Background points randomly placed throughout modeling area excluding species locations. |
| | Evaluation Data | Acceptable | Models are validated by jackknifing (i.e. leave-one-out). |
| Environmental Predictors | Ecological and predictive relevance | Acceptable | Selection of predictor variables were based on previous modeling experience by the Natural Heritage Network and subsetted using variable importance. |
| | Spatial and temporal alignment | Acceptable | Reasonable attempts to align predictor and presence data were made. |
| Modeling Process | Algorithm choice | Acceptable | Algorithms were carefully chosen for modeling rare species. |
| | Sensitivity | Acceptable | Ensemble used to minimize algorithm sensitivity. |
| | Statistical rigor | Acceptable | Collinearity of predictors recognized and addressed; presence points grouped to minimize sample bias and minimize spatial autocorrelation boost during validation; other assumptions recognized and considered. |
| | Performance | Acceptable | Mean TSS for expert-derived thresholds $\geq 0.6$. |
| | Model review | Acceptable | Model was reviewed by regional, taxonomic experts. |
| Model Products | Mapped products | Acceptable | Continuous models plus thresholded models available to users. |
| | Interpretation support products | Ideal | All standards met. |
| | Reproducibility | Ideal | All standards met. |
| | Iterative | Interpret with Caution | Model not re-run with new or modified data. |

## Model Comments

The standard variables (nlcdopn1, nlcdopn10, nlcdopn100, impsur1, impsur10, impsur100, ntm_1_01, ntm_1_02, ntm_1_06, ntm_1_08, ntm_1_09, ntm_2_01, ntm_2_02, ntm_2_05, ntm_2_06, ntm_3_01, ntm_3_03, ntm_3_09, ntm_3_12, ntm_4_01, ntm_4_02, ntm_4_03, ntm_4_05, ntm_4_06, ntm_5_01, ntm_6_01, ntm_6_02, ntm_6_03, ntm_6_04, nlcdshb1, nlcdshb10, nlcdshb100, soil_ph, poresize, clay, bulkdens) were excluded from this model.
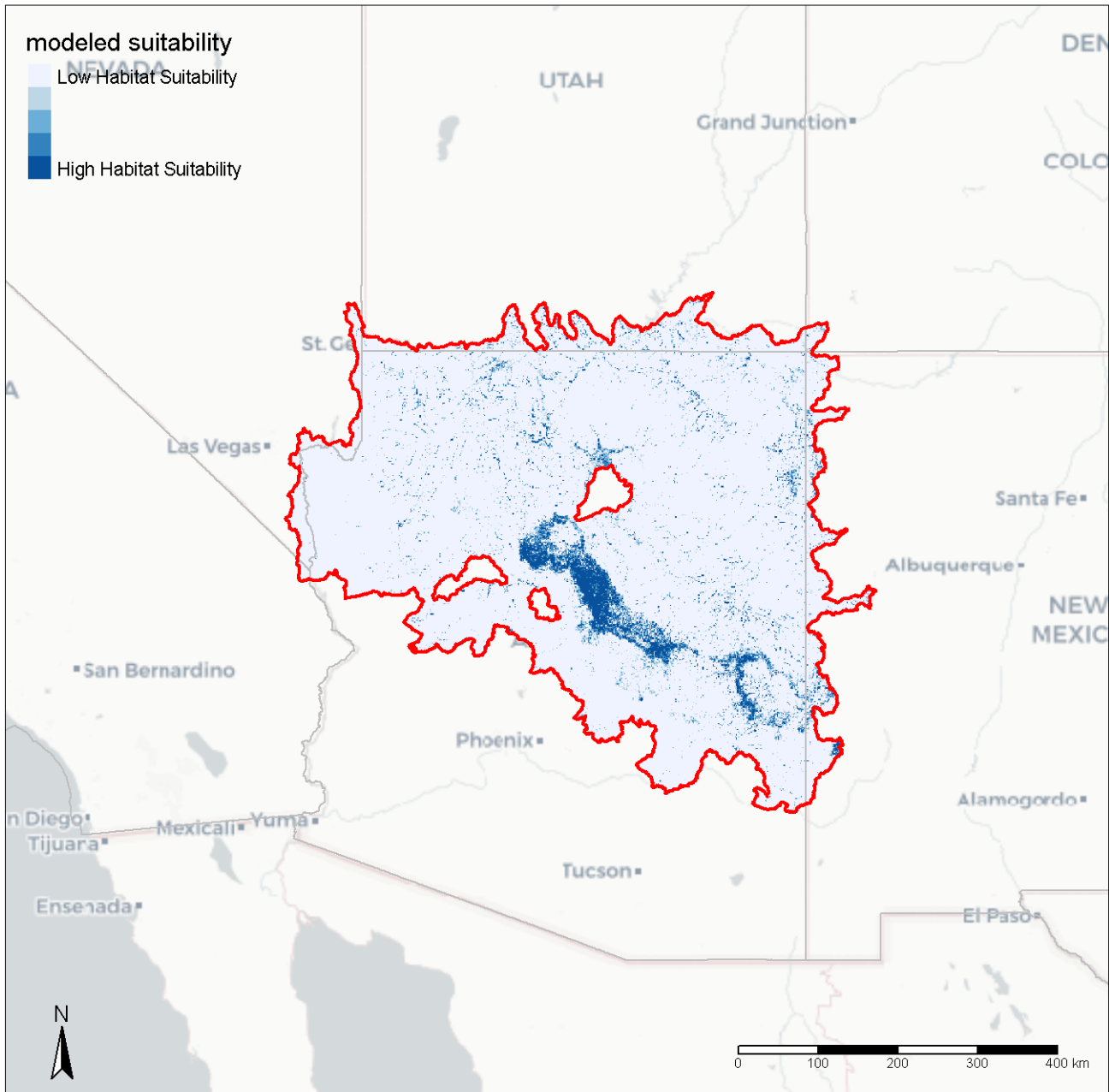
Figure 4. A generalized view of the model predictions throughout the modeled area. If an ensemble was created this map shows simply an average of all model predictions. State boundaries are depicted as a thin gray line. The modeled area is outlined in red. Basemap: CartoDB.Positron (©OpenStreetMap, contributors: ©CARTO).

