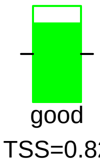
Phemeranthus piedmontanus



Species Distribution Model (SDM) assessment metrics and metadata
Common name: Piedmont Fameflower
Date: 06 Sep 2017
Code: phempied

ability to find new sites

This SDM incorporates the number of known and background locations indicated in Table 1, modeled with the random forests routine [1, 2] in the R statistical environment [3, 4]. We validated the model by jackknifing (also called leave-one-out, see [5, 6, 7]) by element occurrence for a total of 11 groups. The statistics in Table 2 report the mean and variance for these jackknifing runs.

Table 1. Input statistics. Polys = input polygons; EOs = element occurrences (known locations); BG points = background points placed throughout study area excluding known species locations; PR points = presence points placed throughout all polygons.

Name	Number
polys	17
EOs	11
BG points	27166
PR points	95

Table 2. Validation statistics for jackknife trials. Overall Accuracy = Correct Classification Rate, TSS = True Skill Statistic, AUC = area under the ROC curve; see [6, 8, 9].

Name	Mean	SD	SEM
Overall Accuracy	0.91	0.20	0.06
Specificity	0.82	0.40	0.12
Sensitivity	1.00	0.00	0.00
TSS	0.82	0.40	0.12
Kappa	0.82	0.40	0.12
AUC	0.91	0.30	0.09

Validation runs used 58 environmental variables, the most important of 78 variables (top 75 percent). Each tree was built with 2 variables tried at each split (mtry) and 1000 trees built. The final model was built using 2000 trees, all presence and background points, with an mtry of 2, and the same number of environmental variables.

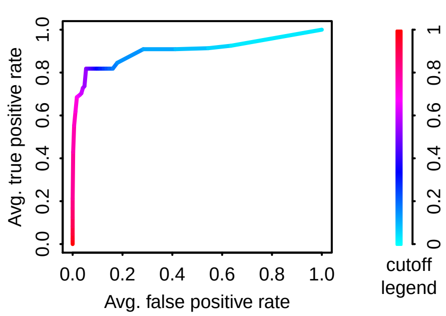


Figure 1. ROC plot for all 11 validation runs, averaged along cutoffs.

