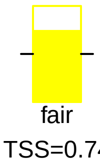
Anatrytone logan

Species Distribution Model (SDM) assessment metrics and metadata
Common name: Delaware Skipper
Date: 17 Nov 2017
Code: anatloga



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 43 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); Groups = element occurrence BG points = background points; PR points = presence points placed throughout all polygons.

Name	Number
polys	46
EOs	43
BG points	11473
PR points	2550

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

Name	Mean	SD	SEM
Overall Accuracy	0.87	0.17	0.03
Specificity	0.86	0.32	0.05
Sensitivity	0.88	0.10	0.02
TSS	0.74	0.33	0.05
Kappa	0.74	0.33	0.05
AUC	0.94	0.11	0.02

Validation runs used 60 environmental variables, the most important of 89 variables (top 75 percent). Each tree was built with 2 variables tried at each split (mtry) and 750 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.

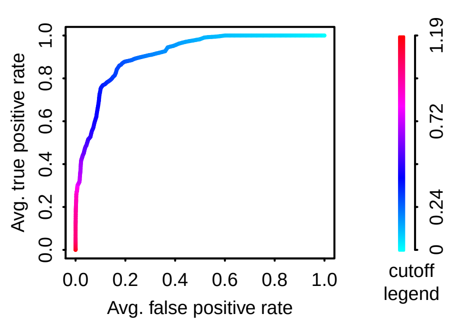


Figure 1. ROC plot for all 43 validation runs, averaged along cutoffs.

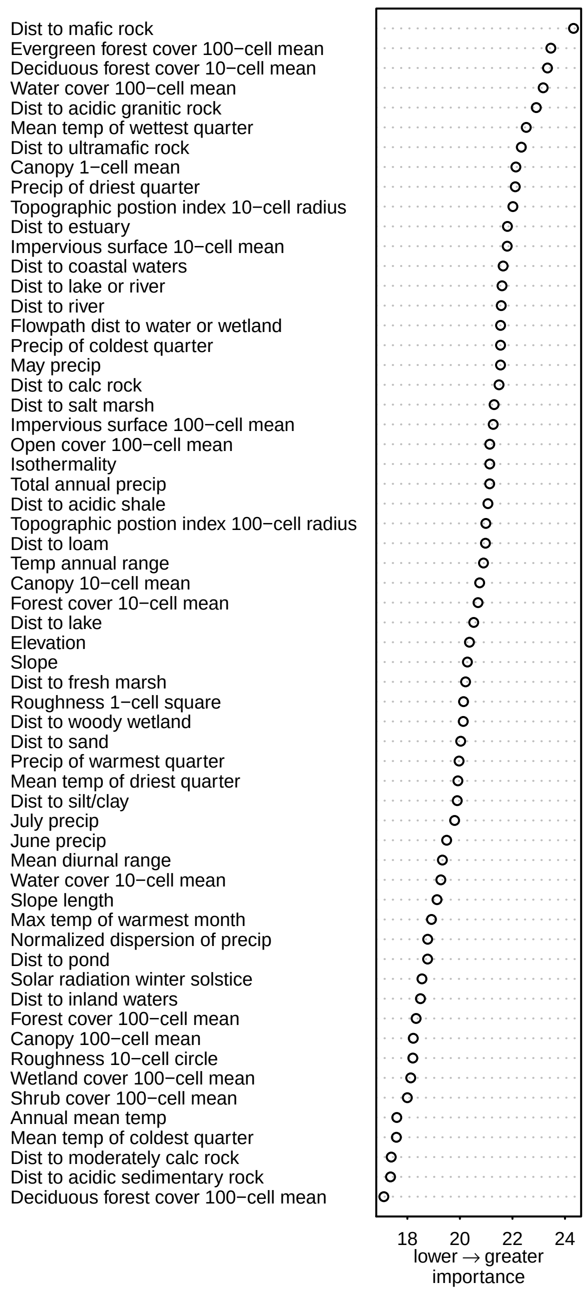


Figure 2. Relative importance of each environmental variable based on the full model using all sites as input. Abbreviations used: calc = calcareous, CP = coastal plain, dist = distance, fresh = freshwater, precip = precipitation, temp = temperature, max = maximum, min = minimum.

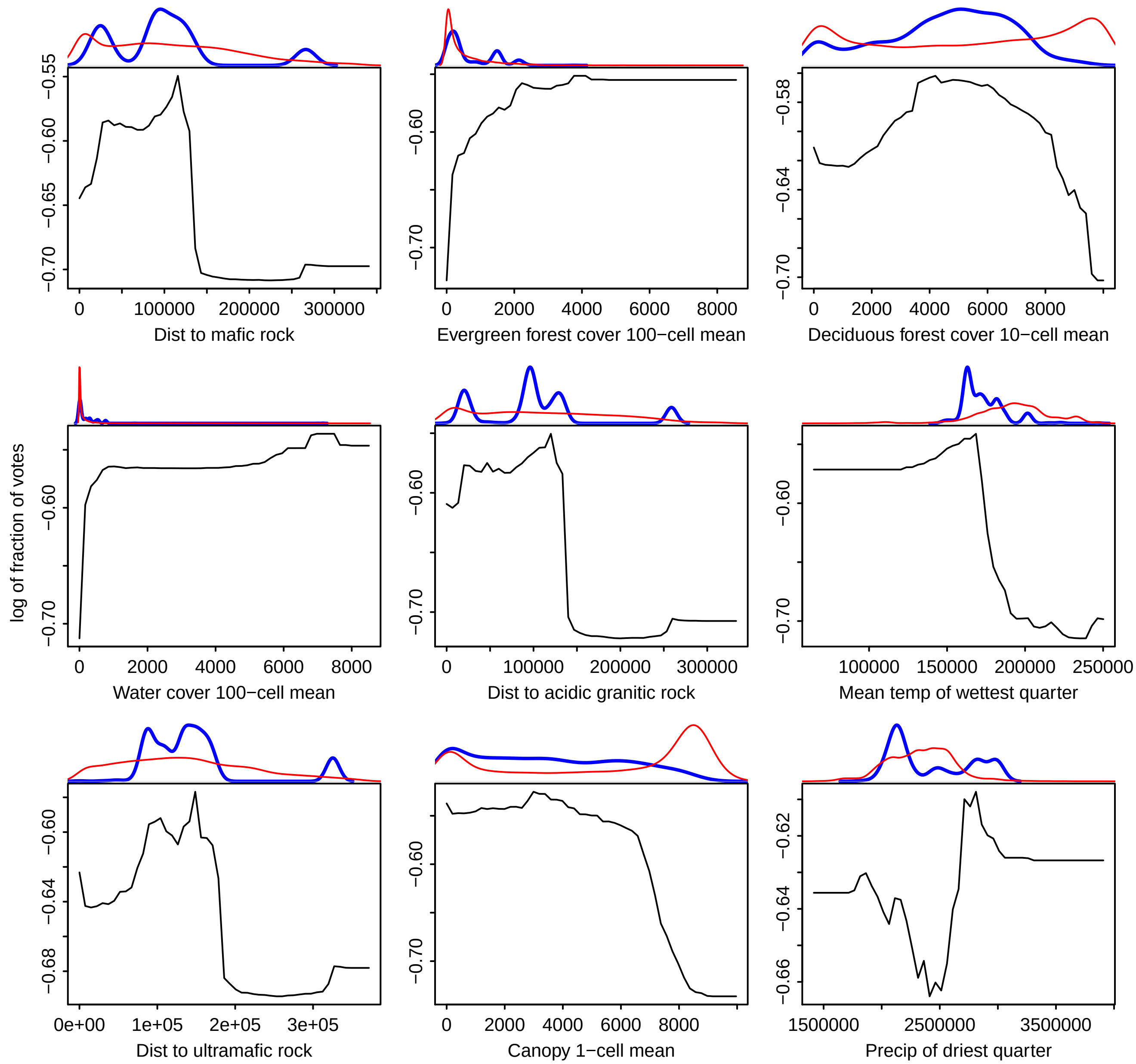


Figure 3. Partial dependence plots for the 9 environmental variables with the most influence on the model. Each plot shows the effect of the variable on the probability of appropriate habitat with the effects of the other variables removed [3]. Peaks in the line indicate where this variable had the strongest 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	0.588	100(43)	100(46)	99.5	The probability at which the absolute value of sensitivity minus specificity is minimized.
F-measure with alpha set to 0.01	0.562	100(43)	100(46)	100	The harmonic average of precision and recall, with strong weighting towards recall (classifying presence points as suitable habitat).
Maximum of sensitivity plus specificity	0.562	100(43)	100(46)	100	The probability at which the sum of sensitivity (true positive rate) and specificity (true negative rate) is maximized.
Minimum Training Presence	0.562	100(43)	100(46)	100	The lowest probability value assigned to any of the input presence points. 100% of input presence points are classified as suitable habitat.
Minimum Training Presence by Element Occurrence	0.836	100(43)	100(46)	71	The lowest probability value assigned to any of the input presence element occurrences. This calculation first summarizes EOs by their maximum and then finds the minimum of these values.
Minimum Training Presence by Polygon	0.836	100(43)	100(46)	71	The lowest probability value assigned to any of the input presence polygons.
Tenth percentile of training presence	0.715	100(43)	100(46)	90	The probability at which 90% of the input presence points are classified as suitable habitat and 10% are classified as unsuitable.

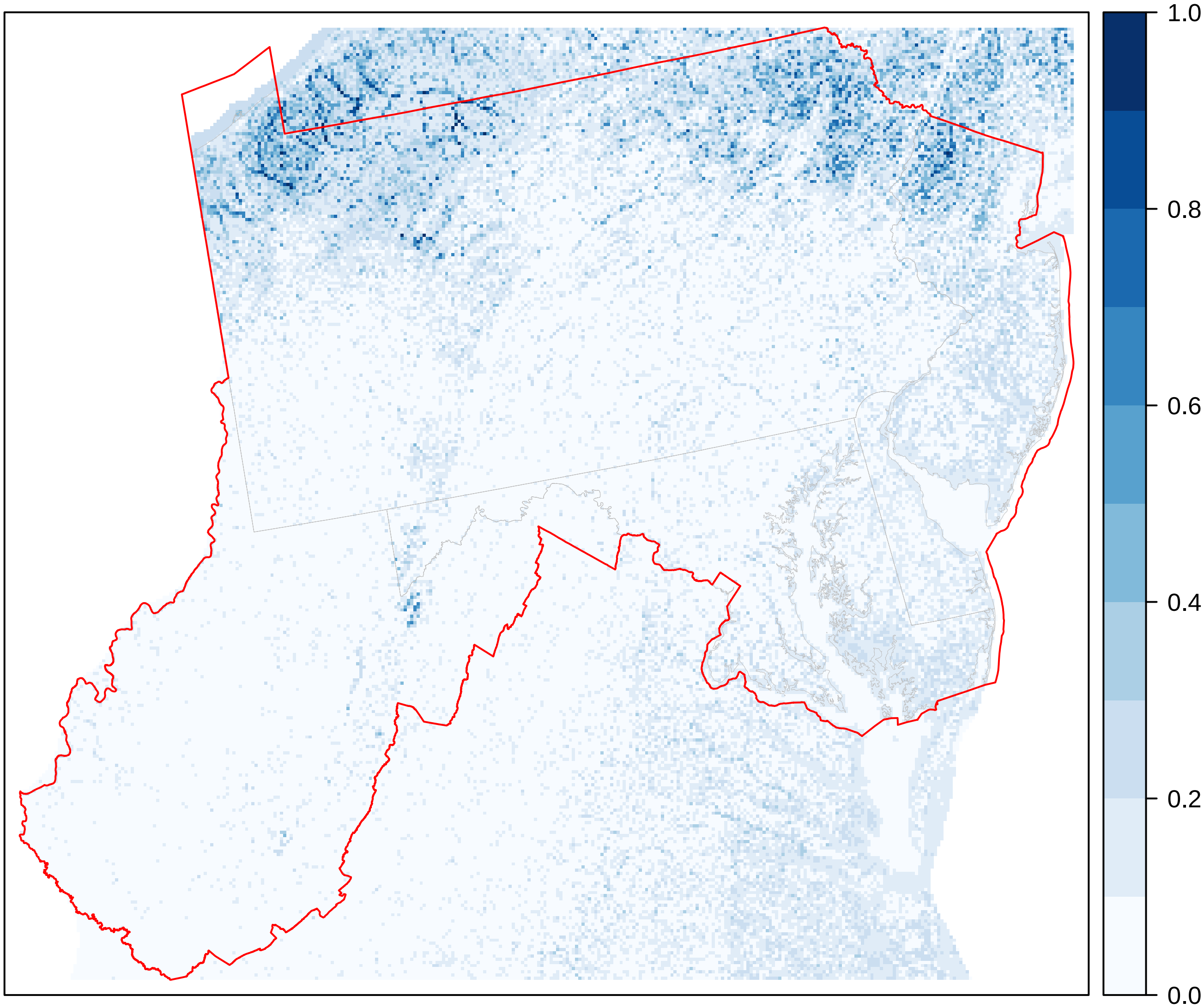


Figure 5. A generalized view of the model predictions throughout the study area. State boundaries are shown in black. The study area is outlined in red.

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

- Maryland Natural Heritage Program, Maryland Department of Natural Resources, Wildlife and Heritage Service
- Pennsylvania Natural Heritage Program
- West 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.

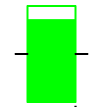
Please cite this document and its associated SDM as:
Pennsylvania Natural Heritage Program. 2017. Species distribution model for Delaware Skipper (*Anatrytone logan*). Created on 17 Nov 2017. Western Pennsylvania Conservancy, Pittsburgh, PA.

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 3.4.1 (2017-06-30).
[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.

Boloria selene myrina

Species Distribution Model (SDM) assessment metrics and metadata
Common name: Silver-bordered Fritillary
Date: 30 Jan 2018
Code: bolosele



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 57 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); Groups = element occurrence BG points = background points; PR points = presence points placed throughout all polygons.

Name	Number
polys	211
EOs	57
BG points	11473
PR points	13471

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

Name	Mean	SD	SEM
Overall Accuracy	0.93	0.10	0.01
Specificity	0.93	0.21	0.03
Sensitivity	0.92	0.04	0.00
TSS	0.85	0.20	0.03
Kappa	0.85	0.20	0.03
AUC	0.98	0.05	0.01

Validation runs used 60 environmental variables, the most important of 89 variables (top 75 percent). Each tree was built with 2 variables tried at each split (mtry) and 750 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.

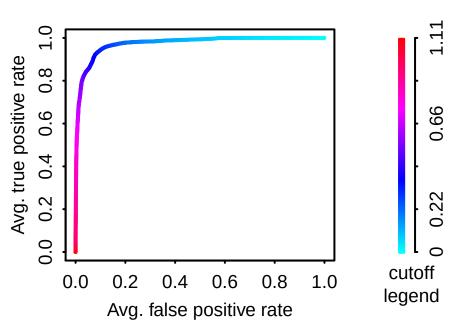


Figure 1. ROC plot for all 57 validation runs, averaged along cutoffs.

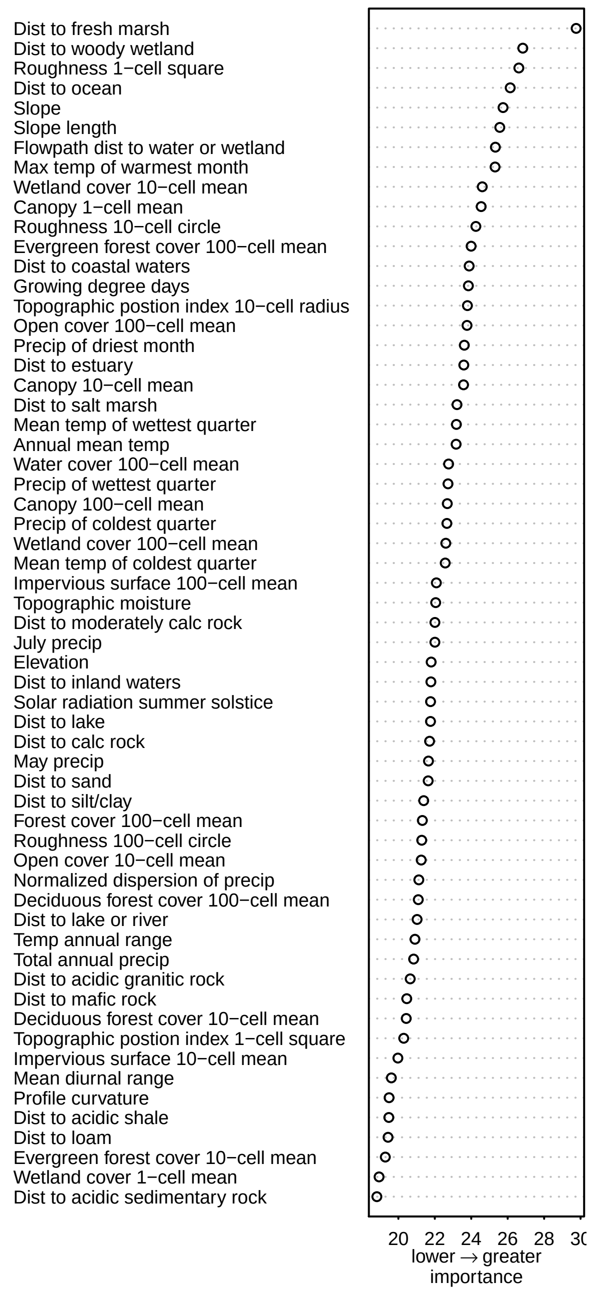


Figure 2. Relative importance of each environmental variable based on the full model using all sites as input. Abbreviations used: calc = calcareous, CP = coastal plain, dist = distance, fresh = freshwater, precip = precipitation, temp = temperature, max = maximum, min = minimum.

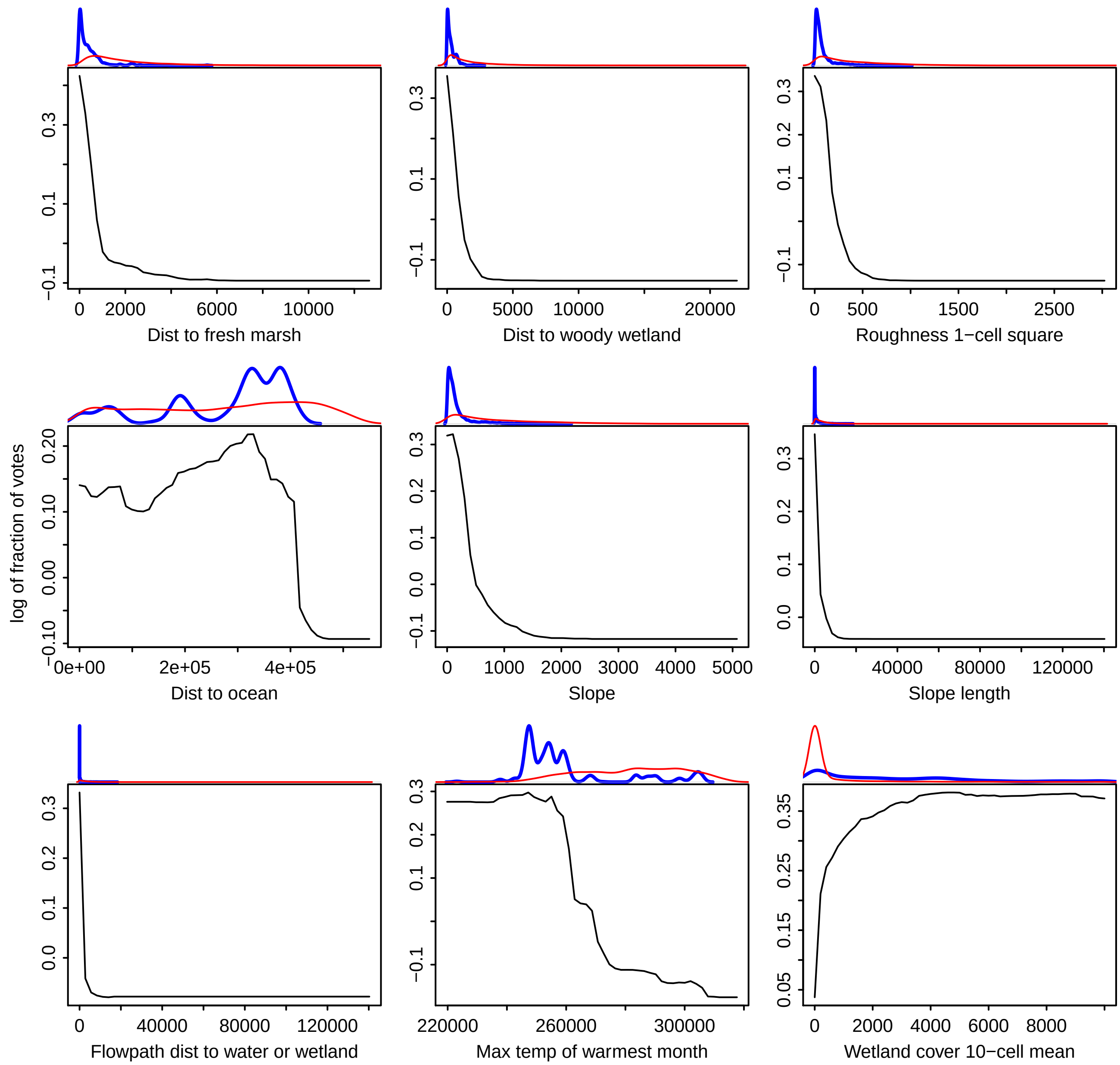


Figure 3. Partial dependence plots for the 9 environmental variables with the most influence on the model. Each plot shows the effect of the variable on the probability of appropriate habitat with the effects of the other variables removed [3]. Peaks in the line indicate where this variable had the strongest 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	0.603	100(57)	99.5(210)	98.9	The probability at which the absolute value of sensitivity minus specificity is minimized.
F-measure with alpha set to 0.01	0.409	100(57)	100(211)	100	The harmonic average of precision and recall, with strong weighting towards recall (classifying presence points as suitable habitat).
Maximum of sensitivity plus specificity	0.568	100(57)	100(211)	99.6	The probability at which the sum of sensitivity (true positive rate) and specificity (true negative rate) is maximized.
Minimum Training Presence	0.409	100(57)	100(211)	100	The lowest probability value assigned to any of the input presence points. 100% of input presence points are classified as suitable habitat.
Minimum Training Presence by Element Occurrence	0.900	100(57)	80.1(169)	71.4	The lowest probability value assigned to any of the input presence element occurrences. This calculation first summarizes EOs by their maximum and then finds the minimum of these values.
Minimum Training Presence by Polygon	0.569	100(57)	100(211)	99.6	The lowest probability value assigned to any of the input presence polygons.
Tenth percentile of training presence	0.788	100(57)	96.2(203)	90	The probability at which 90% of the input presence points are classified as suitable habitat and 10% are classified as unsuitable.