This distribution model would not have been possible without data sharing among organizations. Other data sets and sources may have been evaluated, but this final model includes data from these sources:

- Arizona Game and Fish Department

This model was built using a methodology developed through collaboration among the Florida Natural Areas Inventory, the New York Natural Heritage Program, the Pennsylvania Natural Heritage Program, and the Virginia Natural Heritage Program, all member programs of the NatureServe Network. It is one of a suite of species distribution models developed using the same methods, scripts, and environmental data sets. Our goal was to be consistent and transparent in our methodology, validation, and output.

**References**

[1] Breiman, L. 2001. Random forests. Machine Learning 45:5-32.

[2] Iverson, L. R., A. M. Prasad, and A. Liaw. 2004. New machine learning tools for predictive vegetation mapping after climate change: Bagging and Random Forest perform better than Regression Tree Analysis. Landscape ecology of trees and forests.Proceedings of the twelfth annual IALE (UK) conference, Cirencester, UK, 21-24 June 2004 317-320.

[3] Liaw, A. and M. Wiener. 2002. Classification and regression by randomForest. R News 2:18-22. Version 4.6-14.

[4] R Core Team. 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/. R version 4.1.0 (2021-05-18).

[5] Fielding, A. H. and J. F. Bell. 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environmental Conservation 24:38-49.

[6] Fielding, A. H. 2002. What are the appropriate characteristics of an accuracy measure? Pages 271-280 in Predicting Species Occurrences, issues of accuracy and scale. J. M. Scott, P. J. Helglund, M. L. Morrison, J. B. Haufler, M. G. Raphael, W. A. Wall, F. B. Samson, eds. Island Press, Washington.

[7] Pearson, R.G. 2007. Species Distribution Modeling for Conservation Educators and Practitioners. Synthesis. American Museum of Natural History. Available at http://ncep.amnh.org.

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

[9] Vaughan, I. P. and S. J. Ormerod. 2005. The continuing challenges of testing species distribution models. Journal of Applied Ecology 42:720-730.

[10] Sing, T., O. Sander, N. Beerenwinkel, T. Lengauer. 2005. ROCR: visualizing classifier performance in R. Bioinformatics 21(20):3940-3941.

[11] Liu, C., P. M. Berry, T. P. Dawson, and R. G. Pearson. 2005. Selecting thresholds of occurrence in the prediction of species distributions. Ecography 28:385–393.

[12] Liu, C., G. Newell, and M. White. 2015. On the selection of thresholds for predicting species occurrence with presence-only data. Ecology and Evolution 6:337–348.

[13] Sofaer, H. R., C. S. Jarnevich1, I. S. Pearse, R. L. Smyth, S. Auer, G. L. Cook, T. C. Edwards, Jr., G. F. Guala, T. G. Howard, J. T. Morisette, and H. Hamilton. 2019. The development and delivery of species distribution models to inform decision-making. BioScience 69:544-557.