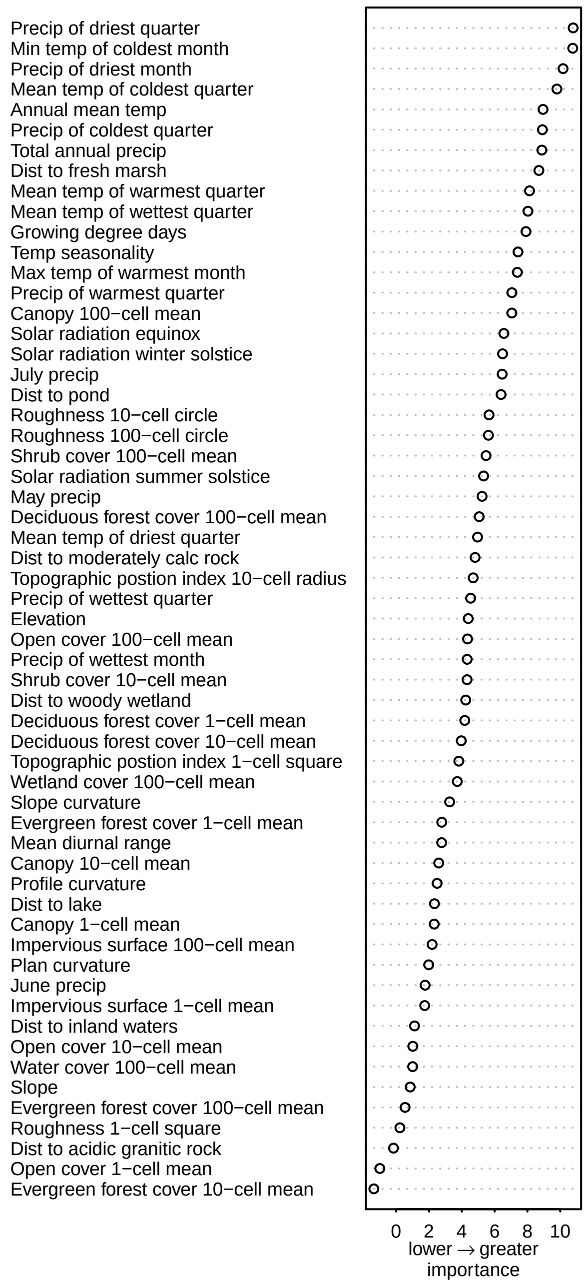


Figure 2. Relative importance of each environmental variable based on the full model using all background and presence points as input. Abbreviations used: calc = calcareous, CP = coastal plain, dist = distance, fresh = freshwater, precip = precipitation, temp = temperature, max = maximum, min = minimum.