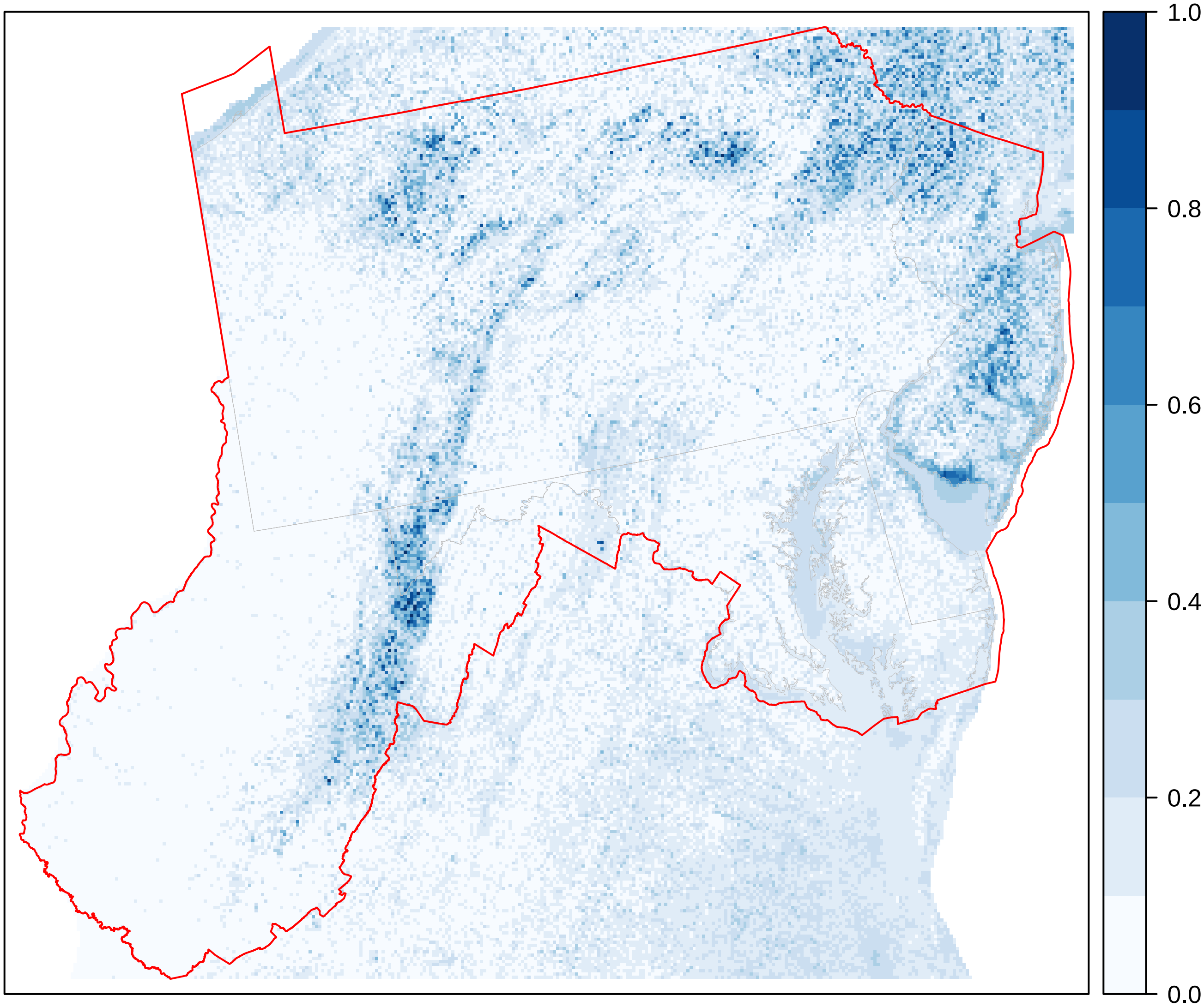


Figure 5. A generalized view of the model predictions throughout the study area. State boundaries are shown in black. The study area is outlined in red.

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

- Maryland Natural Heritage Program, Maryland Department of Natural Resources, Wildlife and Heritage Service
- New Jersey Department of Environmental Protection, Division of Fish and Wildlife, New Jersey Endangered & Nongame Species Program
- Pennsylvania Natural Heritage Program
- West 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.

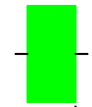
Please cite this document and its associated SDM as:
Pennsylvania Natural Heritage Program. 2018. Species distribution model for Silver-bordered Fritillary (*Boloria selene myrina*). Created on 30 Jan 2018. Western Pennsylvania Conservancy, Pittsburgh, PA.

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 3.4.3 (2017-11-30).
[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.

Carterocephalus palaemon

Species Distribution Model (SDM) assessment metrics and metadata
Common name: Arctic Skipper
Date: 18 Nov 2017
Code: cartpala



good
TSS=0.98
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 7 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); Groups = element occurrence BG points = background points; PR points = presence points placed throughout all polygons.

Name	Number
polys	12
EOs	7
BG points	11473
PR points	1727

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

Name	Mean	SD	SEM
Overall Accuracy	0.99	0.01	0.00
Specificity	1.00	0.00	0.00
Sensitivity	0.98	0.02	0.01
TSS	0.98	0.01	0.01
Kappa	0.98	0.01	0.01
AUC	1.00	0.00	0.00

Validation runs used 54 environmental variables, the most important of 81 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.

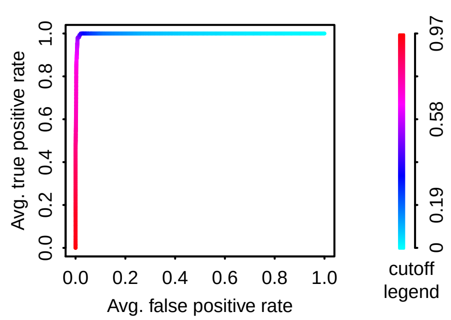


Figure 1. ROC plot for all 7 validation runs, averaged along cutoffs.

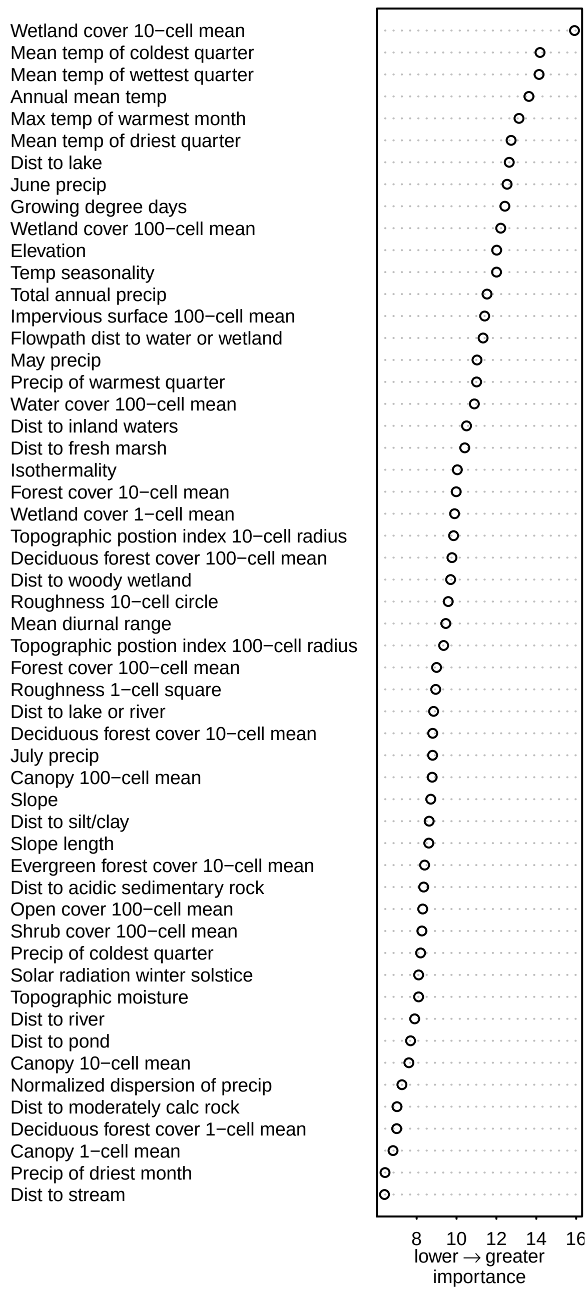


Figure 2. Relative importance of each environmental variable based on the full model using all sites as input. Abbreviations used: calc = calcareous, CP = coastal plain, dist = distance, fresh = freshwater, precip = precipitation, temp = temperature, max = maximum, min = minimum.

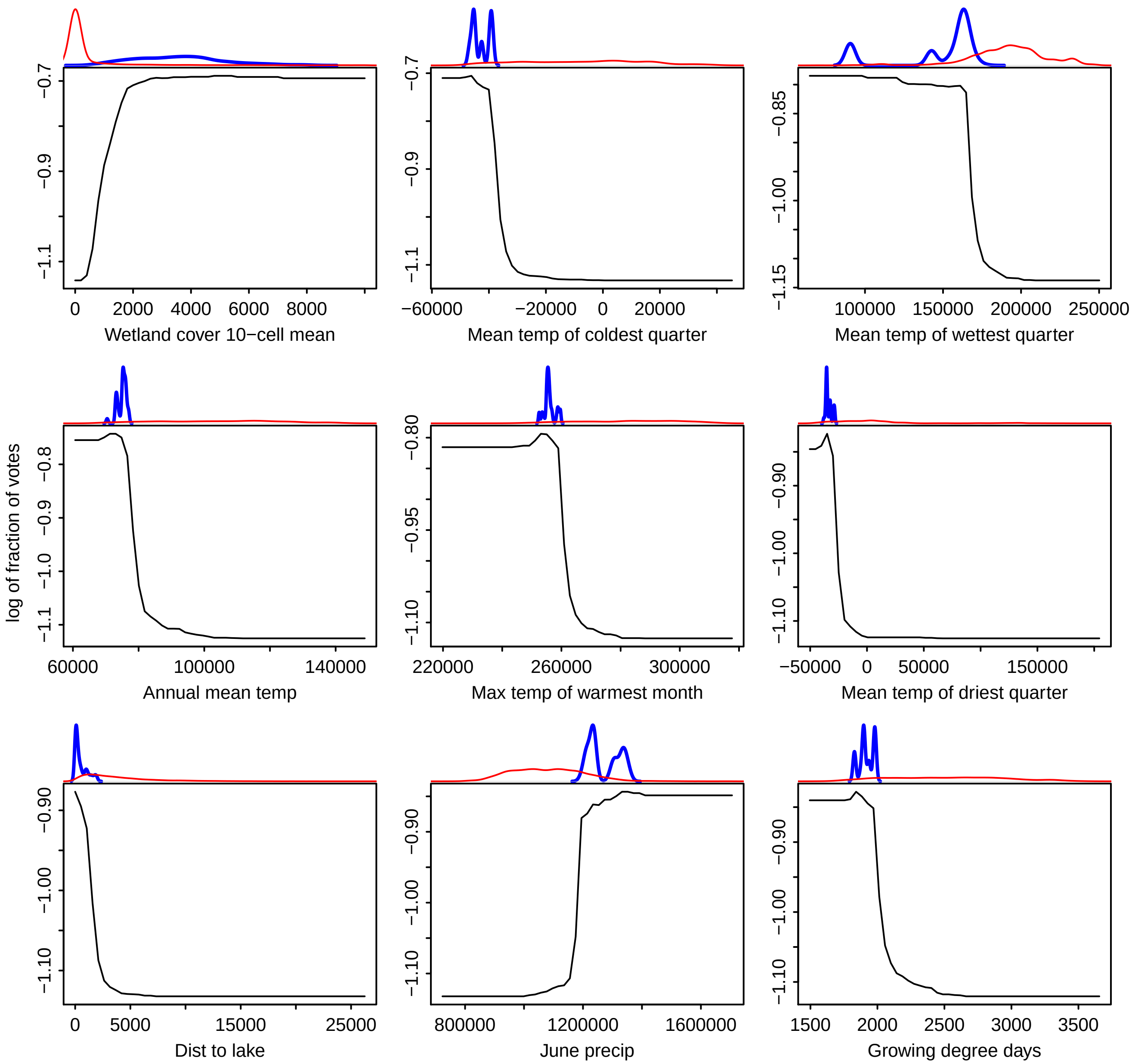


Figure 3. Partial dependence plots for the 9 environmental variables with the most influence on the model. Each plot shows the effect of the variable on the probability of appropriate habitat with the effects of the other variables removed [3]. Peaks in the line indicate where this variable had the strongest 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	0.658	100(7)	100(12)	99.8	The probability at which the absolute value of sensitivity minus specificity is minimized.
F-measure with alpha set to 0.01	0.635	100(7)	100(12)	100	The harmonic average of precision and recall, with strong weighting towards recall (classifying presence points as suitable habitat).
Maximum of sensitivity plus specificity	0.635	100(7)	100(12)	100	The probability at which the sum of sensitivity (true positive rate) and specificity (true negative rate) is maximized.
Minimum Training Presence	0.635	100(7)	100(12)	100	The lowest probability value assigned to any of the input presence points. 100% of input presence points are classified as suitable habitat.
Minimum Training Presence by Element Occurrence	0.985	100(7)	58.3(7)	19.6	The lowest probability value assigned to any of the input presence element occurrences. This calculation first summarizes EOs by their maximum and then finds the minimum of these values.
Minimum Training Presence by Polygon	0.931	100(7)	100(12)	72.6	The lowest probability value assigned to any of the input presence polygons.
Tenth percentile of training presence	0.857	100(7)	100(12)	90	The probability at which 90% of the input presence points are classified as suitable habitat and 10% are classified as unsuitable.

```
## Error in nrow(sdm.customComments.subset):  object 'sdm.customComments.subset' not found
```

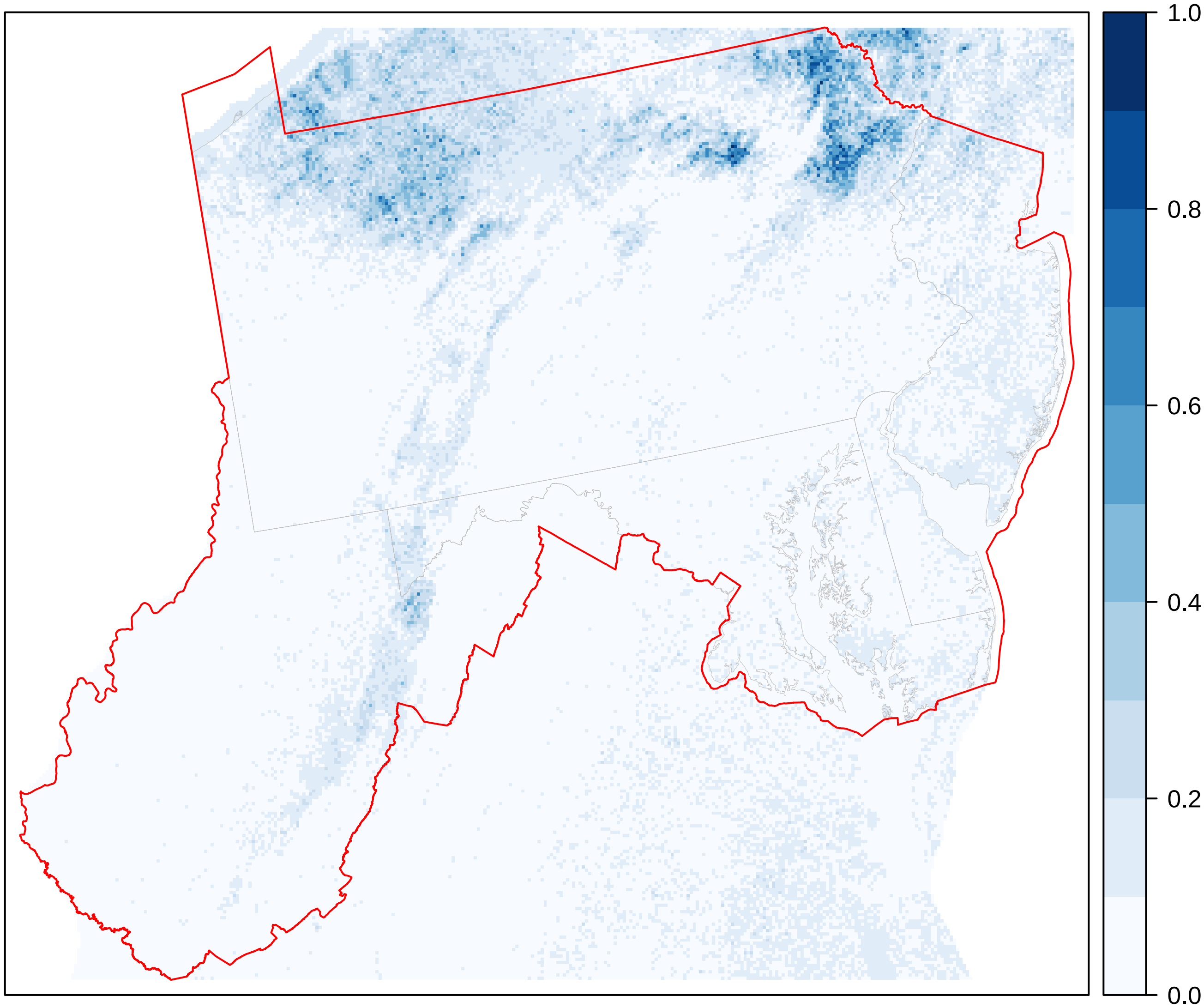


Figure 5. A generalized view of the model predictions throughout the study area. State boundaries are shown in black. The study area is outlined in red.

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

- New Jersey Department of Environmental Protection, Division of Fish and Wildlife, New Jersey Endangered & Nongame Species Program
- Pennsylvania 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.

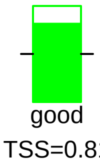
Please cite this document and its associated SDM as:
Pennsylvania Natural Heritage Program. 2017. Species distribution model for Arctic Skipper (*Carterocephalus palaemon*). Created on 18 Nov 2017. Western Pennsylvania Conservancy, Pittsburgh, PA.

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 3.4.1 (2017-06-30).
[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.

Chlosyne harrisii

Species Distribution Model (SDM) assessment metrics and metadata
Common name: Harris' Checkerspot
Date: 30 Jan 2018
Code: chloharr



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 55 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); Groups = element occurrence BG points = background points; PR points = presence points placed throughout all polygons.

Name	Number
polys	78
EOs	55
BG points	11472
PR points	4480

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

Name	Mean	SD	SEM
Overall Accuracy	0.91	0.12	0.02
Specificity	0.94	0.22	0.03
Sensitivity	0.88	0.09	0.01
TSS	0.81	0.24	0.03
Kappa	0.81	0.24	0.03
AUC	0.98	0.05	0.01

Validation runs used 57 environmental variables, the most important of 85 variables (top 75 percent). Each tree was built with 2 variables tried at each split (mtry) and 750 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.

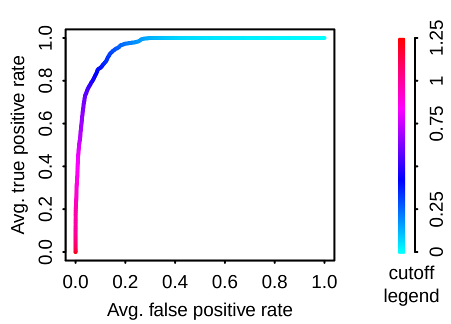


Figure 1. ROC plot for all 55 validation runs, averaged along cutoffs.

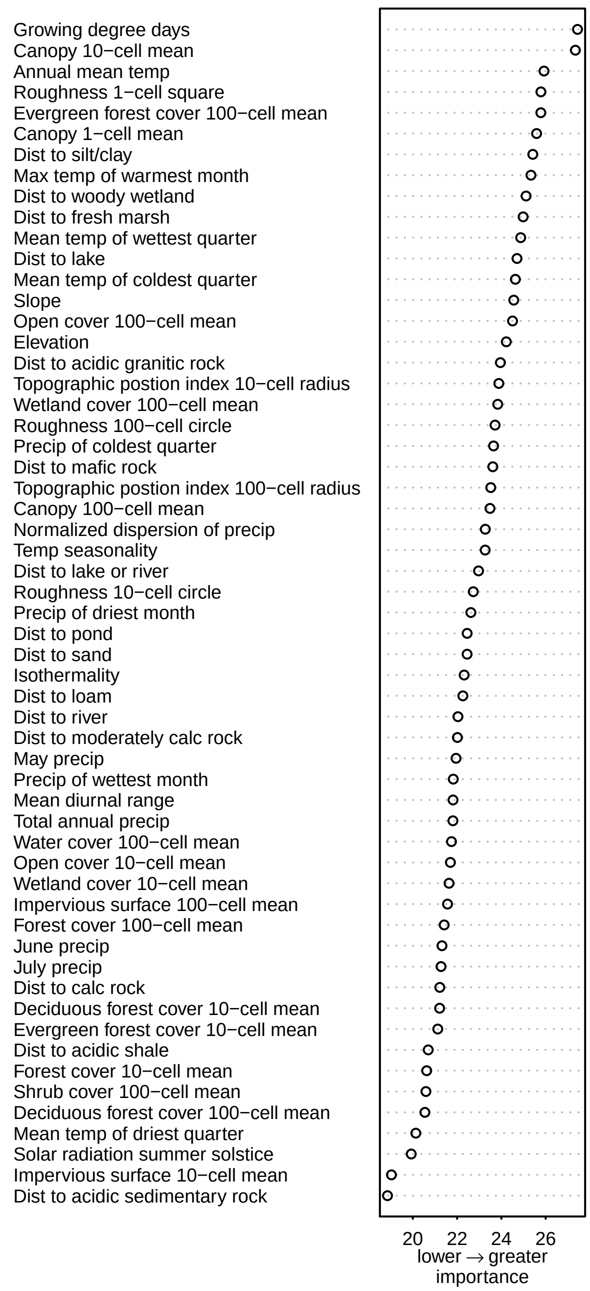


Figure 2. Relative importance of each environmental variable based on the full model using all sites as input. Abbreviations used: calc = calcareous, CP = coastal plain, dist = distance, fresh = freshwater, precip = precipitation, temp = temperature, max = maximum, min = minimum.