Appendix 1. Descriptions for environmental variables included in model.

| Variable Name | Variable Description |
| --- | --- |
| Annual Mean Temperature | Annual Mean Temperature |
| Annual Precipitation | Annual Precipitation |
| Canopy 100-cell mean | mean percent canopy cover in 100-cell radius (30 meter cells) |
| Climatic Water Deficit | Climatic Water Deficit |
| Coniferous forest cover 100-cell mean | mean coniferous forest cover within 100 cell radius |
| Dist to Carbonate Residual Material | Euclidean distance to surficial geology type: Carbonate Residual Material |
| Dist to Extrusive Volcanic Rock | Euclidean distance to surficial geology type: Extrusive Volcanic Rock |
| Dist to Fine Coastal Zone Sediment | Euclidean distance to surficial geology type: Alluvium and Fine-Textured Coastal Zone Sediment |
| Dist to Non-Carbonate Residual Material | Euclidean distance to surficial geology type: Non-Carbonate Residual Material |
| Dist to Sand Dunes | Euclidean distance to surficial geology type: Eolian Sediment, Coarse-Textured (Sand Dunes) |
| Dist to Silicic Residual Material | Euclidean distance to surficial geology type: Silicic Residual Material |
| Dist to fresh marsh | Distance to freshwater emergent wetland |
| Dist to fresh waters | Euclidean distance to nearest stream, river, or other inland waterbody (excluding estuaries) |
| Dist to lake or pond | Euclidean distance to nearest river or other inland waterbody (features represented by polygons) |
| Dist to stream poly | Euclidean distance to nearest stream (features represented by polygons) |
| Dist to woody wetland | Distance to forested palustrine wetland |
| Elevation | Elevation in decimeters (originally in meters) |
| Growing Degree Days | Growing Degree Days |
| Lower slope (flat) 10-cell mean | mean percent Lower slope (flat) in 10-cell mean (30 meter cells) |
| Lower slope (neutral) 10-cell mean | mean percent Lower slope (neutral) in 10-cell mean (30 meter cells) |
| Lower slope (warm) 10-cell mean | mean percent Lower slope (warm) in 10-cell mean (30 meter cells) |
| Mean Temperature of Coldest Quarter | Mean Temperature of Coldest Quarter |
| Mean Temperature of Warmest Quarter | Mean Temperature of Warmest Quarter |
| Precipitation of Driest Quarter | Precipitation of Driest Quarter |
| Precipitation of Wettest Quarter | Precipitation of Wettest Quarter |
| Roughness 10-cell circle | The standard deviation of elevation values within a circular neighborhood with a radius of 10 cells. |
| Topographic postion index 10-cell radius | Topographic position index using elevation values within a circular neighborhood with a radius of 10 cells. |
| Topographic postion index 100-cell radius | Topographic position index using elevation values within a circular neighborhood with a radius of 100 cells. |
| Upper slope (neutral) 10-cell mean | mean percent Upper slope (neutral) in 10-cell mean (30 meter cells) |
| Upper slope (warm) 1-cell mean | mean percent Upper slope (warm) in 1-cell mean (30 meter cells) |
| Upper slope (warm) 10-cell mean | mean percent Upper slope (warm) in 10-cell mean (30 meter cells) |
| Valley (narrow) 1-cell mean | mean percent Valley (narrow) in 1-cell mean (30 meter cells) |
| Valley (narrow) 10-cell mean | mean percent Valley (narrow) in 10-cell mean (30 meter cells) |
| Valley 1-cell mean | mean percent Valley in 1-cell mean (30 meter cells) |
| Valley 10-cell mean | mean percent Valley in 10-cell mean (30 meter cells) |
| Water cover 100-cell mean | mean open water cover within 100 cell radius |
| Woody wetlands 1-cell mean | mean woody wetland cover within 1 cell radius |
| Woody wetlands 10-cell mean | mean woody wetland cover within 10 cell radius |
| Woody wetlands 100-cell mean | mean woody wetland cover within 100 cell radius |

Appendix 2. Model details for reproducibility

- All R Scripts are available at github
- The repository version (repo head) used for this run was: 790073e450793d1cc18f224a967516b2b0ba3723
- The model run name was: lithpipi_20210720_063257
- R version: R version 4.1.0 (2021-05-18)
- Random seed for each model: 720063257

Table 1. Algorithm-specific details.

| Name | value |
|---|---|
| Algorithm = rf | |
| number of predictors used | 35 |
| mtry | 9 |
| number of trees | 2000 |
| type of trees | classification |
| Algorithm = xgb | |
| number of predictors used | 18 |
| iterations | 99 |
| eta | 0.3 |
| max depth | 3 |
| gamma | 0 |
| colsample by tree | 0.8 |
| min child weight | 1 |
| subsample | 0.5 |
| objective | binary:logistic |

Table 2. Reviewer star ratings for all model outputs of the accepted model version. More than one row for an algorithm indicates more than one reviewer. Average star ratings of 3 or more triggered the use of that algorithm in the final output.

| model version | algorithm | star rating |
|---|---|---|
| lithpipi_20210720_063257 | rf | 1 |
| lithpipi_20210720_063257 | xgb | 3 |