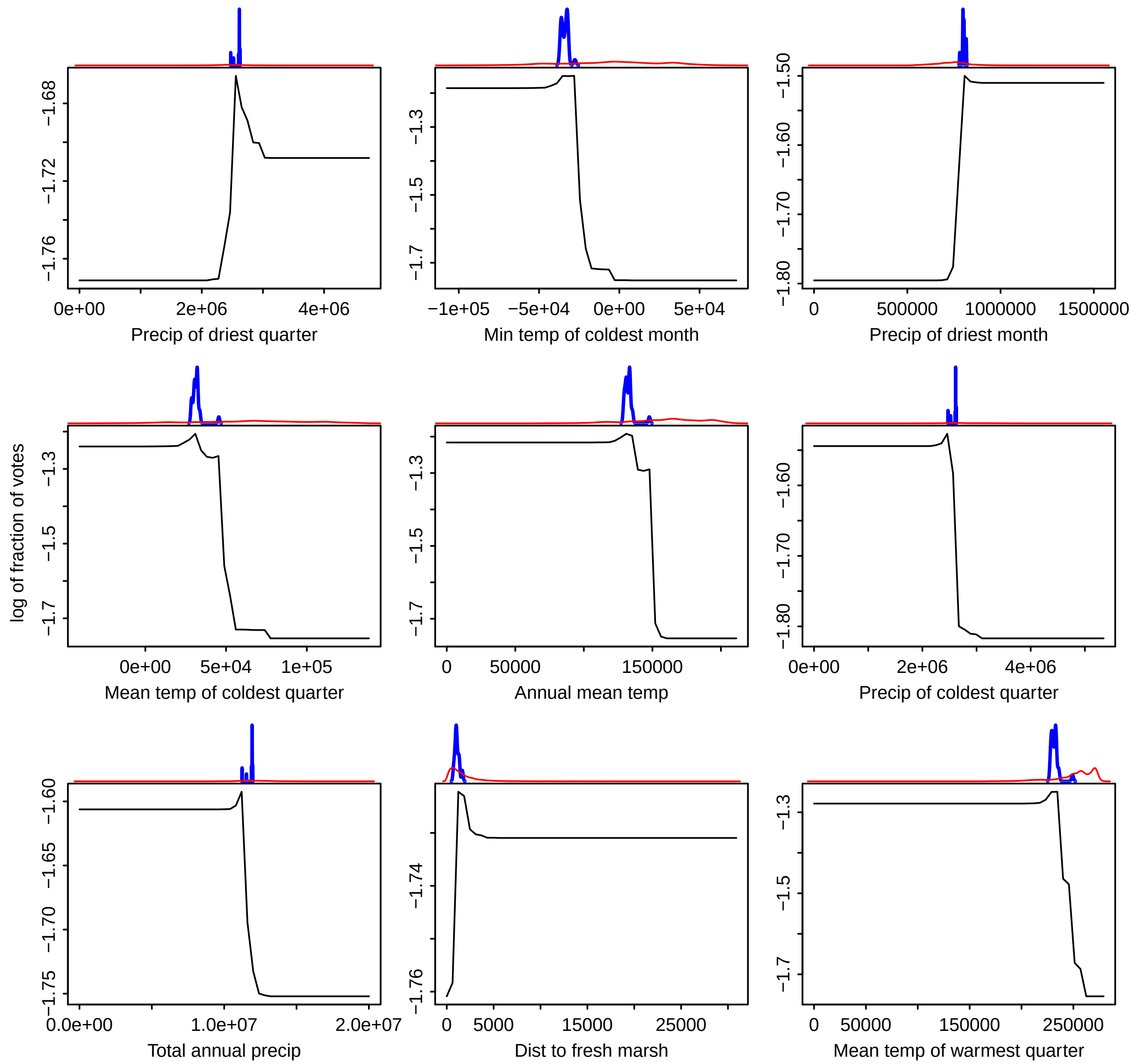


Figure 3. Partial dependence plots for the 9 environmental variables with the most influence on the model. Each plot shows the effect of the variable on the probability of appropriate habitat with the effects of the other variables removed [3]. Peaks in the line indicate where this variable had the strongest influence on predicting appropriate habitat. The distribution of each category (thin red = BG points, thick blue = PR points) is depicted at the top margin.

Element distribution models map places of similar environmental conditions to the submitted locations (PR points). 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. 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 land area may be appropriate to use in this case. For a more conservative depiction of suitable habitat that shows less land area, a higher threshold may be more appropriate. Different thresholds for this model (full model) are described in Table 3.

Table 3. Thresholds calculated from the final model. For discussions of these different thresholds see [11, 12]. The Value column reports the threshold; EOs indicates the percentage (number in brackets) of EOs within which at least one point was predicted as suitable habitat; Polys indicates the percentage (number) of polygons within which at least one point was predicted as having suitable habitat; Pts indicates the percentage of PR points predicted having suitable habitat. Total numbers of EOs, polygons, and PR points used in the final model are reported in Table 1.

Threshold	Value	EOs	Polys	Pts	Description
Equal sensitivity and specificity	0.824	100(11)	100(17)	100	The probability at which the absolute value of the difference between sensitivity and specificity is minimized.
Maximum of sensitivity plus specificity	0.824	100(11)	100(17)	100	The probability at which the sum of sensitivity and specificity is maximized.
Minimum Training Presence	0.824	100(11)	100(17)	100	The highest probability value at which 100% of input presence points remain classified as suitable habitat.
Minimum Training Presence by Polygon	0.950	100(11)	100(17)	94.7	The highest probability value at which 100% of input polygons have at least one presence point classified as suitable habitat.
Minimum Training Presence by Element Occurrence	0.973	100(11)	82.4(14)	71.6	The highest probability value at which 100% of input EOs have at least one presence point classified as suitable habitat.
Tenth percentile of training presence	0.960	100(11)	94.1(16)	89.5	The probability at which 90% of the input presence points are classified as suitable habitat.
F-measure with alpha set to 0.01	0.824	100(11)	100(17)	100	The probability value at which the harmonic mean of precision and recall, with strong weighting towards recall, is maximized.