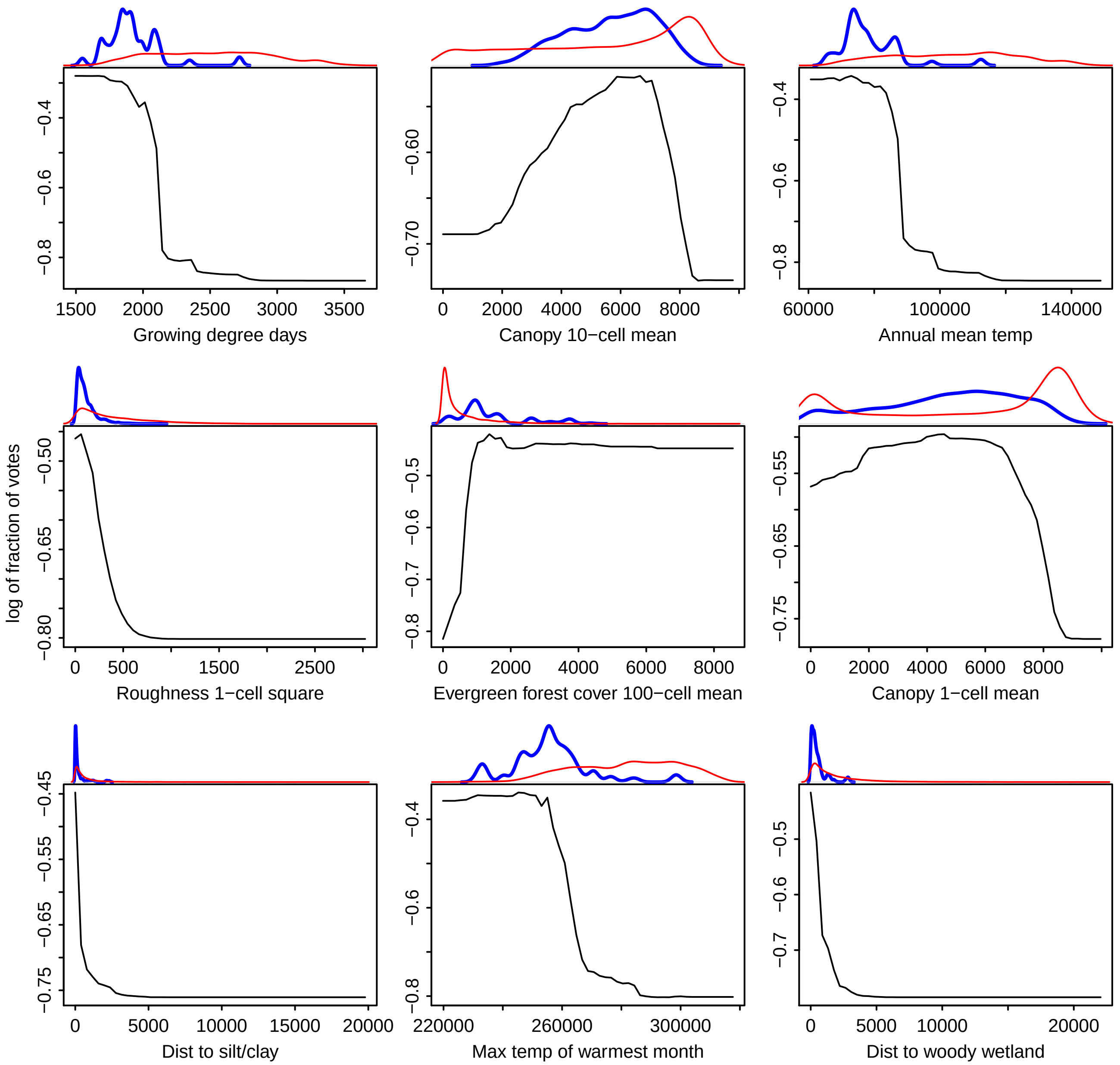


Figure 3. Partial dependence plots for the 9 environmental variables with the most influence on the model. Each plot shows the effect of the variable on the probability of appropriate habitat with the effects of the other variables removed [3]. Peaks in the line indicate where this variable had the strongest 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	0.669	100(55)	100(78)	99.4	The probability at which the absolute value of sensitivity minus specificity is minimized.
F-measure with alpha set to 0.01	0.529	100(55)	100(78)	100	The harmonic average of precision and recall, with strong weighting towards recall (classifying presence points as suitable habitat).
Maximum of sensitivity plus specificity	0.685	100(55)	100(78)	99.3	The probability at which the sum of sensitivity (true positive rate) and specificity (true negative rate) is maximized.
Minimum Training Presence	0.529	100(55)	100(78)	100	The lowest probability value assigned to any of the input presence points. 100% of input presence points are classified as suitable habitat.
Minimum Training Presence by Element Occurrence	0.929	100(55)	88.5(69)	66.7	The lowest probability value assigned to any of the input presence element occurrences. This calculation first summarizes EOs by their maximum and then finds the minimum of these values.
Minimum Training Presence by Polygon	0.697	100(55)	100(78)	99.1	The lowest probability value assigned to any of the input presence polygons.
Tenth percentile of training presence	0.864	100(55)	96.2(75)	90	The probability at which 90% of the input presence points are classified as suitable habitat and 10% are classified as unsuitable.

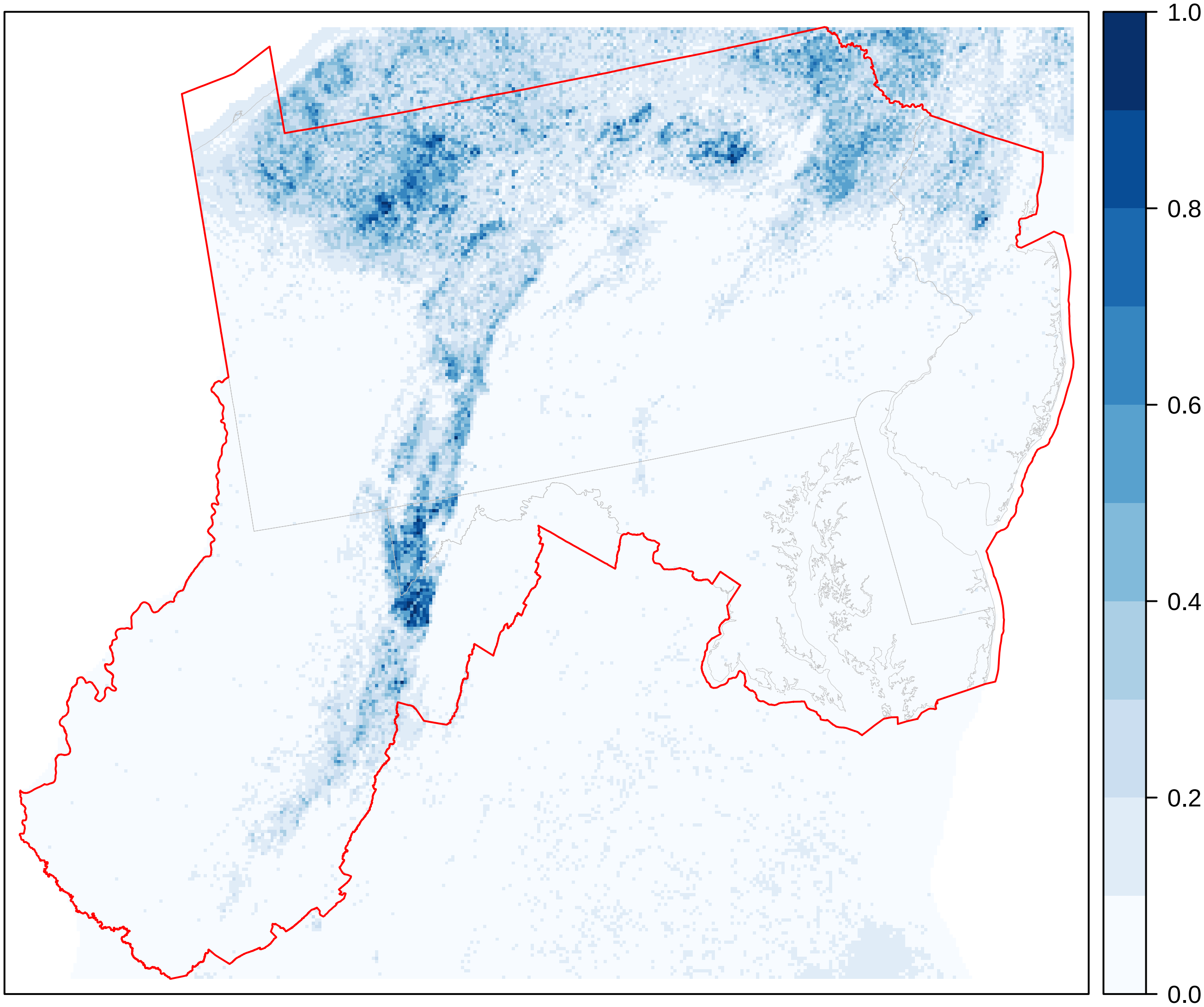


Figure 5. A generalized view of the model predictions throughout the study area. State boundaries are shown in black. The study area is outlined in red.

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

- Maryland Natural Heritage Program, Maryland Department of Natural Resources, Wildlife and Heritage Service
- New Jersey Department of Environmental Protection, Division of Fish and Wildlife, New Jersey Endangered & Nongame Species Program
- Pennsylvania Natural Heritage Program
- West 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.

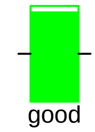
Please cite this document and its associated SDM as:
Pennsylvania Natural Heritage Program. 2018. Species distribution model for Harris' Checkerspot (*Chlosyne harrisii*). Created on 30 Jan 2018. Western Pennsylvania Conservancy, Pittsburgh, PA.

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 3.4.3 (2017-11-30).
[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.

Euphyes bimacula

Species Distribution Model (SDM) assessment metrics and metadata
Common name: Two-spotted Skipper
Date: 30 Jan 2018
Code: euphbima



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 27 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); Groups = element occurrence BG points = background points; PR points = presence points placed throughout all polygons.

Name	Number
polys	28
EOs	27
BG points	11472
PR points	2403

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

Name	Mean	SD	SEM
Overall Accuracy	0.97	0.10	0.02
Specificity	0.95	0.19	0.04
Sensitivity	0.98	0.02	0.00
TSS	0.93	0.19	0.04
Kappa	0.93	0.19	0.04
AUC	0.99	0.02	0.00

Validation runs used 58 environmental variables, the most important of 86 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.

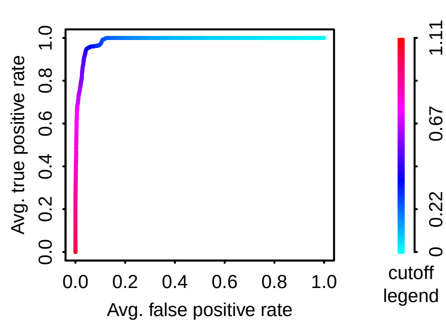


Figure 1. ROC plot for all 27 validation runs, averaged along cutoffs.

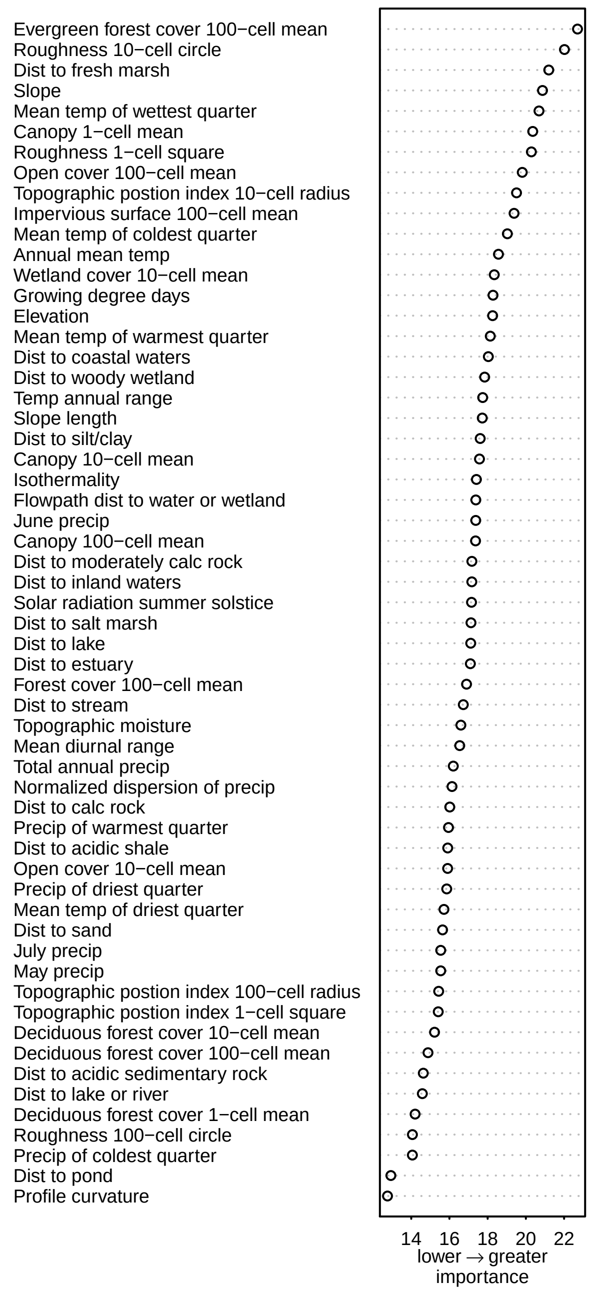


Figure 2. Relative importance of each environmental variable based on the full model using all sites as input. Abbreviations used: calc = calcareous, CP = coastal plain, dist = distance, fresh = freshwater, precip = precipitation, temp = temperature, max = maximum, min = minimum.

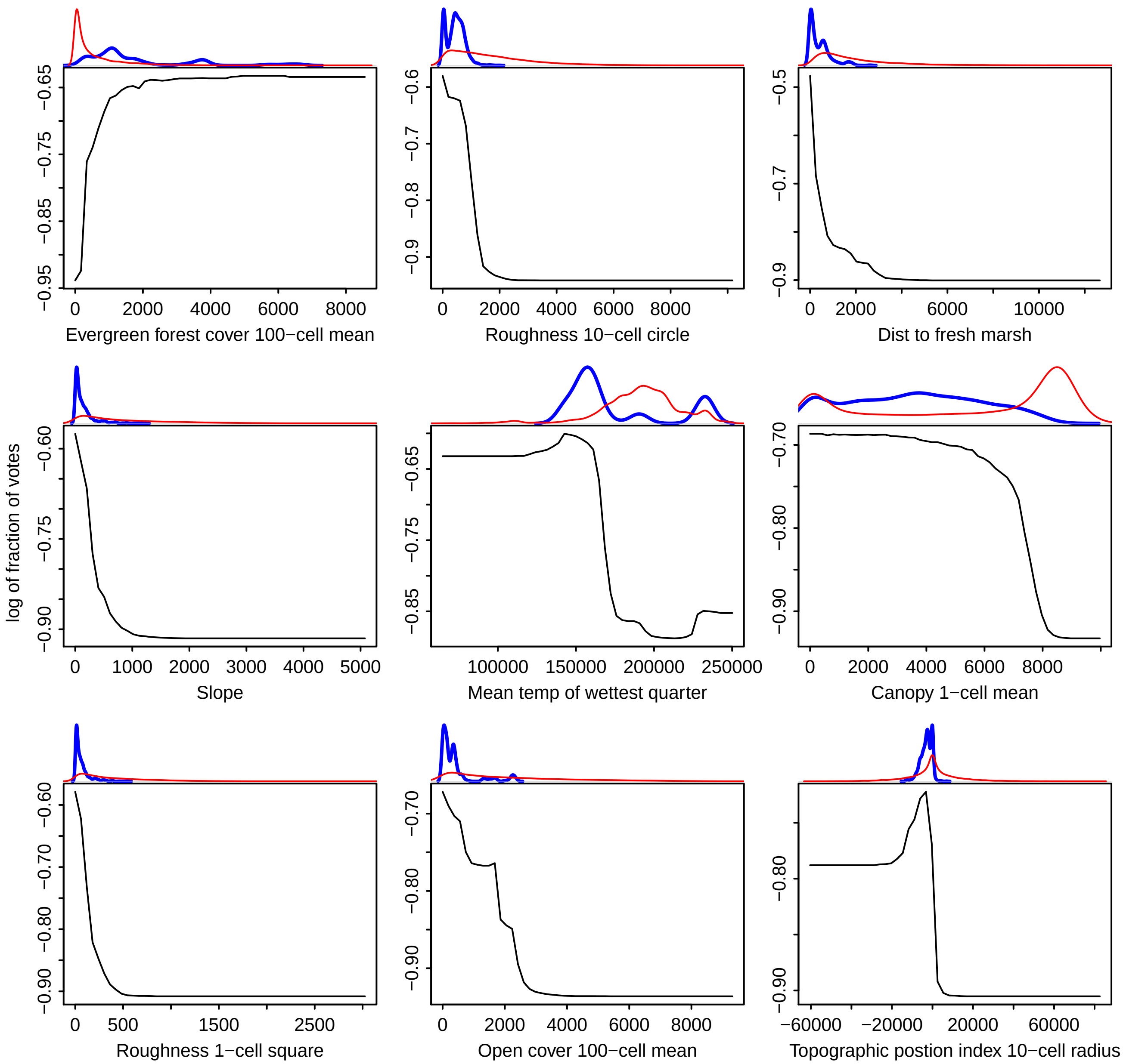


Figure 3. Partial dependence plots for the 9 environmental variables with the most influence on the model. Each plot shows the effect of the variable on the probability of appropriate habitat with the effects of the other variables removed [3]. Peaks in the line indicate where this variable had the strongest 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	0.631	100(27)	100(28)	99.6	The probability at which the absolute value of sensitivity minus specificity is minimized.
F-measure with alpha set to 0.01	0.577	100(27)	100(28)	100	The harmonic average of precision and recall, with strong weighting towards recall (classifying presence points as suitable habitat).
Maximum of sensitivity plus specificity	0.672	100(27)	100(28)	99.5	The probability at which the sum of sensitivity (true positive rate) and specificity (true negative rate) is maximized.
Minimum Training Presence	0.577	100(27)	100(28)	100	The lowest probability value assigned to any of the input presence points. 100% of input presence points are classified as suitable habitat.
Minimum Training Presence by Element Occurrence	0.964	100(27)	100(28)	40.7	The lowest probability value assigned to any of the input presence element occurrences. This calculation first summarizes EOs by their maximum and then finds the minimum of these values.
Minimum Training Presence by Polygon	0.964	100(27)	100(28)	40.7	The lowest probability value assigned to any of the input presence polygons.
Tenth percentile of training presence	0.867	100(27)	100(28)	90	The probability at which 90% of the input presence points are classified as suitable habitat and 10% are classified as unsuitable.

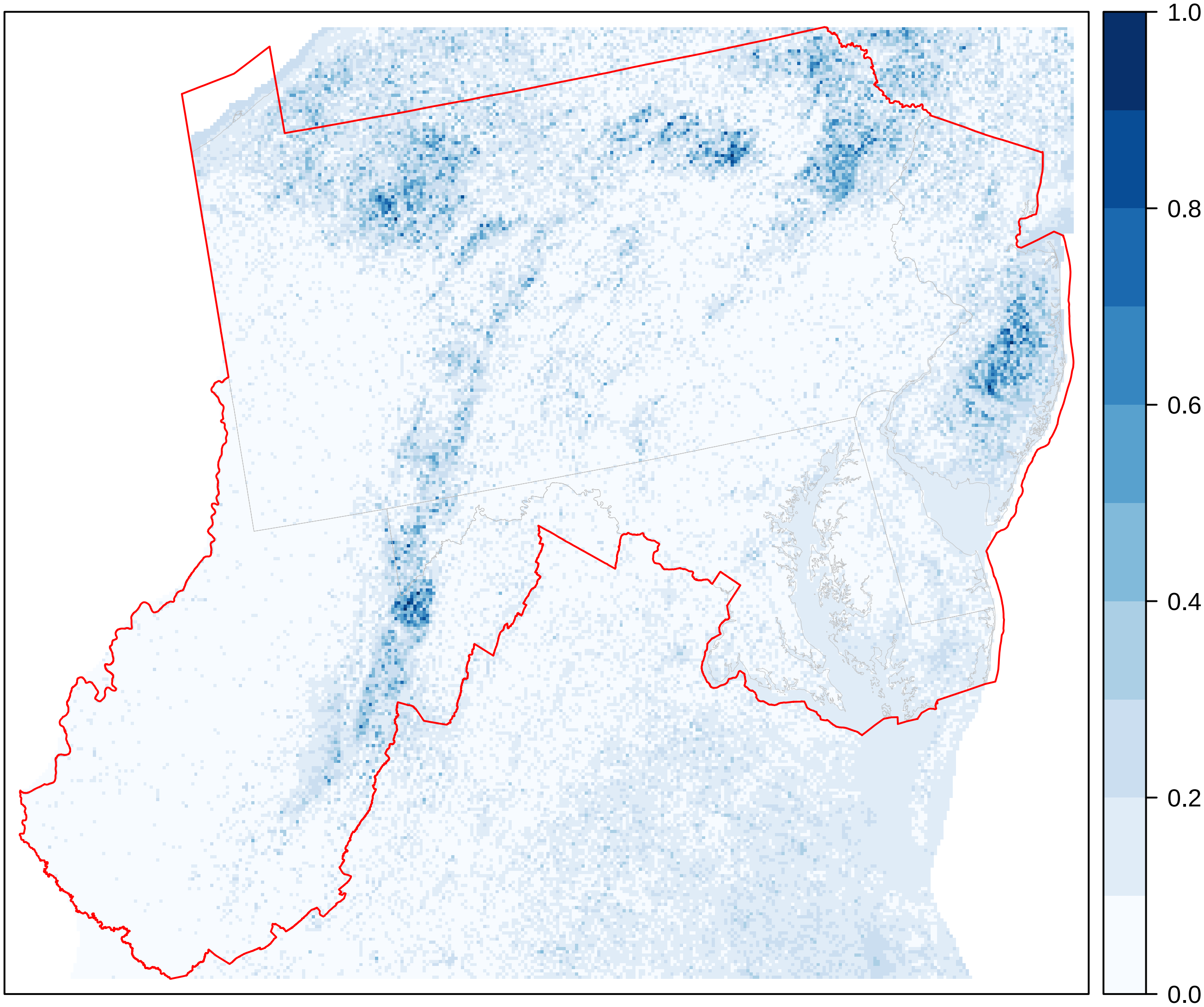


Figure 5. A generalized view of the model predictions throughout the study area. State boundaries are shown in black. The study area is outlined in red.

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

- Maryland Natural Heritage Program, Maryland Department of Natural Resources, Wildlife and Heritage Service
- New Jersey Department of Environmental Protection, Division of Fish and Wildlife, New Jersey Endangered & Nongame Species Program
- Pennsylvania 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.

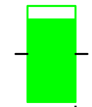
Please cite this document and its associated SDM as:
Pennsylvania Natural Heritage Program. 2018. Species distribution model for Two-spotted Skipper (*Euphyes bimacula*). Created on 30 Jan 2018. Western Pennsylvania Conservancy, Pittsburgh, PA.

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 3.4.3 (2017-11-30).
[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.

Euphyes conspicua

Species Distribution Model (SDM) assessment metrics and metadata
Common name: Black Dash
Date: 09 Dec 2017
Code: euphcons



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 76 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); Groups = element occurrence BG points = background points; PR points = presence points placed throughout all polygons.

Name	Number
polys	113
EOs	76
BG points	11473
PR points	4432

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

Name	Mean	SD	SEM
Overall Accuracy	0.93	0.10	0.01
Specificity	0.93	0.20	0.02
Sensitivity	0.93	0.08	0.01
TSS	0.86	0.21	0.02
Kappa	0.86	0.21	0.02
AUC	0.99	0.03	0.00

Validation runs used 60 environmental variables, the most important of 89 variables (top 75 percent). Each tree was built with 2 variables tried at each split (mtry) and 750 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.

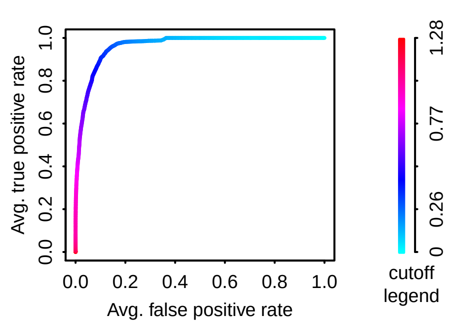


Figure 1. ROC plot for all 76 validation runs, averaged along cutoffs.

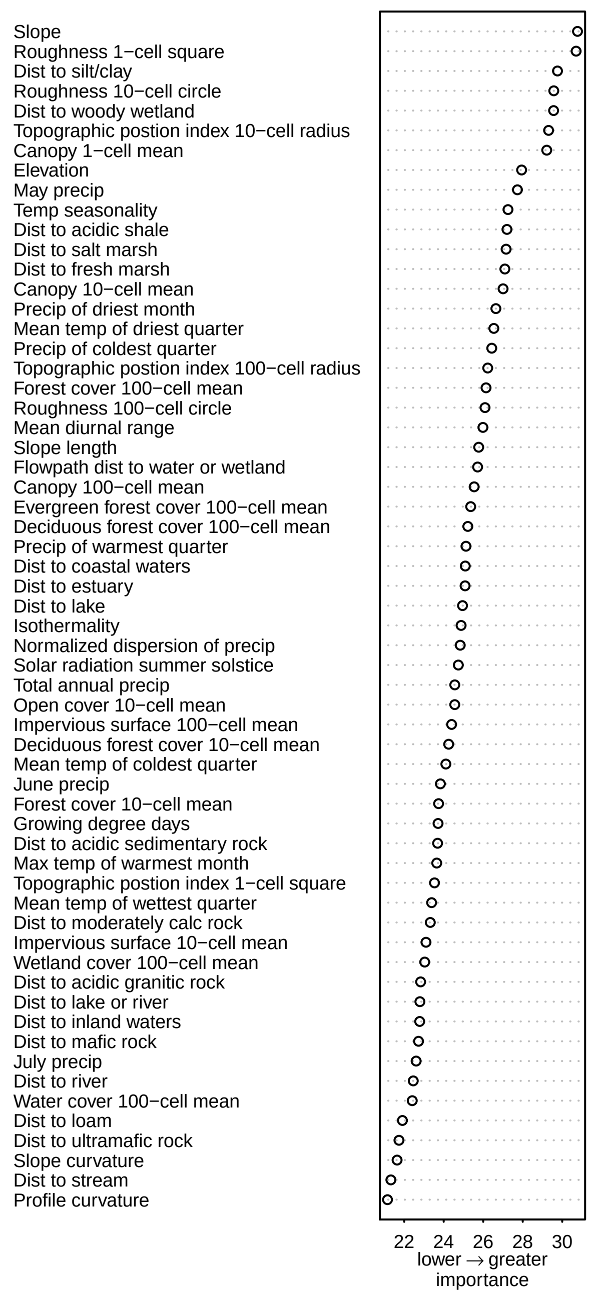


Figure 2. Relative importance of each environmental variable based on the full model using all sites as input. Abbreviations used: calc = calcareous, CP = coastal plain, dist = distance, fresh = freshwater, precip = precipitation, temp = temperature, max = maximum, min = minimum.

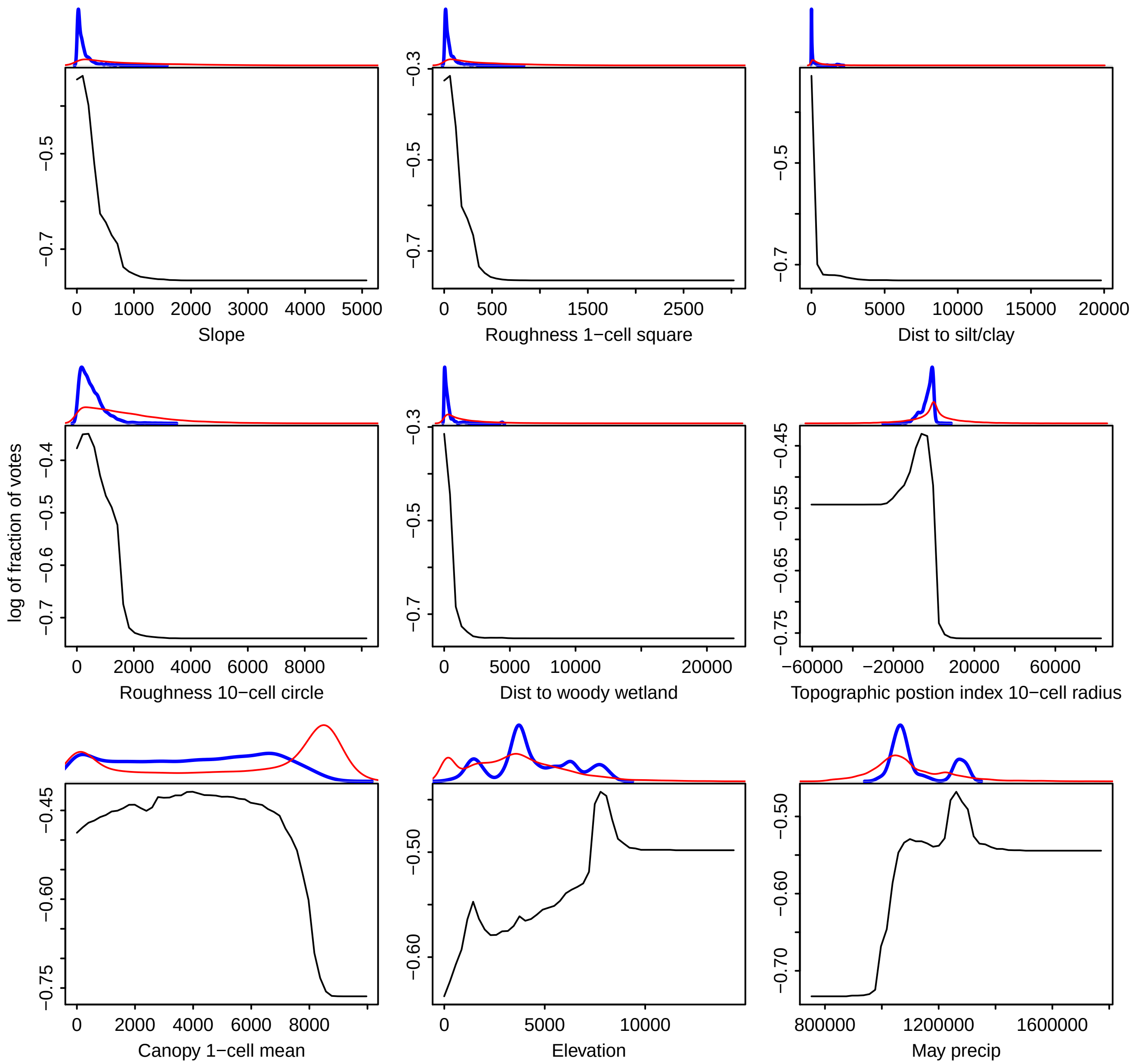


Figure 3. Partial dependence plots for the 9 environmental variables with the most influence on the model. Each plot shows the effect of the variable on the probability of appropriate habitat with the effects of the other variables removed [3]. Peaks in the line indicate where this variable had the strongest 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	0.601	100(76)	100(113)	99	The probability at which the absolute value of sensitivity minus specificity is minimized.
F-measure with alpha set to 0.01	0.419	100(76)	100(113)	100	The harmonic average of precision and recall, with strong weighting towards recall (classifying presence points as suitable habitat).
Maximum of sensitivity plus specificity	0.613	100(76)	100(113)	98.9	The probability at which the sum of sensitivity (true positive rate) and specificity (true negative rate) is maximized.
Minimum Training Presence	0.419	100(76)	100(113)	100	The lowest probability value assigned to any of the input presence points. 100% of input presence points are classified as suitable habitat.
Minimum Training Presence by Element Occurrence	0.867	100(76)	92(104)	78.2	The lowest probability value assigned to any of the input presence element occurrences. This calculation first summarizes EOs by their maximum and then finds the minimum of these values.
Minimum Training Presence by Polygon	0.638	100(76)	100(113)	98.3	The lowest probability value assigned to any of the input presence polygons.
Tenth percentile of training presence	0.794	100(76)	96.5(109)	90	The probability at which 90% of the input presence points are classified as suitable habitat and 10% are classified as unsuitable.

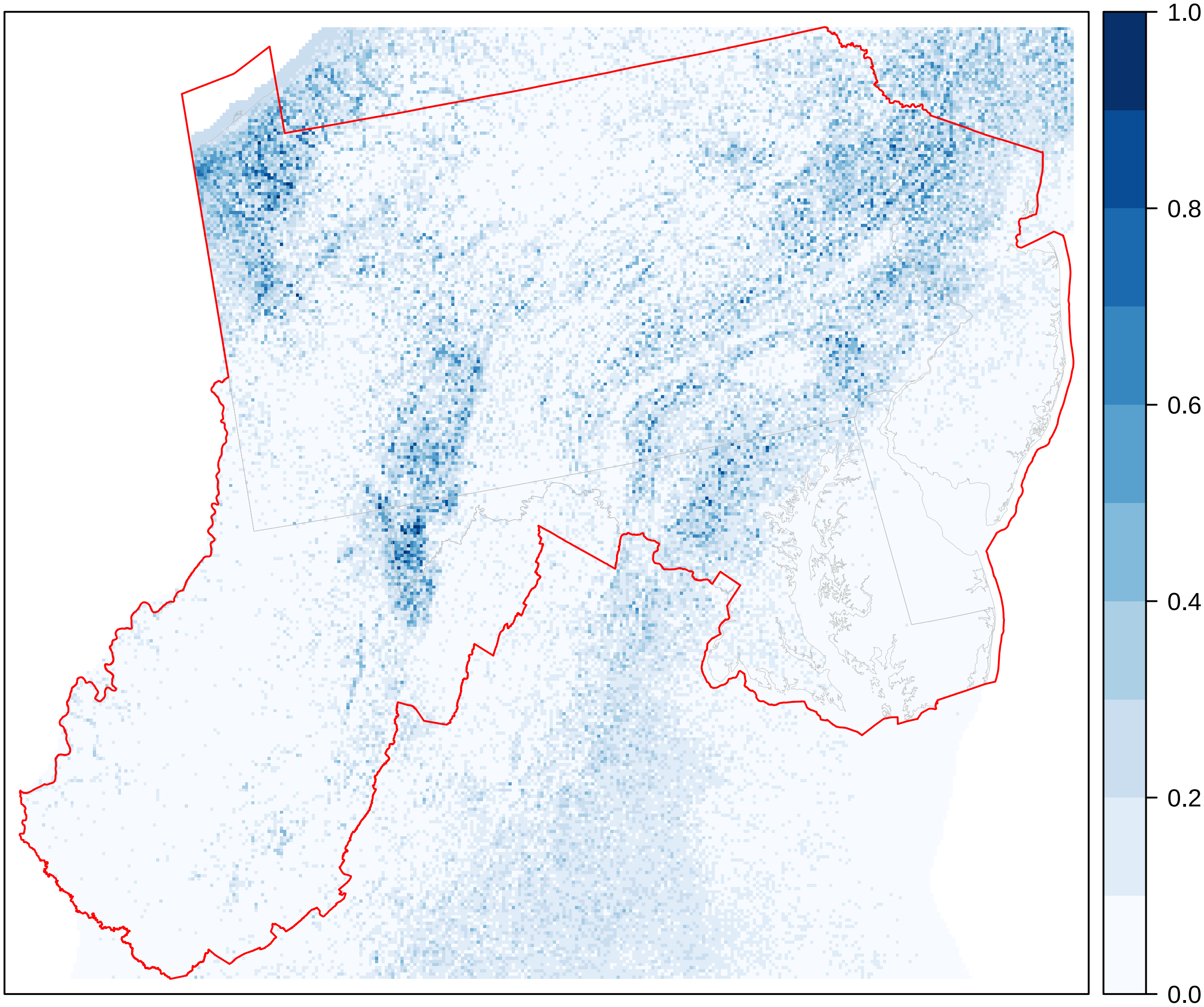


Figure 5. A generalized view of the model predictions throughout the study area. State boundaries are shown in black. The study area is outlined in red.

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

- Maryland Natural Heritage Program, Maryland Department of Natural Resources, Wildlife and Heritage Service
- Pennsylvania 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.

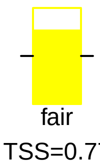
Please cite this document and its associated SDM as:
Pennsylvania Natural Heritage Program. 2017. Species distribution model for Black Dash (*Euphyes conspicua*). Created on 09 Dec 2017. Western Pennsylvania Conservancy, Pittsburgh, PA.

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 3.4.2 (2017-09-28).
[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.

Euphyes dion

Species Distribution Model (SDM) assessment metrics and metadata
Common name: Dion Skipper
Date: 19 Nov 2017
Code: euphdion



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 17 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); Groups = element occurrence BG points = background points; PR points = presence points placed throughout all polygons.

Name	Number
polys	22
EOs	17
BG points	11473
PR points	1781

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

Name	Mean	SD	SEM
Overall Accuracy	0.88	0.18	0.04
Specificity	0.81	0.37	0.09
Sensitivity	0.95	0.06	0.01
TSS	0.77	0.36	0.09
Kappa	0.77	0.36	0.09
AUC	0.95	0.10	0.02

Validation runs used 60 environmental variables, the most important of 88 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.

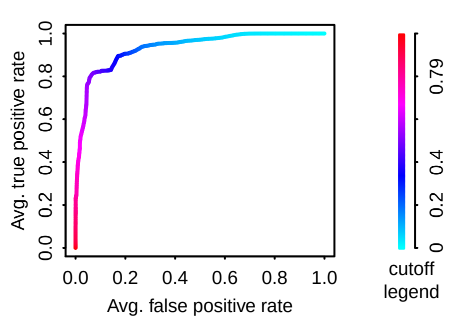


Figure 1. ROC plot for all 17 validation runs, averaged along cutoffs.

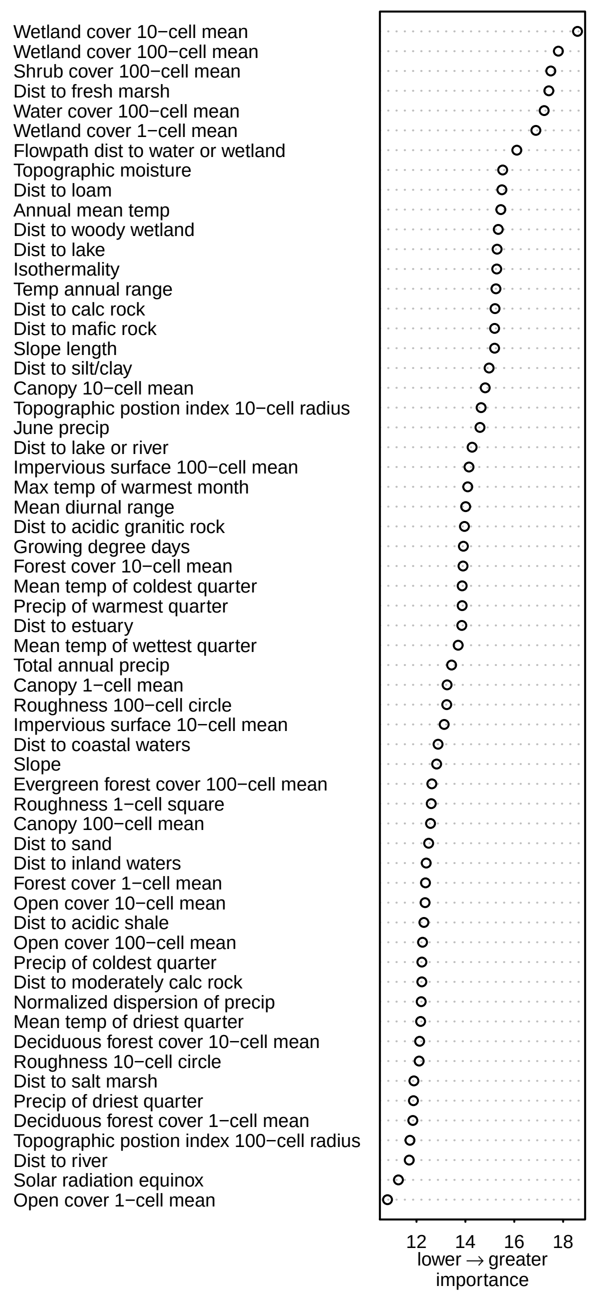


Figure 2. Relative importance of each environmental variable based on the full model using all sites as input. Abbreviations used: calc = calcareous, CP = coastal plain, dist = distance, fresh = freshwater, precip = precipitation, temp = temperature, max = maximum, min = minimum.

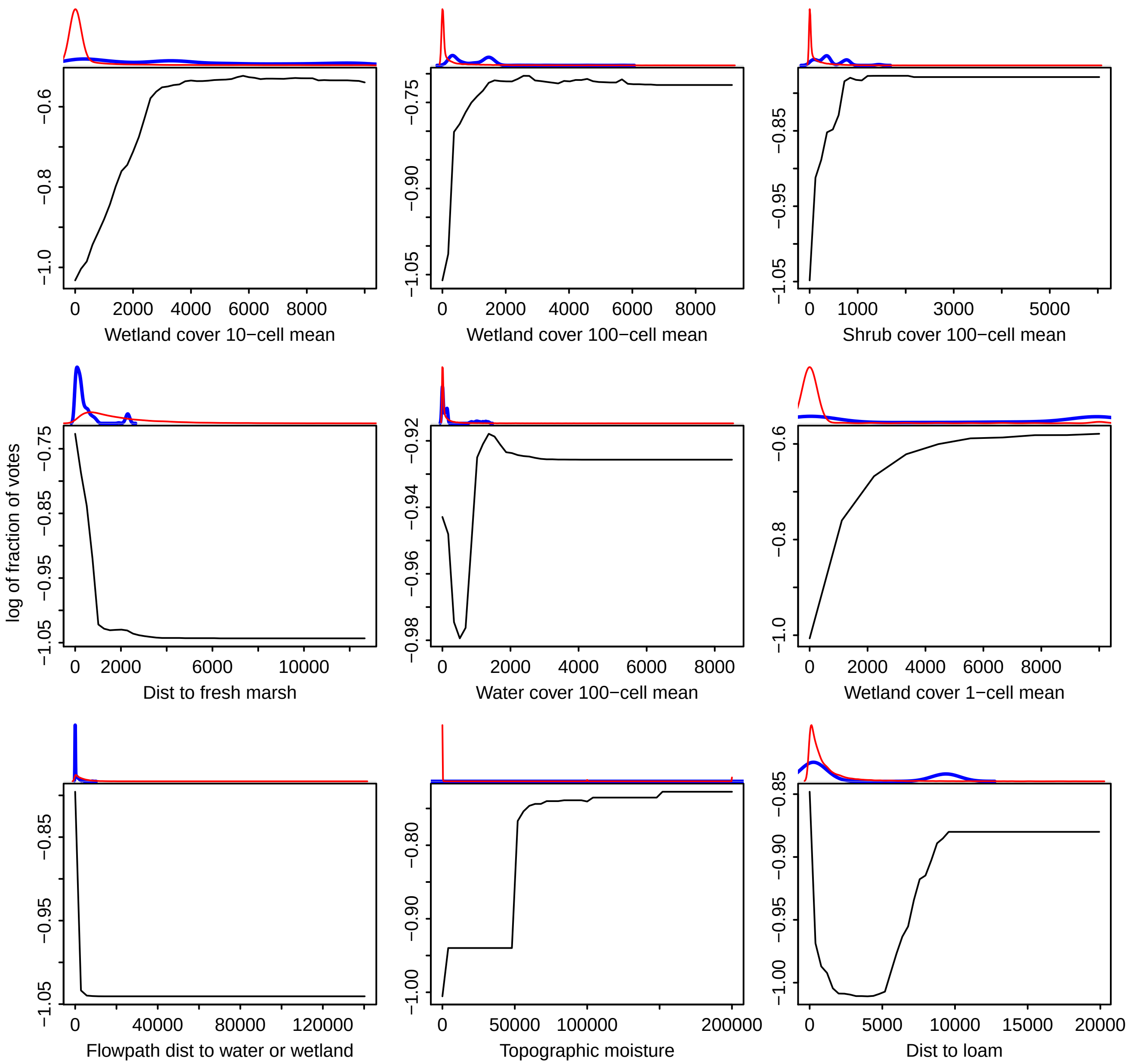


Figure 3. Partial dependence plots for the 9 environmental variables with the most influence on the model. Each plot shows the effect of the variable on the probability of appropriate habitat with the effects of the other variables removed [3]. Peaks in the line indicate where this variable had the strongest 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	0.654	100(17)	100(22)	98.9	The probability at which the absolute value of sensitivity minus specificity is minimized.
F-measure with alpha set to 0.01	0.561	100(17)	100(22)	99.9	The harmonic average of precision and recall, with strong weighting towards recall (classifying presence points as suitable habitat).
Maximum of sensitivity plus specificity	0.625	100(17)	100(22)	99.7	The probability at which the sum of sensitivity (true positive rate) and specificity (true negative rate) is maximized.
Minimum Training Presence	0.407	100(17)	100(22)	100	The lowest probability value assigned to any of the input presence points. 100% of input presence points are classified as suitable habitat.
Minimum Training Presence by Element Occurrence	0.899	100(17)	95.5(21)	44	The lowest probability value assigned to any of the input presence element occurrences. This calculation first summarizes EOs by their maximum and then finds the minimum of these values.
Minimum Training Presence by Polygon	0.894	100(17)	100(22)	45.2	The lowest probability value assigned to any of the input presence polygons.
Tenth percentile of training presence	0.738	100(17)	100(22)	90	The probability at which 90% of the input presence points are classified as suitable habitat and 10% are classified as unsuitable.

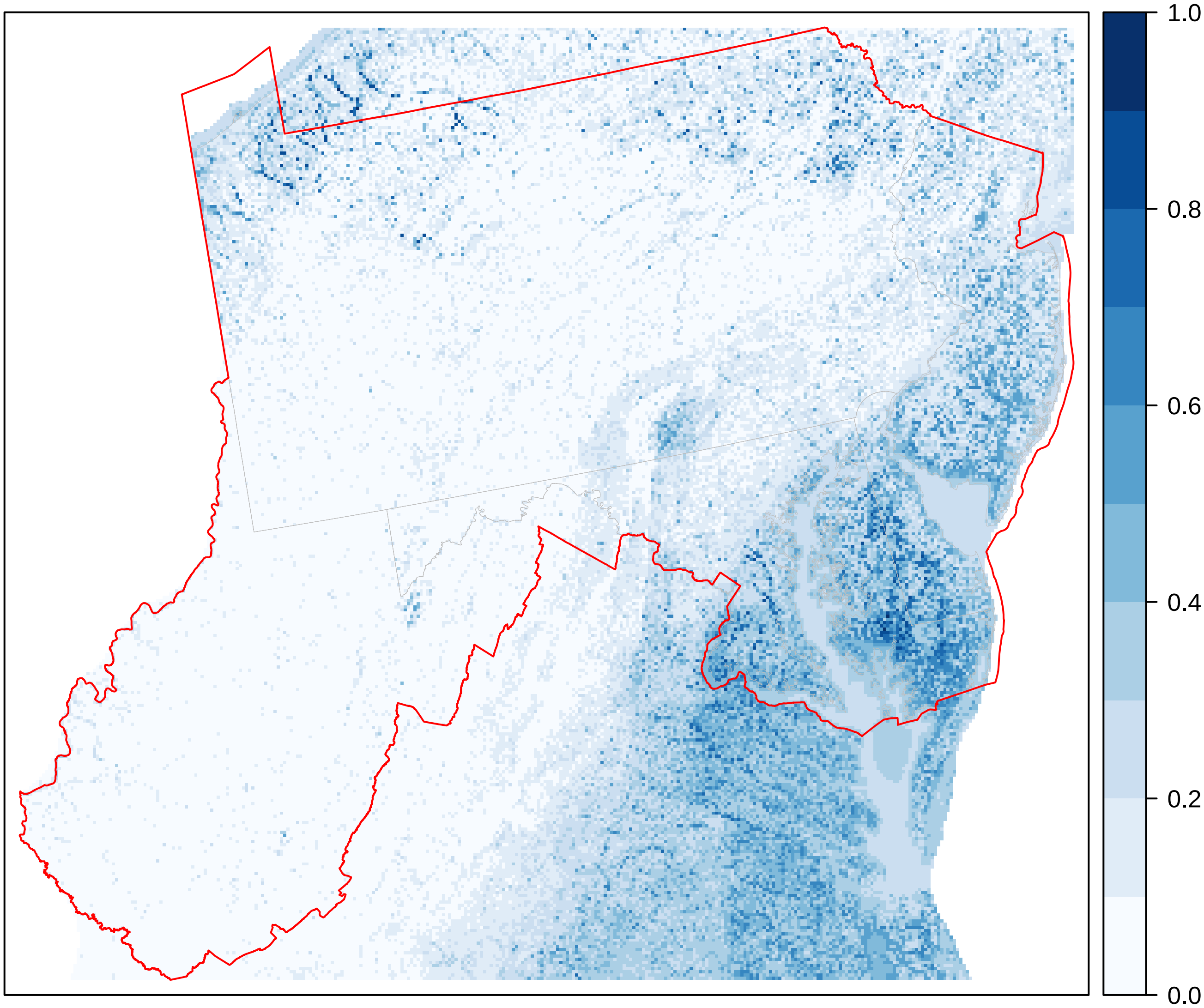


Figure 5. A generalized view of the model predictions throughout the study area. State boundaries are shown in black. The study area is outlined in red.