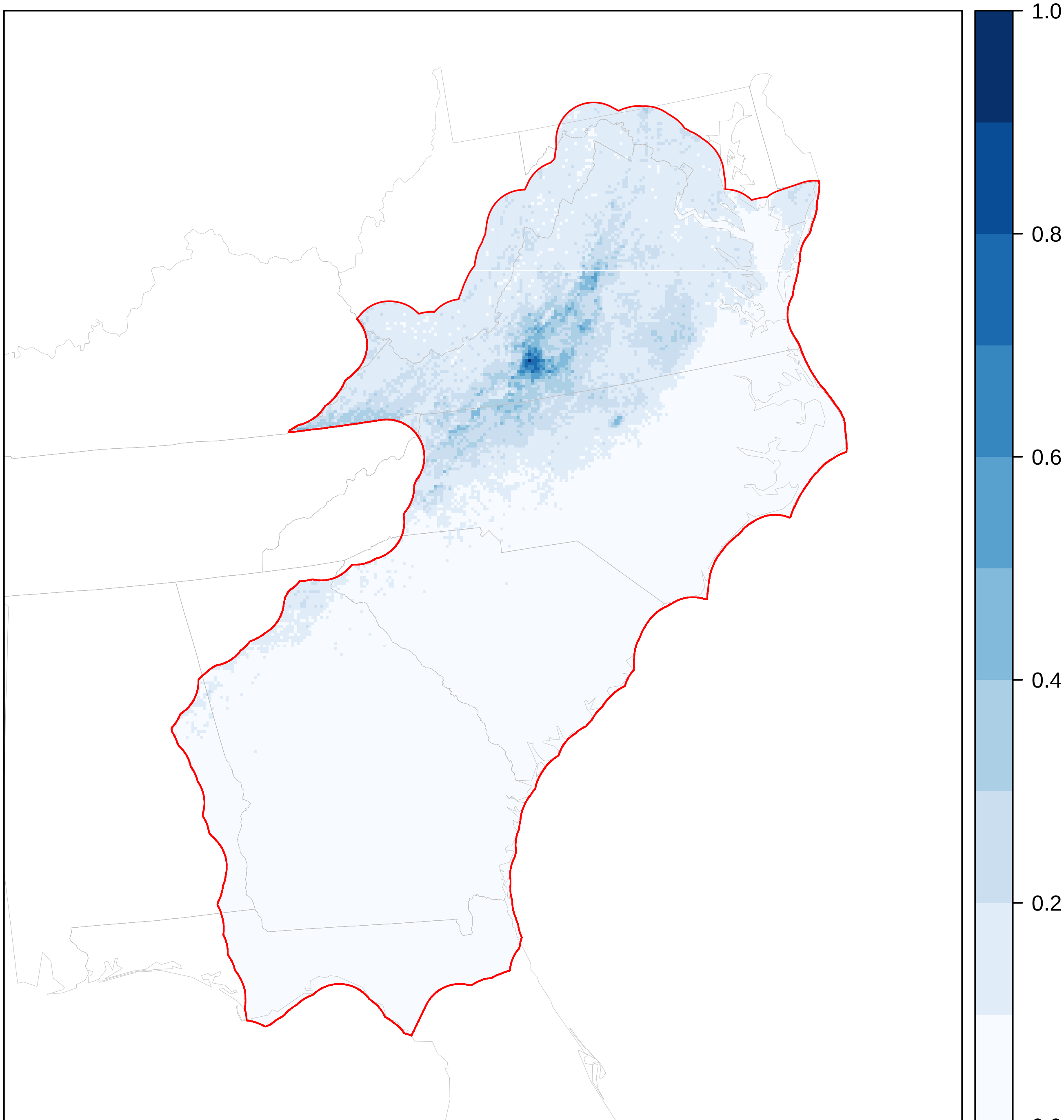


Figure 5. A generalized view of the model predictions throughout the study area. State boundaries are shown in gray. The study area is outlined in red.

This distribution model would not have been possible without data sharing among organizations. The following organizations provided data:

- North Carolina Natural Heritage Program
- Virginia Natural Heritage Program

This model was built using a methodology developed through collaboration among the Florida Natural Areas Inventory, New York Natural Heritage Program, Pennsylvania Natural Heritage Program, and Virginia Natural Heritage Program. It is one of a suite of distribution models developed using the same methods, the same scripts, and the same environmental data sets. Our goal was to be consistent and transparent in our methodology, validation, and output. This work was supported by the US Fish and Wildlife Service, and the South Atlantic Landscape Conservation Cooperative.

Please cite this document and its associated SDM as:
Virginia Natural Heritage Program. 2017. Species distribution model for Piedmont Fameflower (*Phemeranthus piedmontanus*). Created on 06 Sep 2017. Virginia Department of Conservation and Recreation - Division of Natural Heritage, Richmond, VA.

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 4.6-12.
[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 3.3.3 (2017-03-06).
[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.