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

- Maryland Natural Heritage Program, Maryland Department of Natural Resources, Wildlife and Heritage Service
- Pennsylvania 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.

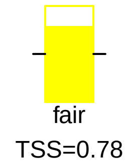
Please cite this document and its associated SDM as:
Pennsylvania Natural Heritage Program. 2017. Species distribution model for Dion Skipper (*Euphyes dion*). Created on 19 Nov 2017. Western Pennsylvania Conservancy, Pittsburgh, PA.

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 3.4.1 (2017-06-30).
[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.

Euphydryas phaeton

Species Distribution Model (SDM) assessment metrics and metadata
Common name: Baltimore Checkerspot
Date: 27 Nov 2017
Code: euphphae



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 134 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); Groups = element occurrence BG points = background points; PR points = presence points placed throughout all polygons.

Name	Number
polys	186
EOs	134
BG points	11473
PR points	8300

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

Name	Mean	SD	SEM
Overall Accuracy	0.89	0.12	0.01
Specificity	0.93	0.21	0.02
Sensitivity	0.85	0.10	0.01
TSS	0.78	0.23	0.02
Kappa	0.78	0.23	0.02
AUC	0.96	0.10	0.01

Validation runs used 60 environmental variables, the most important of 89 variables (top 75 percent). Each tree was built with 1 variables tried at each split (mtry) and 750 trees built. The final model was built using 2000 trees, all presence and background points, with an mtry of 1, and the same number of environmental variables.

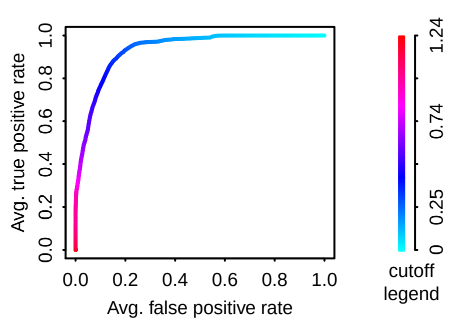


Figure 1. ROC plot for all 134 validation runs, averaged along cutoffs.

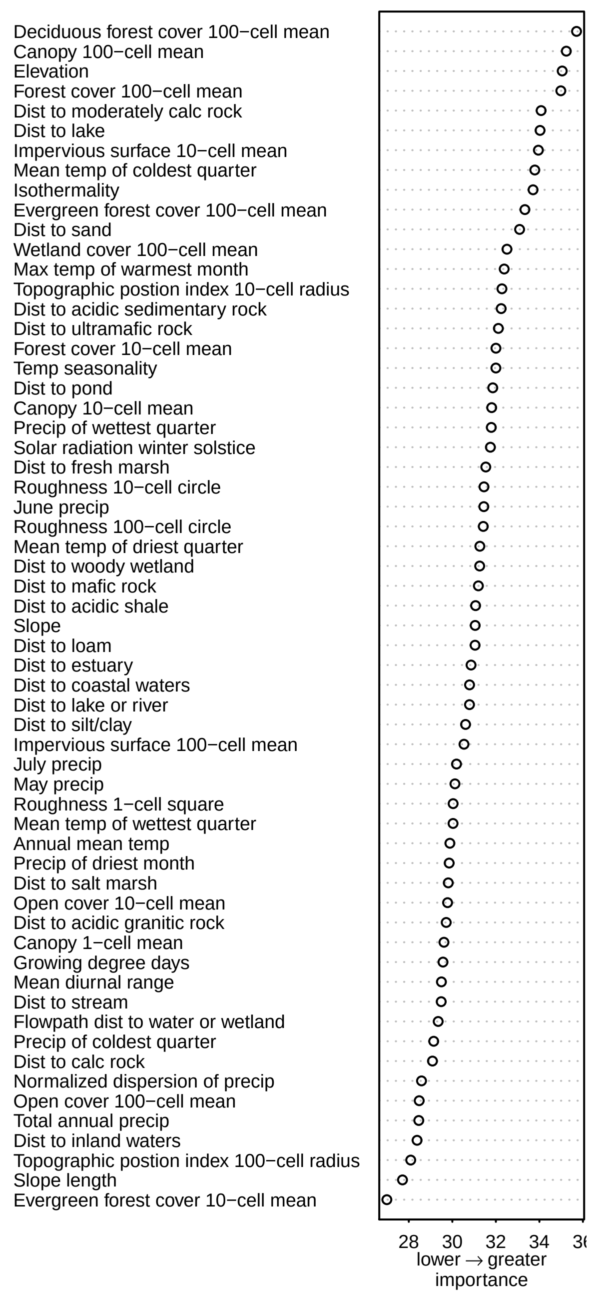


Figure 2. Relative importance of each environmental variable based on the full model using all sites as input. Abbreviations used: calc = calcareous, CP = coastal plain, dist = distance, fresh = freshwater, precip = precipitation, temp = temperature, max = maximum, min = minimum.

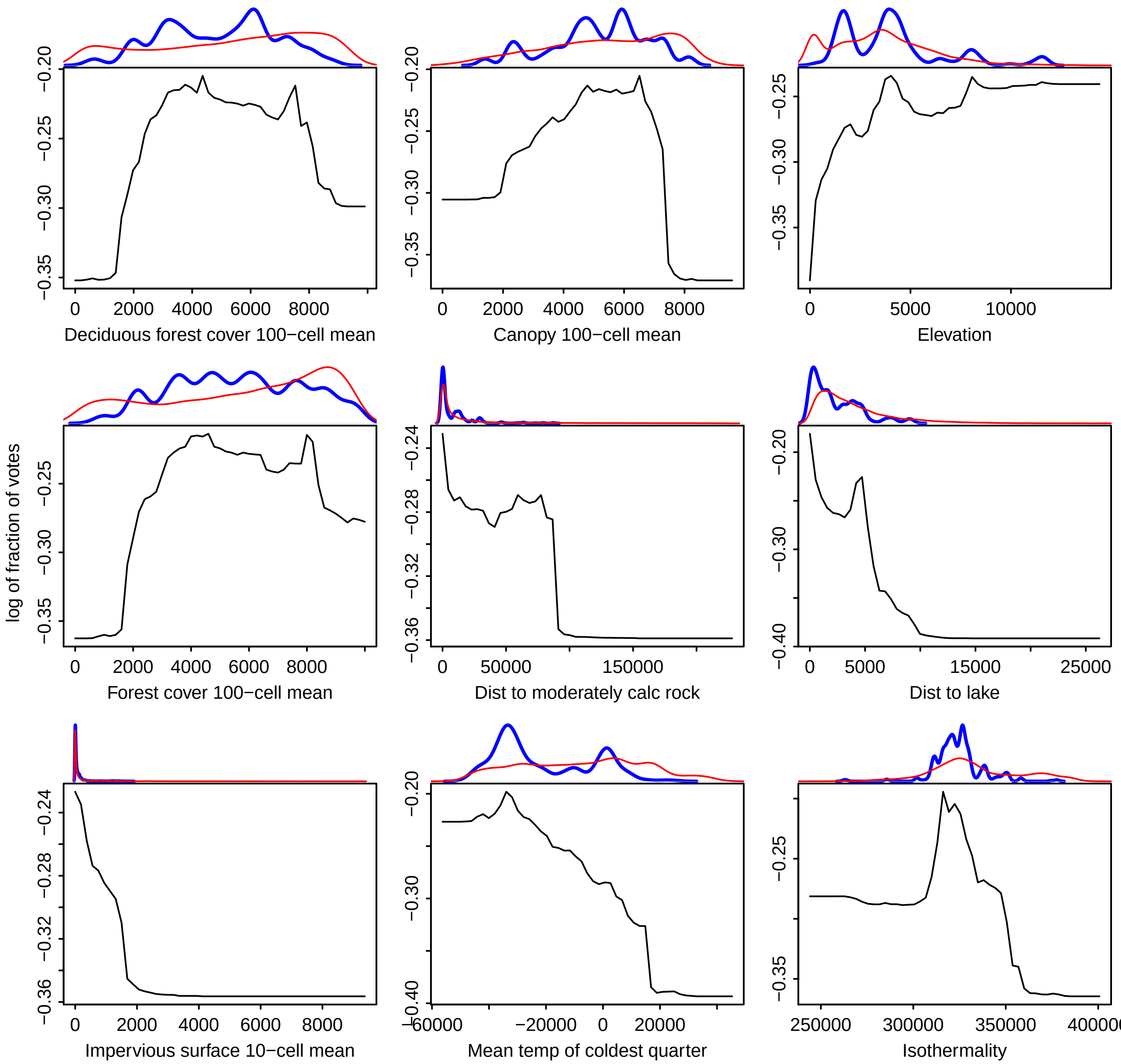


Figure 3. Partial dependence plots for the 9 environmental variables with the most influence on the model. Each plot shows the effect of the variable on the probability of appropriate habitat with the effects of the other variables removed [3]. Peaks in the line indicate where this variable had the strongest 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	0.479	100(134)	100(186)	96.8	The probability at which the absolute value of sensitivity minus specificity is minimized.
F-measure with alpha set to 0.01	0.298	100(134)	100(186)	100	The harmonic average of precision and recall, with strong weighting towards recall (classifying presence points as suitable habitat).
Maximum of sensitivity plus specificity	0.485	100(134)	100(186)	96.7	The probability at which the sum of sensitivity (true positive rate) and specificity (true negative rate) is maximized.
Minimum Training Presence	0.226	100(134)	100(186)	100	The lowest probability value assigned to any of the input presence points. 100% of input presence points are classified as suitable habitat.
Minimum Training Presence by Element Occurrence	0.651	100(134)	98.9(184)	86.7	The lowest probability value assigned to any of the input presence element occurrences. This calculation first summarizes EOs by their maximum and then finds the minimum of these values.
Minimum Training Presence by Polygon	0.622	100(134)	100(186)	89.4	The lowest probability value assigned to any of the input presence polygons.
Tenth percentile of training presence	0.616	100(134)	100(186)	90	The probability at which 90% of the input presence points are classified as suitable habitat and 10% are classified as unsuitable.

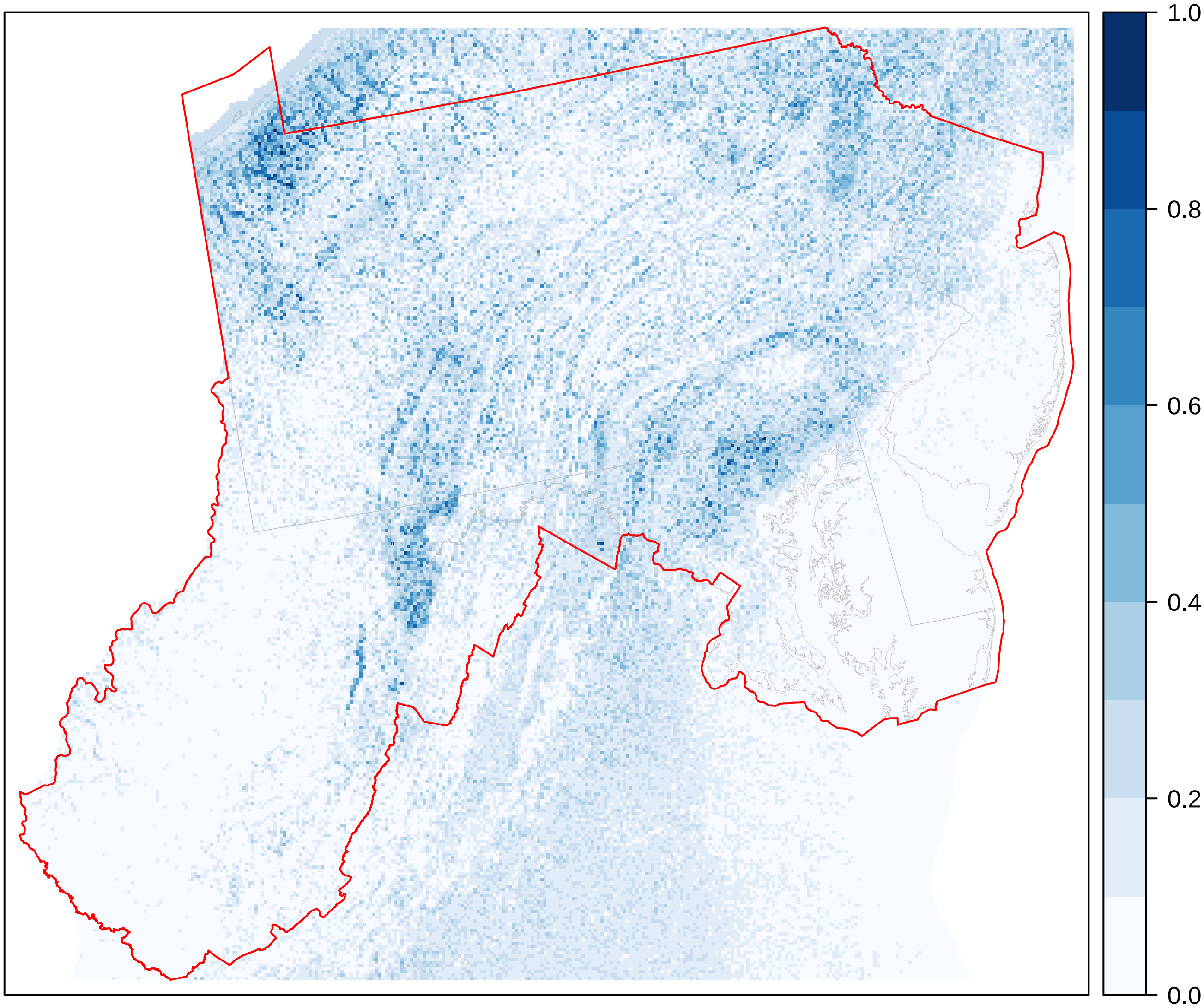


Figure 5. A generalized view of the model predictions throughout the study area. State boundaries are shown in black. The study area is outlined in red.

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

- Maryland Natural Heritage Program, Maryland Department of Natural Resources, Wildlife and Heritage Service
- Pennsylvania Natural Heritage Program
- West 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.

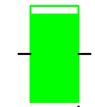
Please cite this document and its associated SDM as:
Pennsylvania Natural Heritage Program. 2017. Species distribution model for Baltimore Checkerspot (*Euphydryas phaeton*). Created on 27 Nov 2017. Western Pennsylvania Conservancy, Pittsburgh, PA.

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 3.4.1 (2017-06-30).
[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.

Lethe eurydice

Species Distribution Model (SDM) assessment metrics and metadata
Common name: Eyed Brown
Date: 01 Feb 2018
Code: letheury



good
TSS=0.91
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 9 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); Groups = element occurrence BG points = background points; PR points = presence points placed throughout all polygons.

Name	Number
polys	12
EOs	9
BG points	11473
PR points	1196

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

Name	Mean	SD	SEM
Overall Accuracy	0.96	0.02	0.01
Specificity	0.99	0.03	0.01
Sensitivity	0.92	0.02	0.01
TSS	0.91	0.04	0.01
Kappa	0.91	0.04	0.01
AUC	0.99	0.02	0.01

Validation runs used 57 environmental variables, the most important of 85 variables (top 75 percent). Each tree was built with 1 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 1, and the same number of environmental variables.

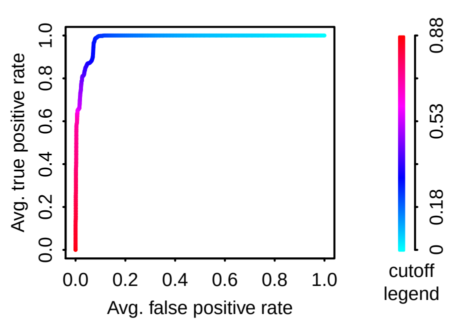


Figure 1. ROC plot for all 9 validation runs, averaged along cutoffs.

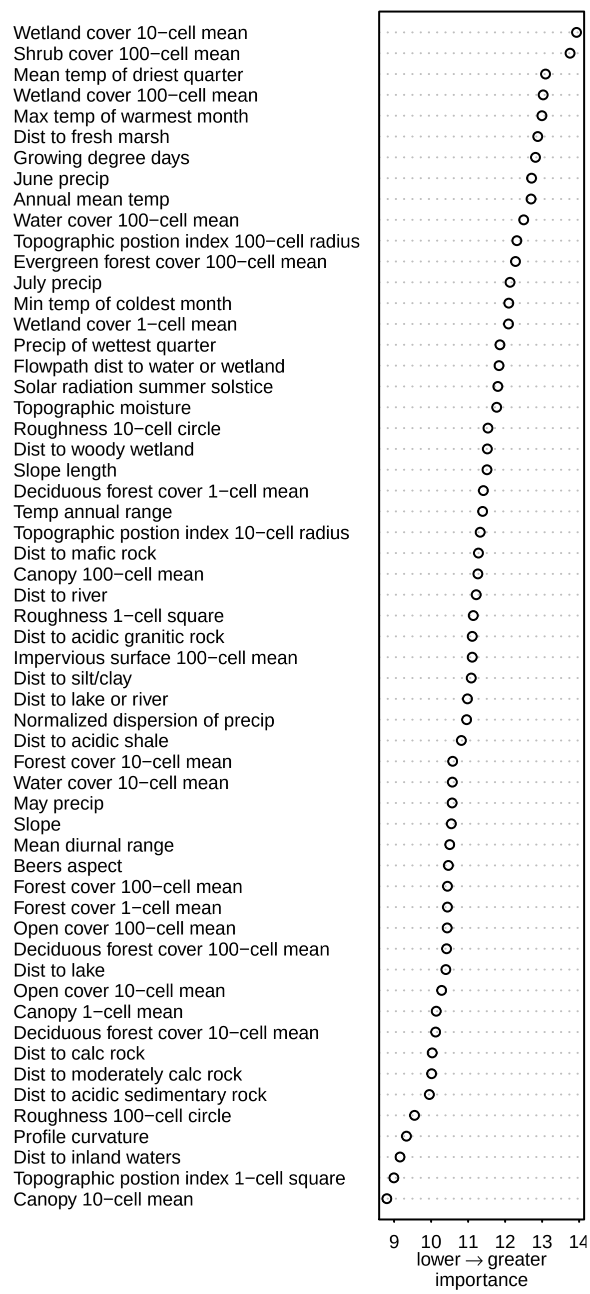


Figure 2. Relative importance of each environmental variable based on the full model using all sites as input. Abbreviations used: calc = calcareous, CP = coastal plain, dist = distance, fresh = freshwater, precip = precipitation, temp = temperature, max = maximum, min = minimum.

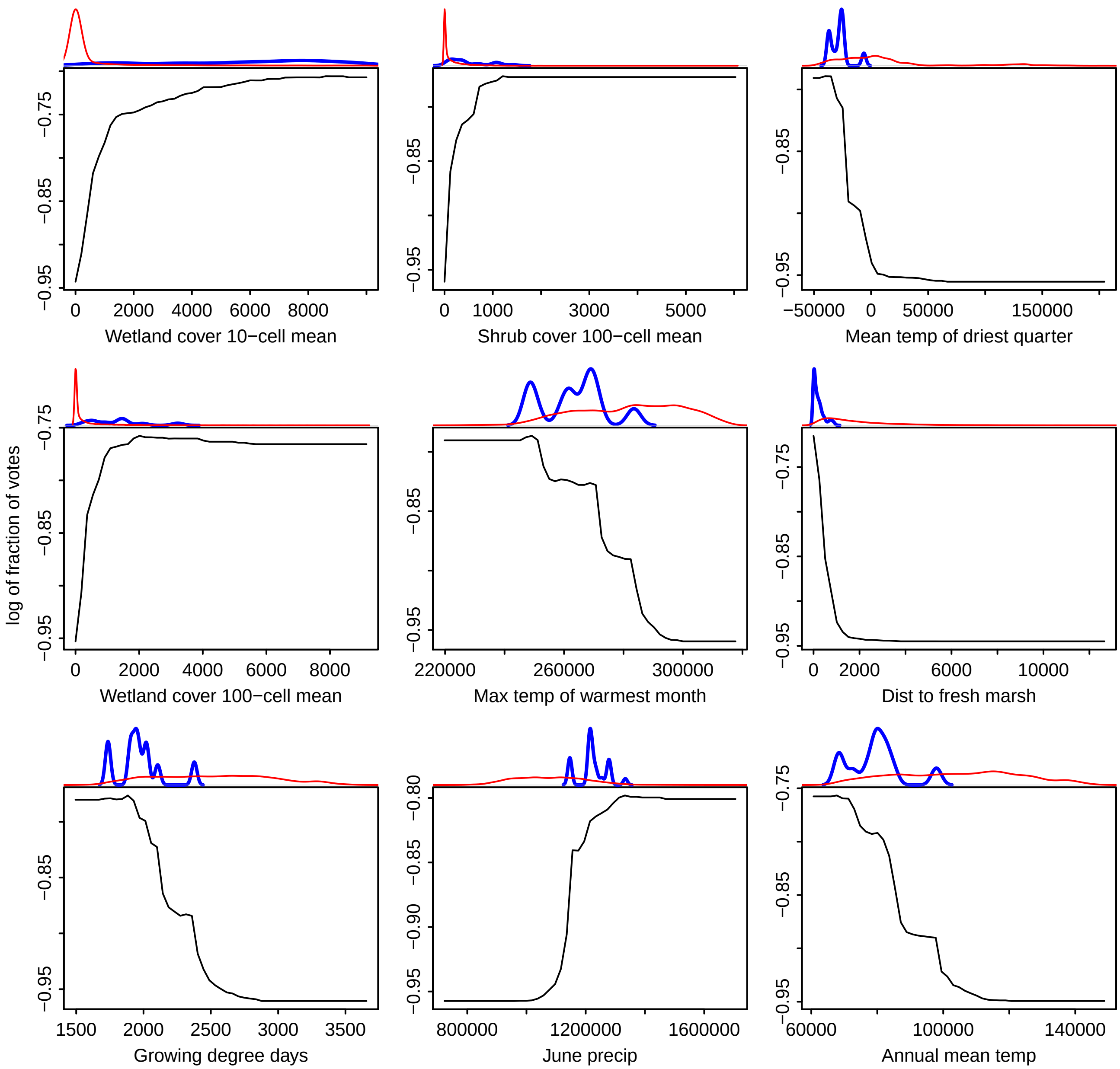


Figure 3. Partial dependence plots for the 9 environmental variables with the most influence on the model. Each plot shows the effect of the variable on the probability of appropriate habitat with the effects of the other variables removed [3]. Peaks in the line indicate where this variable had the strongest 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	0.675	100(9)	100(12)	99.7	The probability at which the absolute value of sensitivity minus specificity is minimized.
F-measure with alpha set to 0.01	0.603	100(9)	100(12)	100	The harmonic average of precision and recall, with strong weighting towards recall (classifying presence points as suitable habitat).
Maximum of sensitivity plus specificity	0.683	100(9)	100(12)	99.7	The probability at which the sum of sensitivity (true positive rate) and specificity (true negative rate) is maximized.
Minimum Training Presence	0.603	100(9)	100(12)	100	The lowest probability value assigned to any of the input presence points. 100% of input presence points are classified as suitable habitat.
Minimum Training Presence by Element Occurrence	0.960	100(9)	100(12)	57.4	The lowest probability value assigned to any of the input presence element occurrences. This calculation first summarizes EOs by their maximum and then finds the minimum of these values.
Minimum Training Presence by Polygon	0.960	100(9)	100(12)	57.4	The lowest probability value assigned to any of the input presence polygons.
Tenth percentile of training presence	0.860	100(9)	100(12)	90	The probability at which 90% of the input presence points are classified as suitable habitat and 10% are classified as unsuitable.

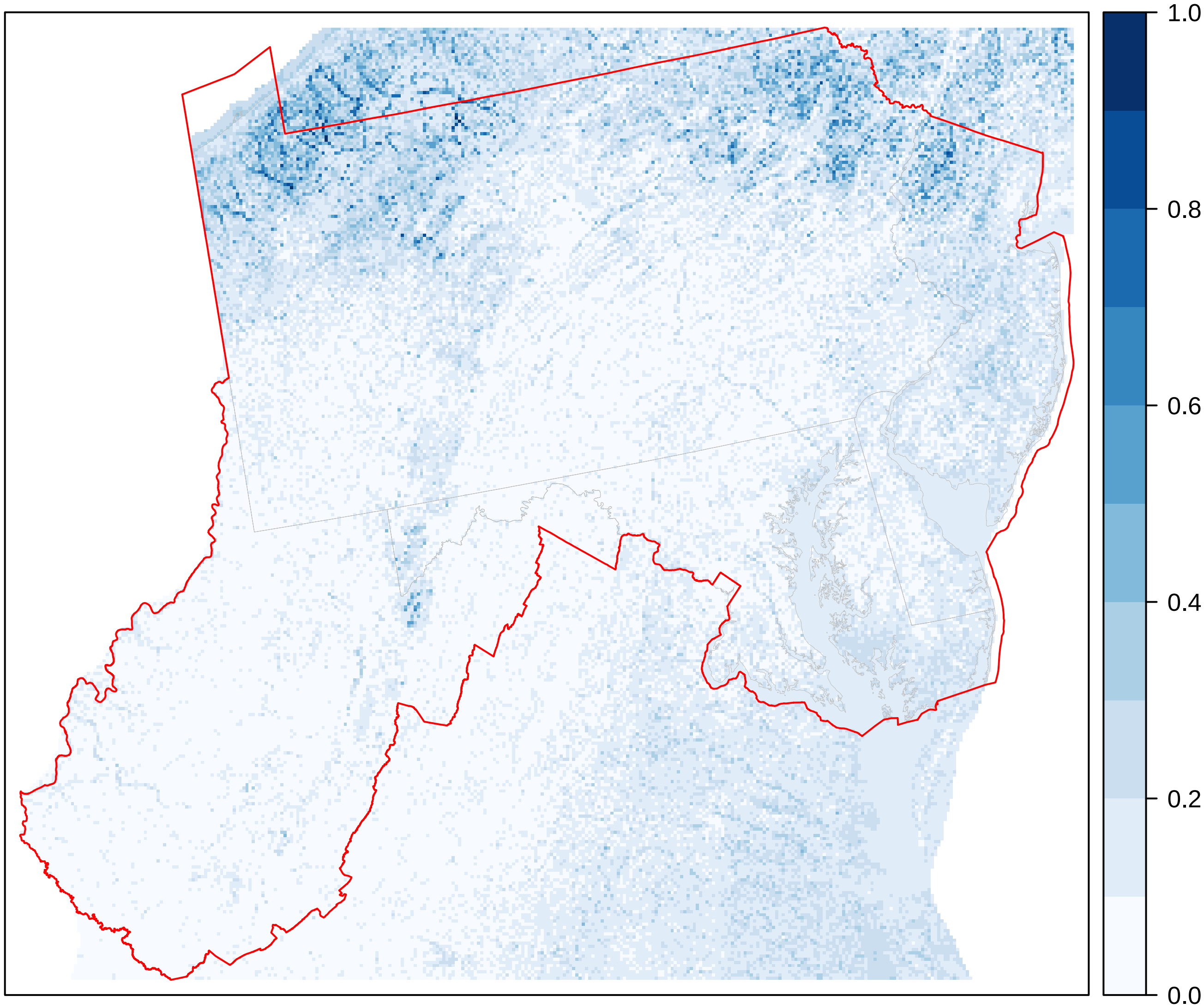


Figure 5. A generalized view of the model predictions throughout the study area. State boundaries are shown in black. The study area is outlined in red.

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

- New Jersey Department of Environmental Protection, Division of Fish and Wildlife, New Jersey Endangered & Nongame Species Program
- Pennsylvania 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.

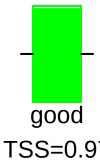
Please cite this document and its associated SDM as:
Pennsylvania Natural Heritage Program. 2018. Species distribution model for Eyed Brown (*Lethe eurydice*). Created on 01 Feb 2018. Western Pennsylvania Conservancy, Pittsburgh, PA.

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 3.4.3 (2017-11-30).
[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.

Lycaena epixanthe

Species Distribution Model (SDM) assessment metrics and metadata
Common name: Bog Copper
Date: 04 Dec 2017
Code: lycaepix



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 51 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); Groups = element occurrence BG points = background points; PR points = presence points placed throughout all polygons.

Name	Number
polys	61
EOs	51
BG points	11473
PR points	4075

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

Name	Mean	SD	SEM
Overall Accuracy	0.98	0.02	0.00
Specificity	0.99	0.03	0.00
Sensitivity	0.98	0.03	0.00
TSS	0.97	0.05	0.01
Kappa	0.97	0.05	0.01
AUC	1.00	0.01	0.00

Validation runs used 56 environmental variables, the most important of 83 variables (top 75 percent). Each tree was built with 1 variables tried at each split (mtry) and 750 trees built. The final model was built using 2000 trees, all presence and background points, with an mtry of 1, and the same number of environmental variables.

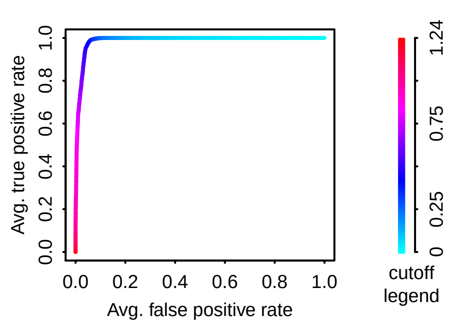


Figure 1. ROC plot for all 51 validation runs, averaged along cutoffs.

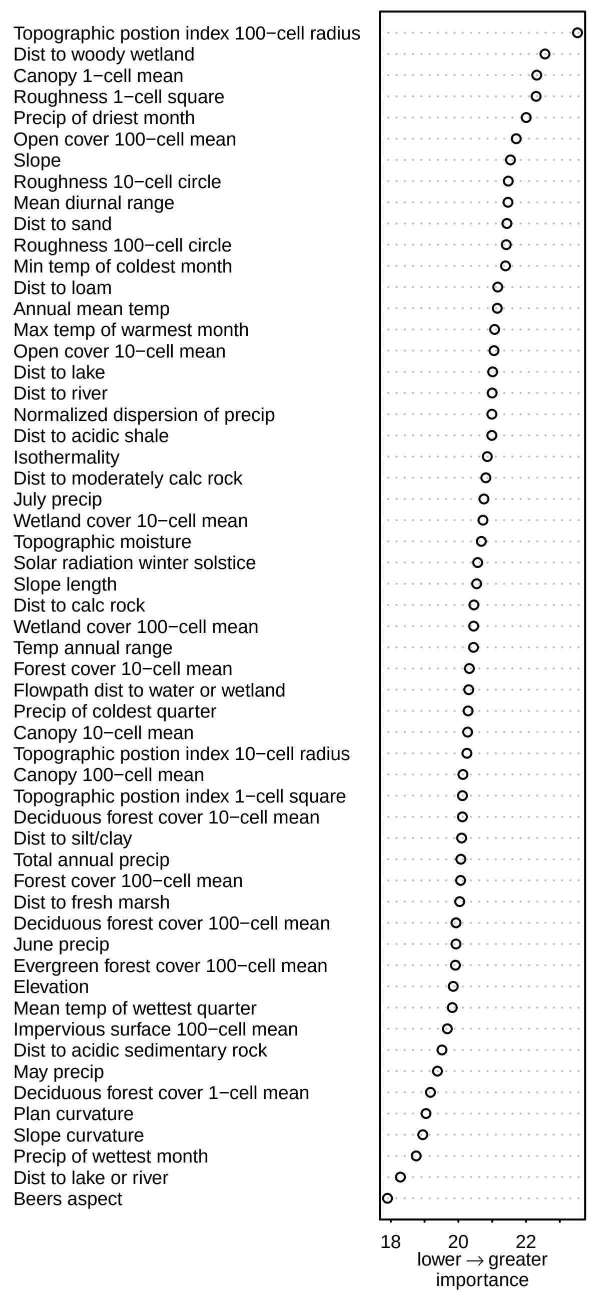


Figure 2. Relative importance of each environmental variable based on the full model using all sites as input. Abbreviations used: calc = calcareous, CP = coastal plain, dist = distance, fresh = freshwater, precip = precipitation, temp = temperature, max = maximum, min = minimum.

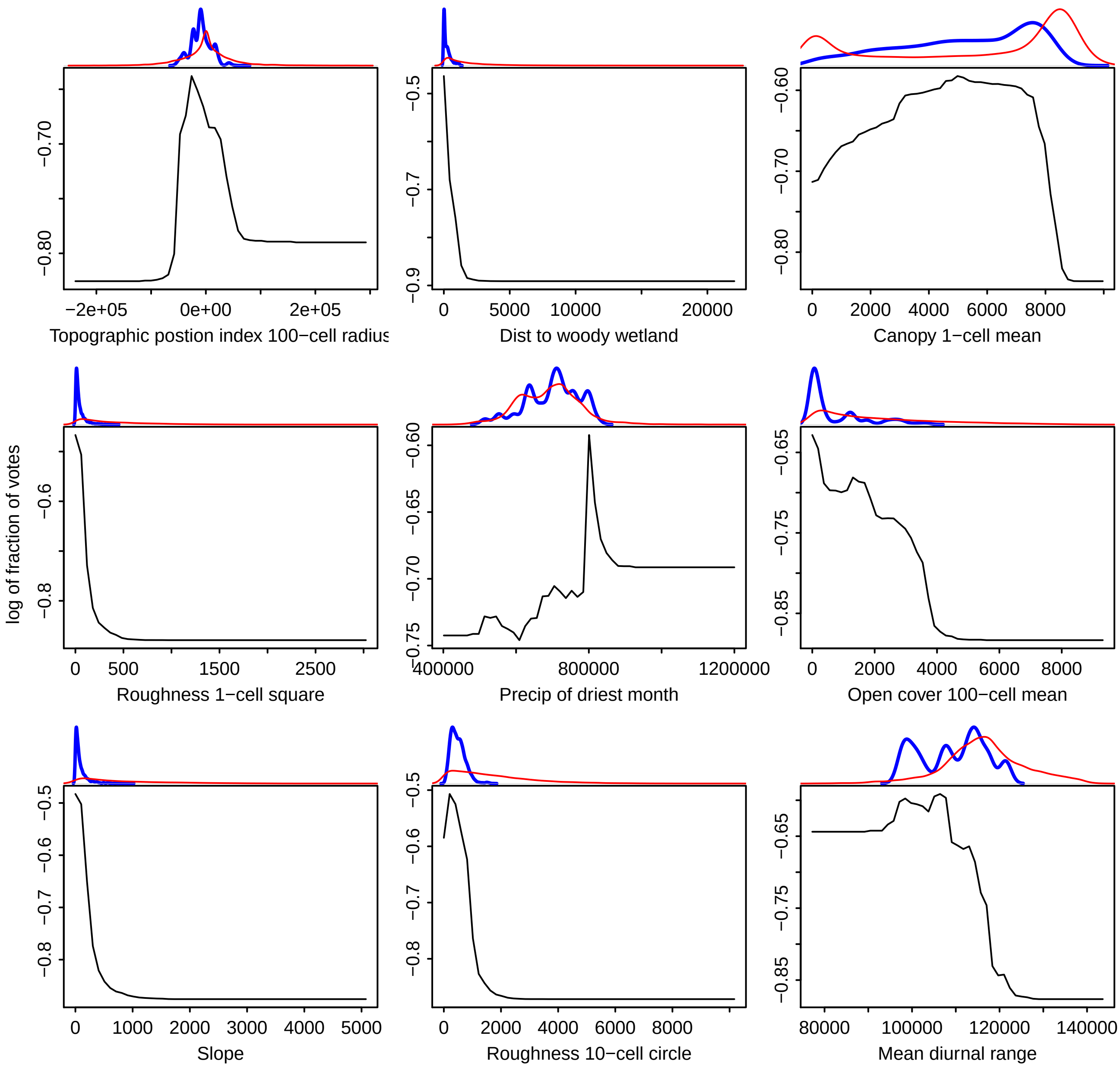


Figure 3. Partial dependence plots for the 9 environmental variables with the most influence on the model. Each plot shows the effect of the variable on the probability of appropriate habitat with the effects of the other variables removed [3]. Peaks in the line indicate where this variable had the strongest 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	0.624	100(51)	100(61)	99.5	The probability at which the absolute value of sensitivity minus specificity is minimized.
F-measure with alpha set to 0.01	0.421	100(51)	100(61)	100	The harmonic average of precision and recall, with strong weighting towards recall (classifying presence points as suitable habitat).
Maximum of sensitivity plus specificity	0.585	100(51)	100(61)	99.7	The probability at which the sum of sensitivity (true positive rate) and specificity (true negative rate) is maximized.
Minimum Training Presence	0.421	100(51)	100(61)	100	The lowest probability value assigned to any of the input presence points. 100% of input presence points are classified as suitable habitat.
Minimum Training Presence by Element Occurrence	0.947	100(51)	93.4(57)	64.1	The lowest probability value assigned to any of the input presence element occurrences. This calculation first summarizes EOs by their maximum and then finds the minimum of these values.
Minimum Training Presence by Polygon	0.759	100(51)	100(61)	97.2	The lowest probability value assigned to any of the input presence polygons.
Tenth percentile of training presence	0.860	100(51)	98.4(60)	90	The probability at which 90% of the input presence points are classified as suitable habitat and 10% are classified as unsuitable.

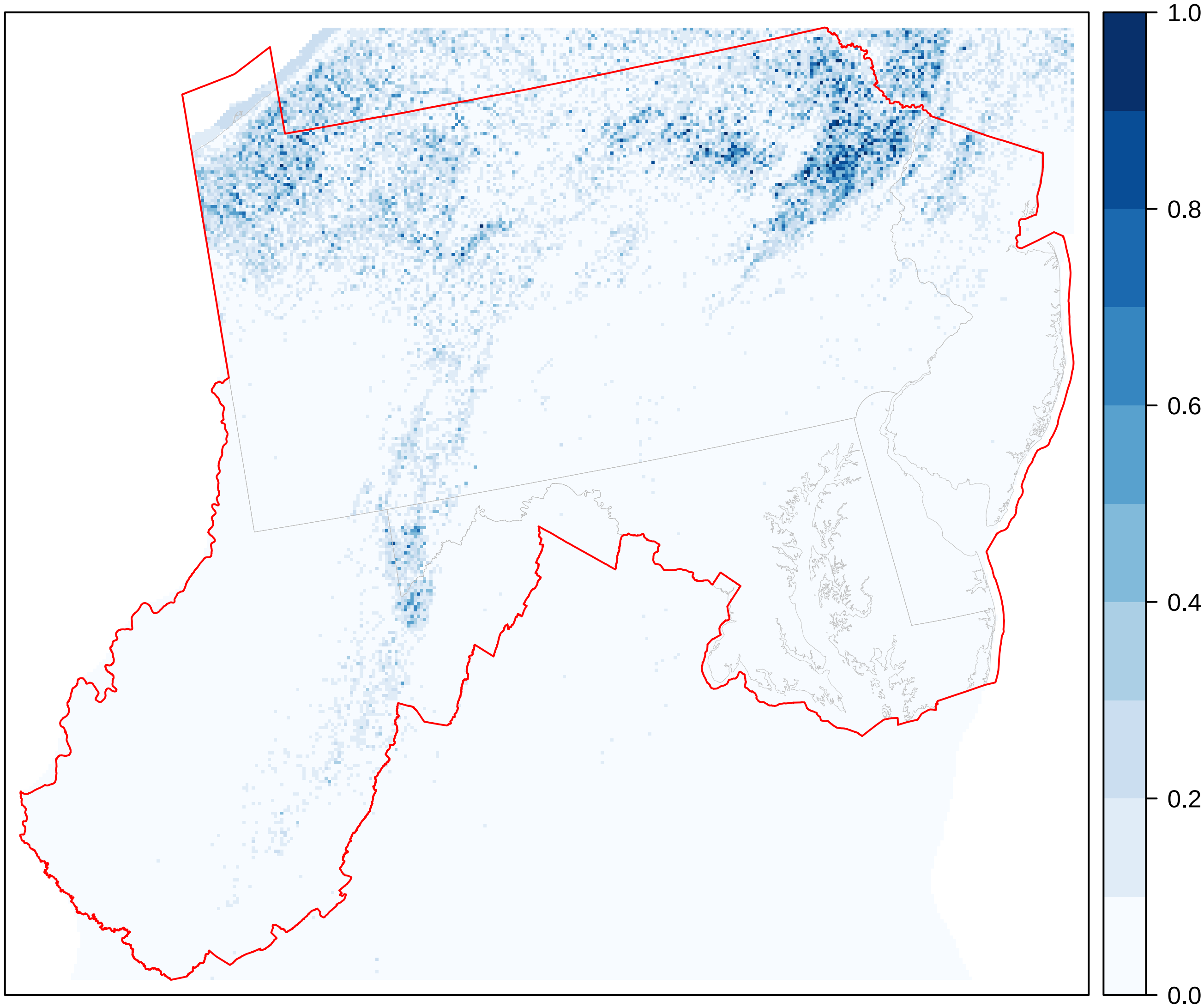


Figure 5. A generalized view of the model predictions throughout the study area. State boundaries are shown in black. The study area is outlined in red.

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

- Maryland Natural Heritage Program, Maryland Department of Natural Resources, Wildlife and Heritage Service
- Pennsylvania Natural Heritage Program
- West 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.

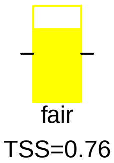
Please cite this document and its associated SDM as:
Pennsylvania Natural Heritage Program. 2017. Species distribution model for Bog Copper (*Lycaena epixanthe*). Created on 04 Dec 2017. Western Pennsylvania Conservancy, Pittsburgh, PA.

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 3.4.2 (2017-09-28).
[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.

Lycaena hyllus

Species Distribution Model (SDM) assessment metrics and metadata
Common name: Bronze Copper
Date: 01 Feb 2018
Code: lycahyll



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 68 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); Groups = element occurrence BG points = background points; PR points = presence points placed throughout all polygons.

Name	Number
polys	92
EOs	68
BG points	11473
PR points	7904

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

Name	Mean	SD	SEM
Overall Accuracy	0.88	0.14	0.02
Specificity	0.89	0.27	0.03
Sensitivity	0.87	0.09	0.01
TSS	0.76	0.28	0.03
Kappa	0.76	0.28	0.03
AUC	0.94	0.15	0.02

Validation runs used 61 environmental variables, the most important of 90 variables (top 75 percent). Each tree was built with 2 variables tried at each split (mtry) and 750 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.

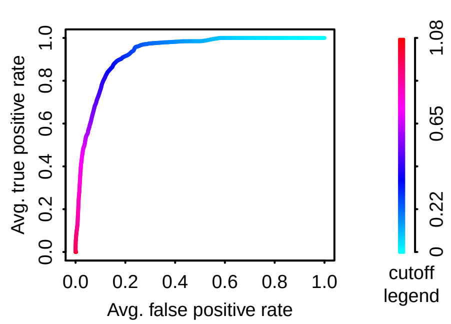


Figure 1. ROC plot for all 68 validation runs, averaged along cutoffs.

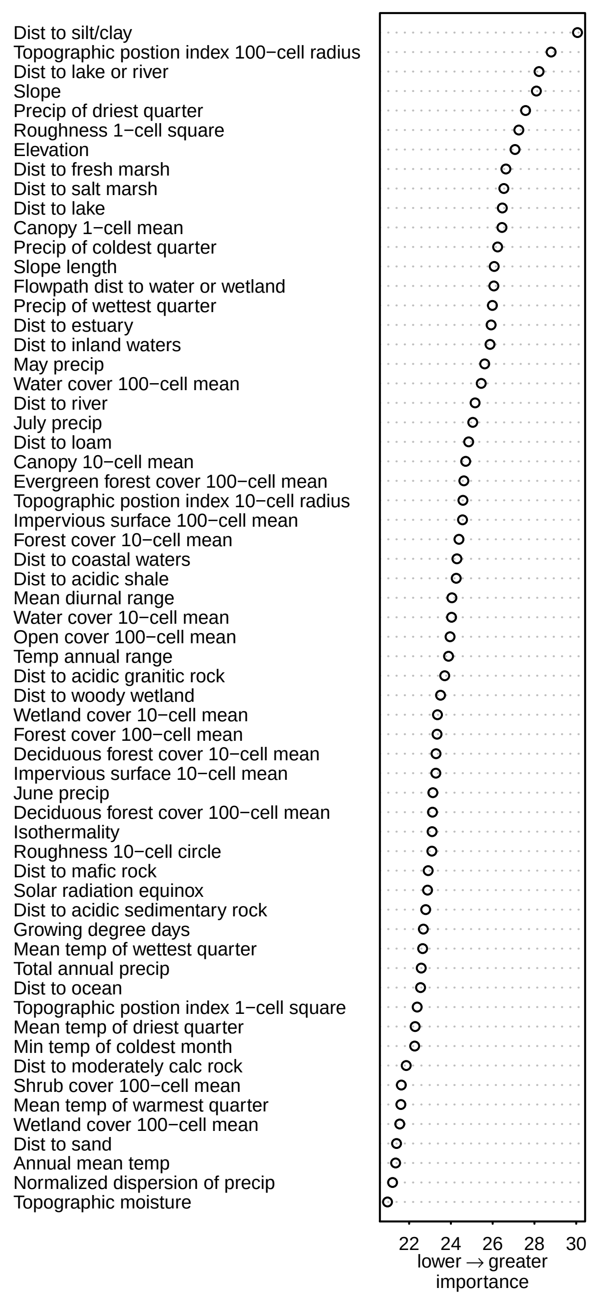


Figure 2. Relative importance of each environmental variable based on the full model using all sites as input. Abbreviations used: calc = calcareous, CP = coastal plain, dist = distance, fresh = freshwater, precip = precipitation, temp = temperature, max = maximum, min = minimum.

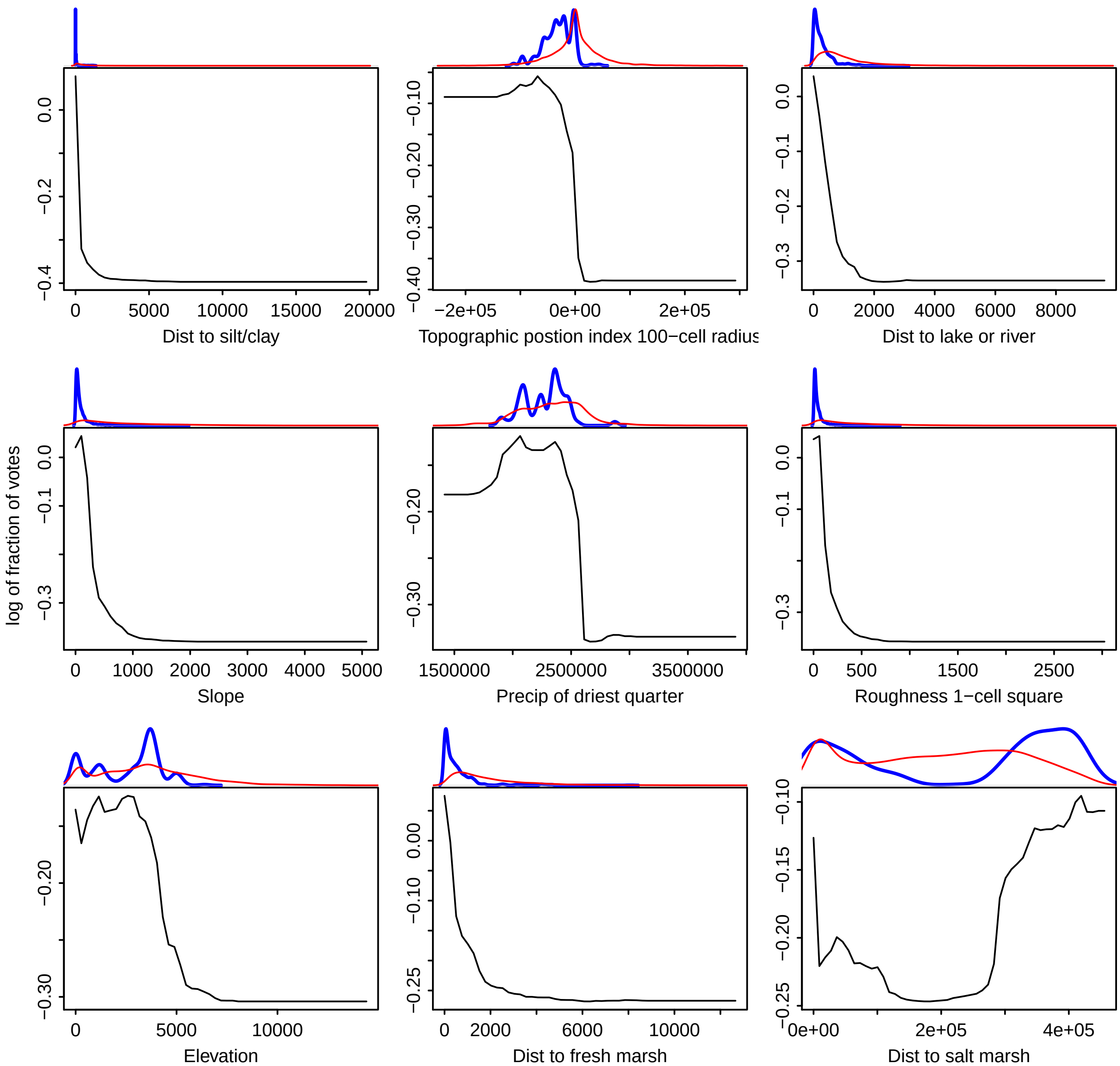


Figure 3. Partial dependence plots for the 9 environmental variables with the most influence on the model. Each plot shows the effect of the variable on the probability of appropriate habitat with the effects of the other variables removed [3]. Peaks in the line indicate where this variable had the strongest 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	0.669	100(68)	100(92)	98.6	The probability at which the absolute value of sensitivity minus specificity is minimized.
F-measure with alpha set to 0.01	0.367	100(68)	100(92)	99.9	The harmonic average of precision and recall, with strong weighting towards recall (classifying presence points as suitable habitat).
Maximum of sensitivity plus specificity	0.674	100(68)	100(92)	98.6	The probability at which the sum of sensitivity (true positive rate) and specificity (true negative rate) is maximized.
Minimum Training Presence	0.291	100(68)	100(92)	100	The lowest probability value assigned to any of the input presence points. 100% of input presence points are classified as suitable habitat.
Minimum Training Presence by Element Occurrence	0.880	100(68)	91.3(84)	77.3	The lowest probability value assigned to any of the input presence element occurrences. This calculation first summarizes EOs by their maximum and then finds the minimum of these values.
Minimum Training Presence by Polygon	0.684	100(68)	100(92)	98.4	The lowest probability value assigned to any of the input presence polygons.
Tenth percentile of training presence	0.818	100(68)	97.8(90)	90	The probability at which 90% of the input presence points are classified as suitable habitat and 10% are classified as unsuitable.

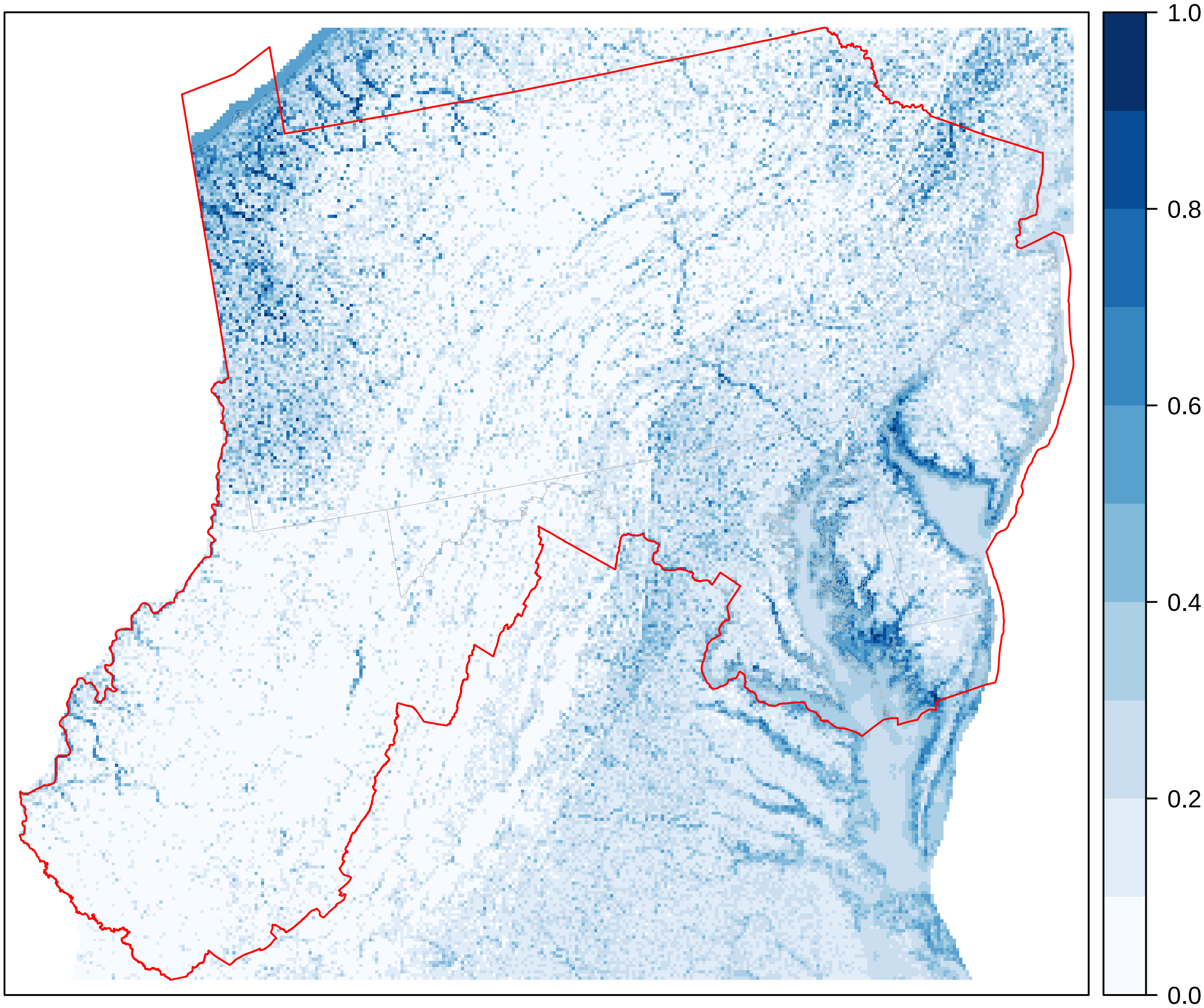


Figure 5. A generalized view of the model predictions throughout the study area. State boundaries are shown in black. The study area is outlined in red.

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

- Maryland Natural Heritage Program, Maryland Department of Natural Resources, Wildlife and Heritage Service
- New Jersey Department of Environmental Protection, Division of Fish and Wildlife, New Jersey Endangered & Nongame Species Program
- Pennsylvania 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.

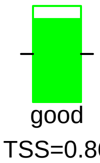
Please cite this document and its associated SDM as:
Pennsylvania Natural Heritage Program. 2018. Species distribution model for Bronze Copper (*Lycaena hyllus*). Created on 01 Feb 2018. Western Pennsylvania Conservancy, Pittsburgh, PA.

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 3.4.3 (2017-11-30).
[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.

Poanes massasoit

Species Distribution Model (SDM) assessment metrics and metadata
Common name: Mulberry Wing
Date: 02 Dec 2017
Code: poanmass



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 28 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); Groups = element occurrence BG points = background points; PR points = presence points placed throughout all polygons.

Name	Number
polys	39
EOs	28
BG points	11473
PR points	906

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

Name	Mean	SD	SEM
Overall Accuracy	0.93	0.11	0.02
Specificity	0.95	0.19	0.04
Sensitivity	0.91	0.08	0.02
TSS	0.86	0.21	0.04
Kappa	0.86	0.21	0.04
AUC	0.99	0.04	0.01

Validation runs used 60 environmental variables, the most important of 89 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.

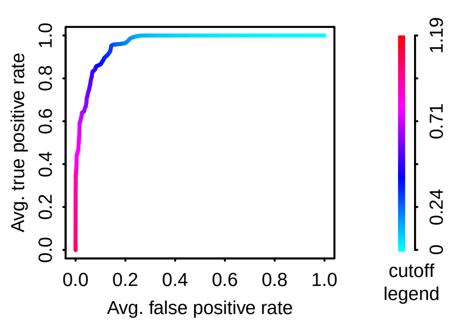


Figure 1. ROC plot for all 28 validation runs, averaged along cutoffs.

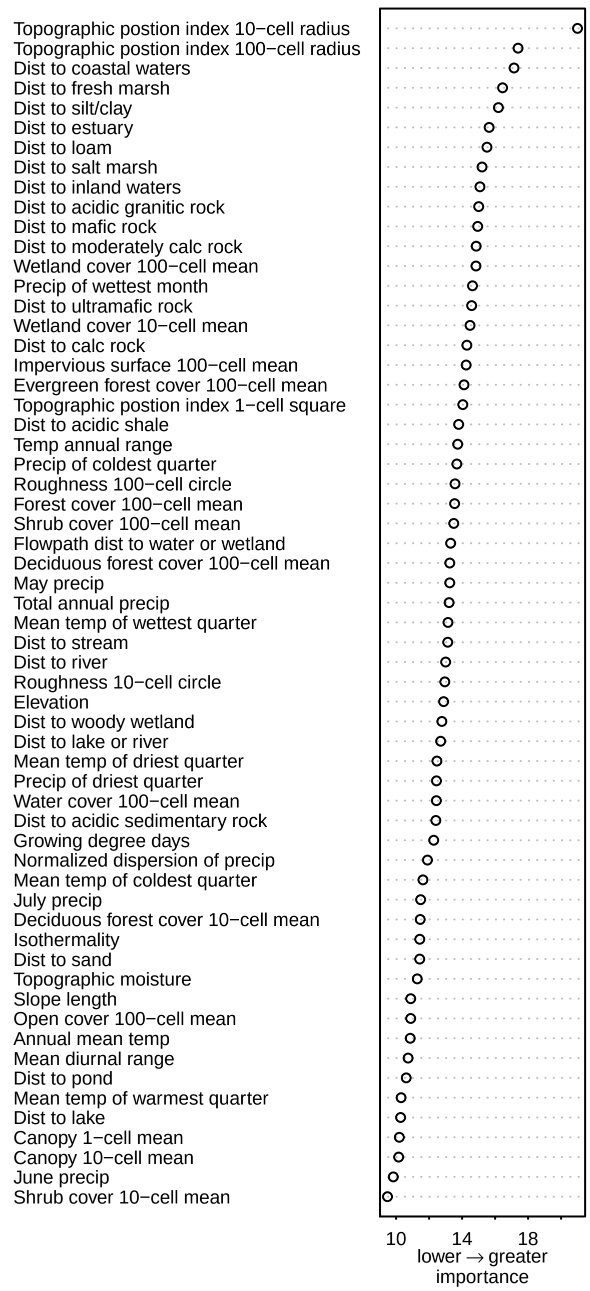


Figure 2. Relative importance of each environmental variable based on the full model using all sites as input. Abbreviations used: calc = calcareous, CP = coastal plain, dist = distance, fresh = freshwater, precip = precipitation, temp = temperature, max = maximum, min = minimum.

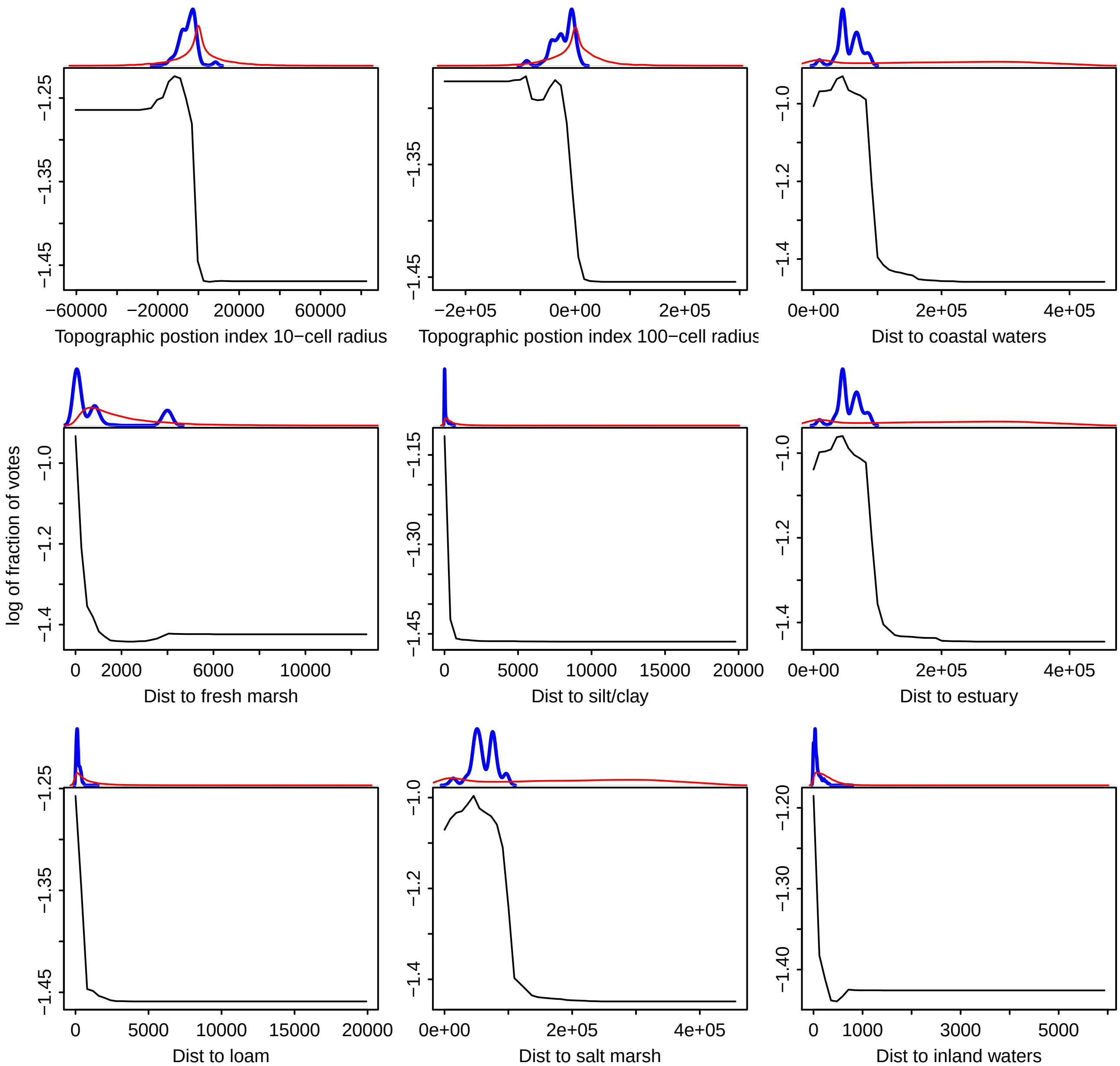


Figure 3. Partial dependence plots for the 9 environmental variables with the most influence on the model. Each plot shows the effect of the variable on the probability of appropriate habitat with the effects of the other variables removed [3]. Peaks in the line indicate where this variable had the strongest 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	0.598	100(28)	97.4(38)	98.6	The probability at which the absolute value of sensitivity minus specificity is minimized.
F-measure with alpha set to 0.01	0.507	100(28)	100(39)	100	The harmonic average of precision and recall, with strong weighting towards recall (classifying presence points as suitable habitat).
Maximum of sensitivity plus specificity	0.535	100(28)	100(39)	99.8	The probability at which the sum of sensitivity (true positive rate) and specificity (true negative rate) is maximized.
Minimum Training Presence	0.507	100(28)	100(39)	100	The lowest probability value assigned to any of the input presence points. 100% of input presence points are classified as suitable habitat.
Minimum Training Presence by Element Occurrence	0.947	100(28)	82.1(32)	56.4	The lowest probability value assigned to any of the input presence element occurrences. This calculation first summarizes EOs by their maximum and then finds the minimum of these values.
Minimum Training Presence by Polygon	0.554	100(28)	100(39)	99	The lowest probability value assigned to any of the input presence polygons.
Tenth percentile of training presence	0.846	100(28)	97.4(38)	90	The probability at which 90% of the input presence points are classified as suitable habitat and 10% are classified as unsuitable.

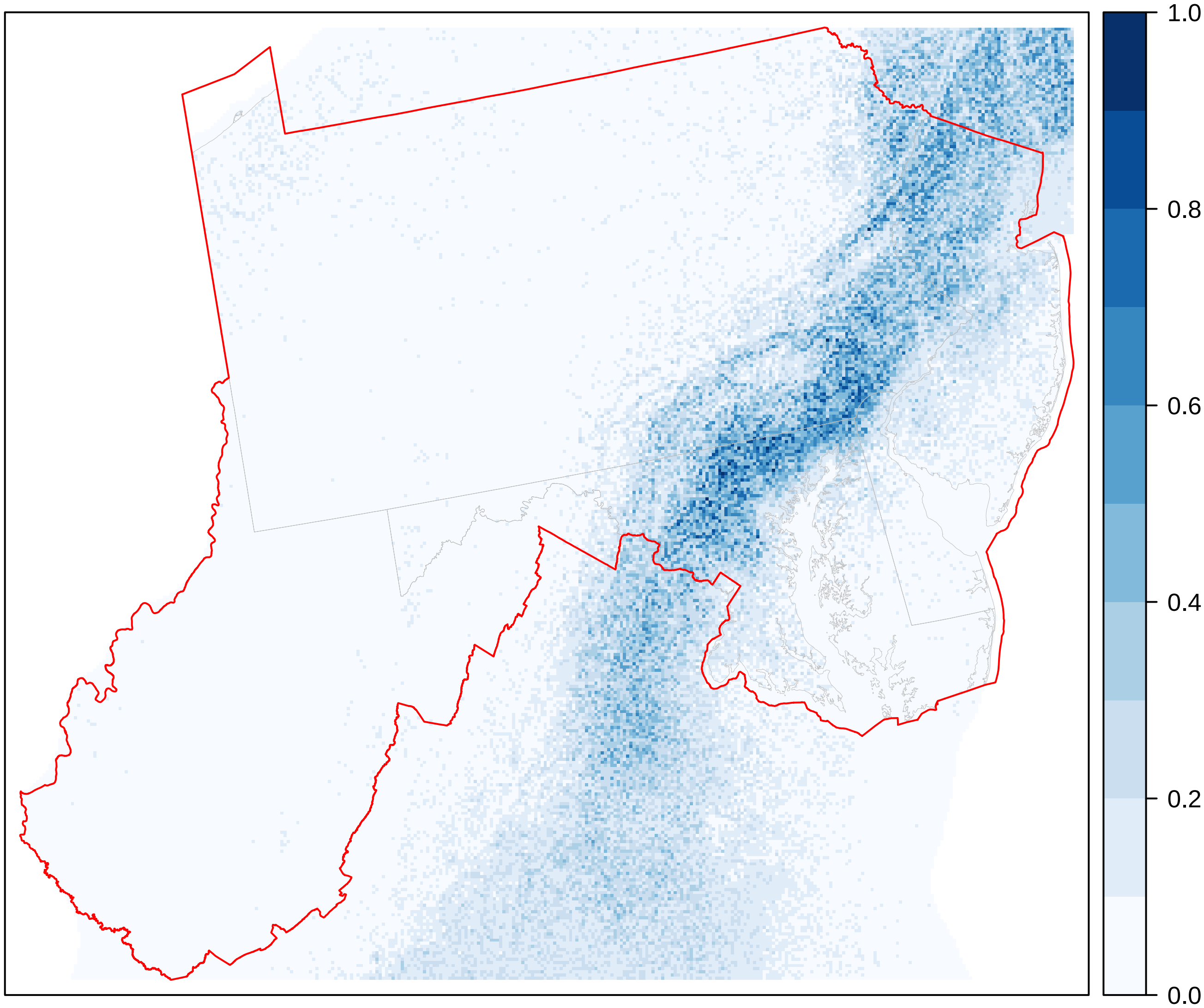


Figure 5. A generalized view of the model predictions throughout the study area. State boundaries are shown in black. The study area is outlined in red.

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

- Maryland Natural Heritage Program, Maryland Department of Natural Resources, Wildlife and Heritage Service
- Pennsylvania 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.

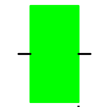
Please cite this document and its associated SDM as:
Pennsylvania Natural Heritage Program. 2017. Species distribution model for Mulberry Wing (*Poanes massasoit*). Created on 02 Dec 2017. Western Pennsylvania Conservancy, Pittsburgh, PA.

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 3.4.2 (2017-09-28).
[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.

Poanes viator viator

Species Distribution Model (SDM) assessment metrics and metadata
Common name: Broad-winged Skipper
Date: 01 Feb 2018
Code: poanvia1



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 8 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); Groups = element occurrence BG points = background points; PR points = presence points placed throughout all polygons.

Name	Number
polys	18
EOs	8
BG points	11473
PR points	1674

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

Name	Mean	SD	SEM
Overall Accuracy	0.99	0.01	0.00
Specificity	1.00	0.01	0.00
Sensitivity	0.98	0.01	0.00
TSS	0.98	0.01	0.00
Kappa	0.98	0.01	0.00
AUC	1.00	0.00	0.00

Validation runs used 54 environmental variables, the most important of 81 variables (top 75 percent). Each tree was built with 4 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 4, and the same number of environmental variables.

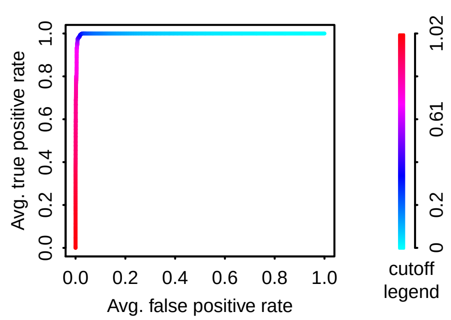


Figure 1. ROC plot for all 8 validation runs, averaged along cutoffs.

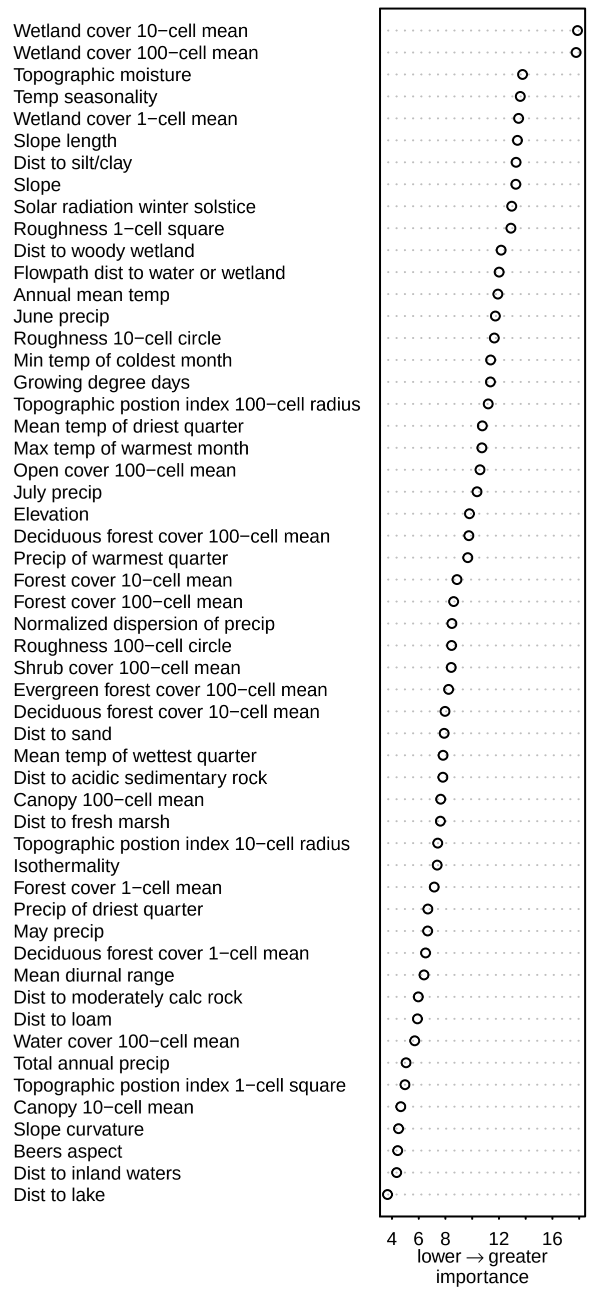


Figure 2. Relative importance of each environmental variable based on the full model using all sites as input. Abbreviations used: calc = calcareous, CP = coastal plain, dist = distance, fresh = freshwater, precip = precipitation, temp = temperature, max = maximum, min = minimum.

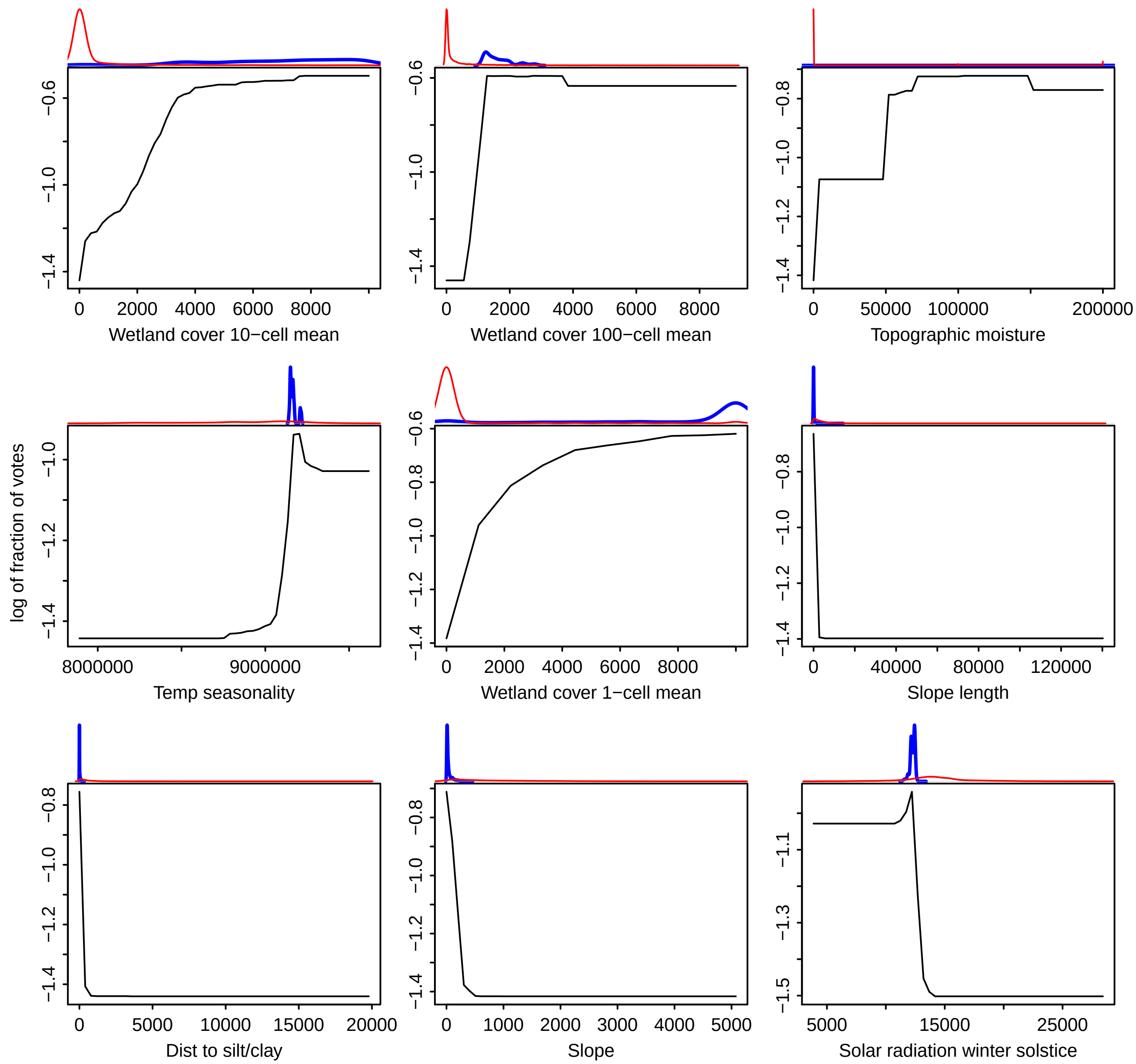


Figure 3. Partial dependence plots for the 9 environmental variables with the most influence on the model. Each plot shows the effect of the variable on the probability of appropriate habitat with the effects of the other variables removed [3]. Peaks in the line indicate where this variable had the strongest 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	0.591	100(8)	100(18)	99.6	The probability at which the absolute value of sensitivity minus specificity is minimized.
F-measure with alpha set to 0.01	0.543	100(8)	100(18)	100	The harmonic average of precision and recall, with strong weighting towards recall (classifying presence points as suitable habitat).
Maximum of sensitivity plus specificity	0.543	100(8)	100(18)	100	The probability at which the sum of sensitivity (true positive rate) and specificity (true negative rate) is maximized.
Minimum Training Presence	0.543	100(8)	100(18)	100	The lowest probability value assigned to any of the input presence points. 100% of input presence points are classified as suitable habitat.
Minimum Training Presence by Element Occurrence	0.997	100(8)	50(9)	7.3	The lowest probability value assigned to any of the input presence element occurrences. This calculation first summarizes EOs by their maximum and then finds the minimum of these values.
Minimum Training Presence by Polygon	0.965	100(8)	100(18)	74.9	The lowest probability value assigned to any of the input presence polygons.
Tenth percentile of training presence	0.879	100(8)	100(18)	90	The probability at which 90% of the input presence points are classified as suitable habitat and 10% are classified as unsuitable.

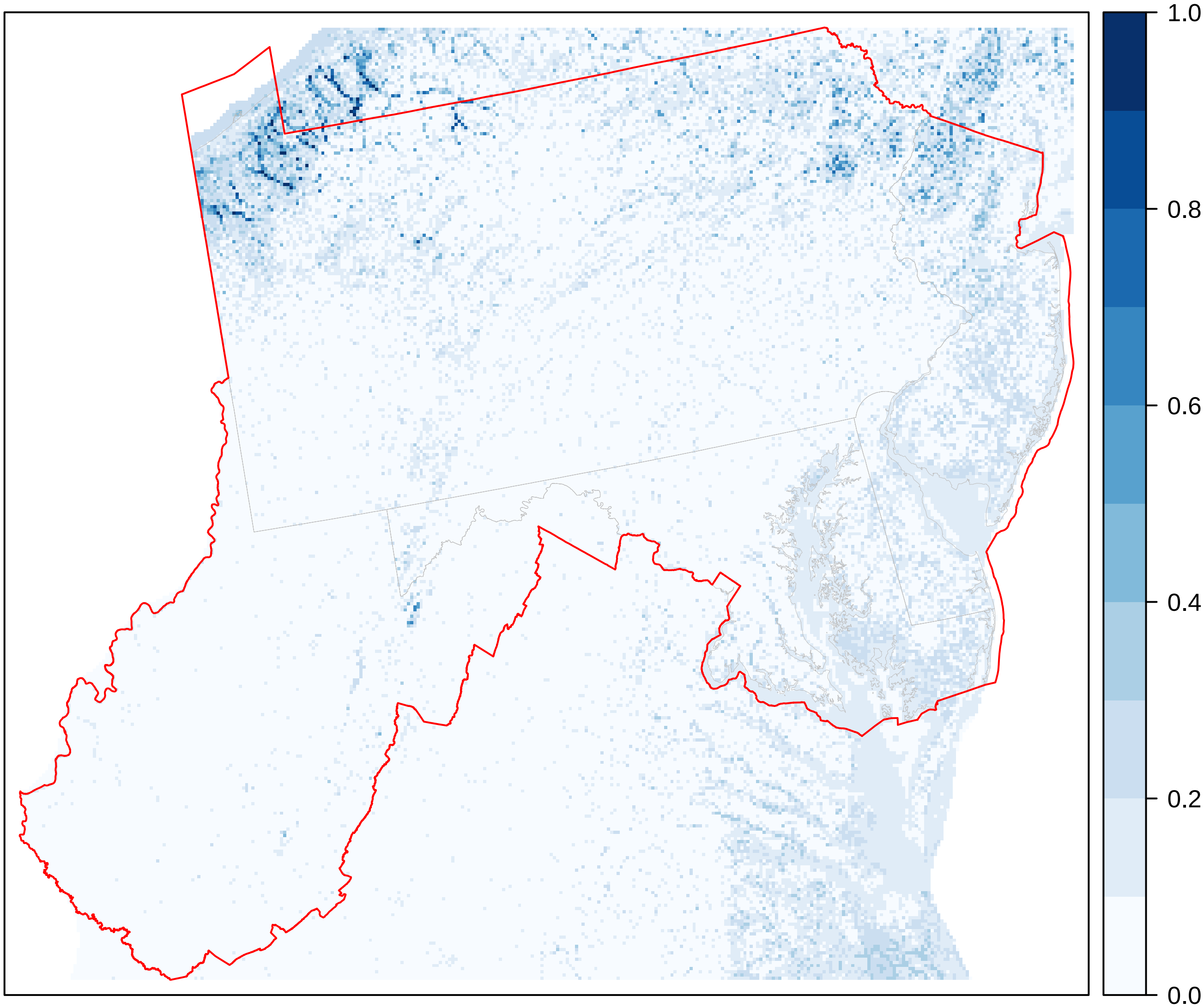


Figure 5. A generalized view of the model predictions throughout the study area. State boundaries are shown in black. The study area is outlined in red.

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

- Maryland Natural Heritage Program, Maryland Department of Natural Resources, Wildlife and Heritage Service
- Pennsylvania 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:
Pennsylvania Natural Heritage Program. 2018. Species distribution model for Broad-winged Skipper (*Poanes viator viator*). Created on 01 Feb 2018. Western Pennsylvania Conservancy, Pittsburgh, PA.

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 3.4.3 (2017-11-30).
[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.

Polites mystic

Species Distribution Model (SDM) assessment metrics and metadata
Common name: Long Dash
Date: 01 Feb 2018
Code: polimyst

fair
TSS=0.78
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 51 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); Groups = element occurrence BG points = background points; PR points = presence points placed throughout all polygons.

Name	Number
polys	69
EOs	51
BG points	11473
PR points	4983

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

Name	Mean	SD	SEM
Overall Accuracy	0.89	0.17	0.02
Specificity	0.85	0.34	0.05
Sensitivity	0.93	0.07	0.01
TSS	0.78	0.34	0.05
Kappa	0.78	0.34	0.05
AUC	0.96	0.10	0.01

Validation runs used 57 environmental variables, the most important of 85 variables (top 75 percent). Each tree was built with 2 variables tried at each split (mtry) and 750 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.

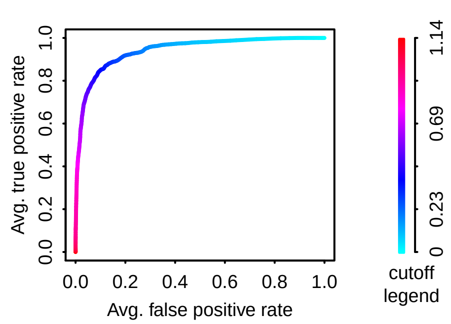


Figure 1. ROC plot for all 51 validation runs, averaged along cutoffs.

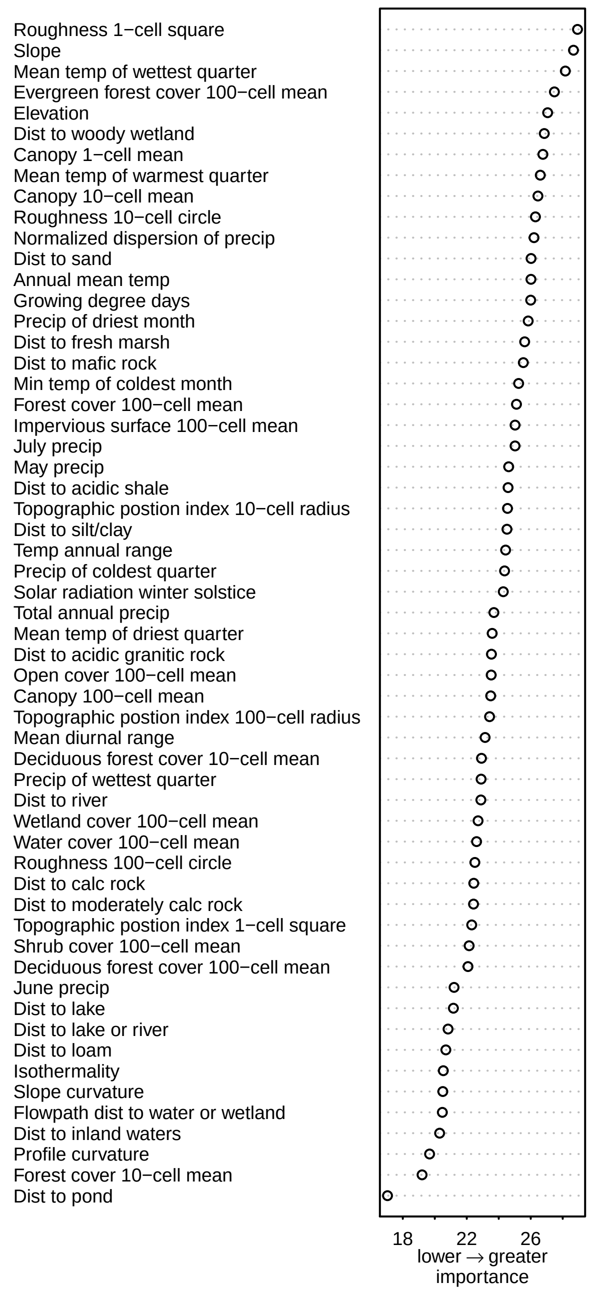


Figure 2. Relative importance of each environmental variable based on the full model using all sites as input. Abbreviations used: calc = calcareous, CP = coastal plain, dist = distance, fresh = freshwater, precip = precipitation, temp = temperature, max = maximum, min = minimum.

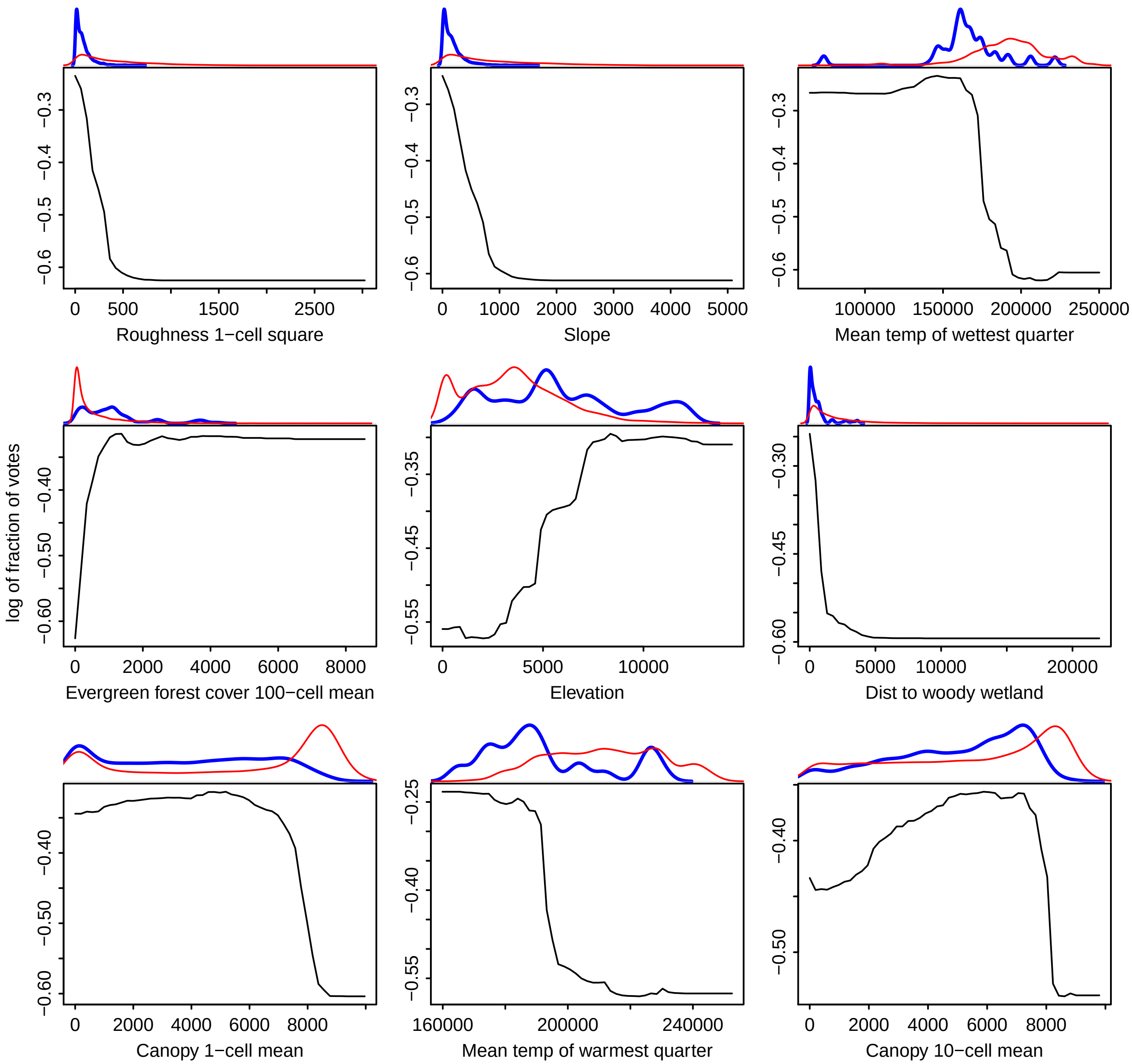


Figure 3. Partial dependence plots for the 9 environmental variables with the most influence on the model. Each plot shows the effect of the variable on the probability of appropriate habitat with the effects of the other variables removed [3]. Peaks in the line indicate where this variable had the strongest 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	0.570	100(51)	98.6(68)	98.7	The probability at which the absolute value of sensitivity minus specificity is minimized.
F-measure with alpha set to 0.01	0.307	100(51)	100(69)	100	The harmonic average of precision and recall, with strong weighting towards recall (classifying presence points as suitable habitat).
Maximum of sensitivity plus specificity	0.582	100(51)	98.6(68)	98.6	The probability at which the sum of sensitivity (true positive rate) and specificity (true negative rate) is maximized.
Minimum Training Presence	0.279	100(51)	100(69)	100	The lowest probability value assigned to any of the input presence points. 100% of input presence points are classified as suitable habitat.
Minimum Training Presence by Element Occurrence	0.836	100(51)	89.9(62)	83.5	The lowest probability value assigned to any of the input presence element occurrences. This calculation first summarizes EOs by their maximum and then finds the minimum of these values.
Minimum Training Presence by Polygon	0.513	100(51)	100(69)	99.2	The lowest probability value assigned to any of the input presence polygons.
Tenth percentile of training presence	0.773	100(51)	92.8(64)	90	The probability at which 90% of the input presence points are classified as suitable habitat and 10% are classified as unsuitable.

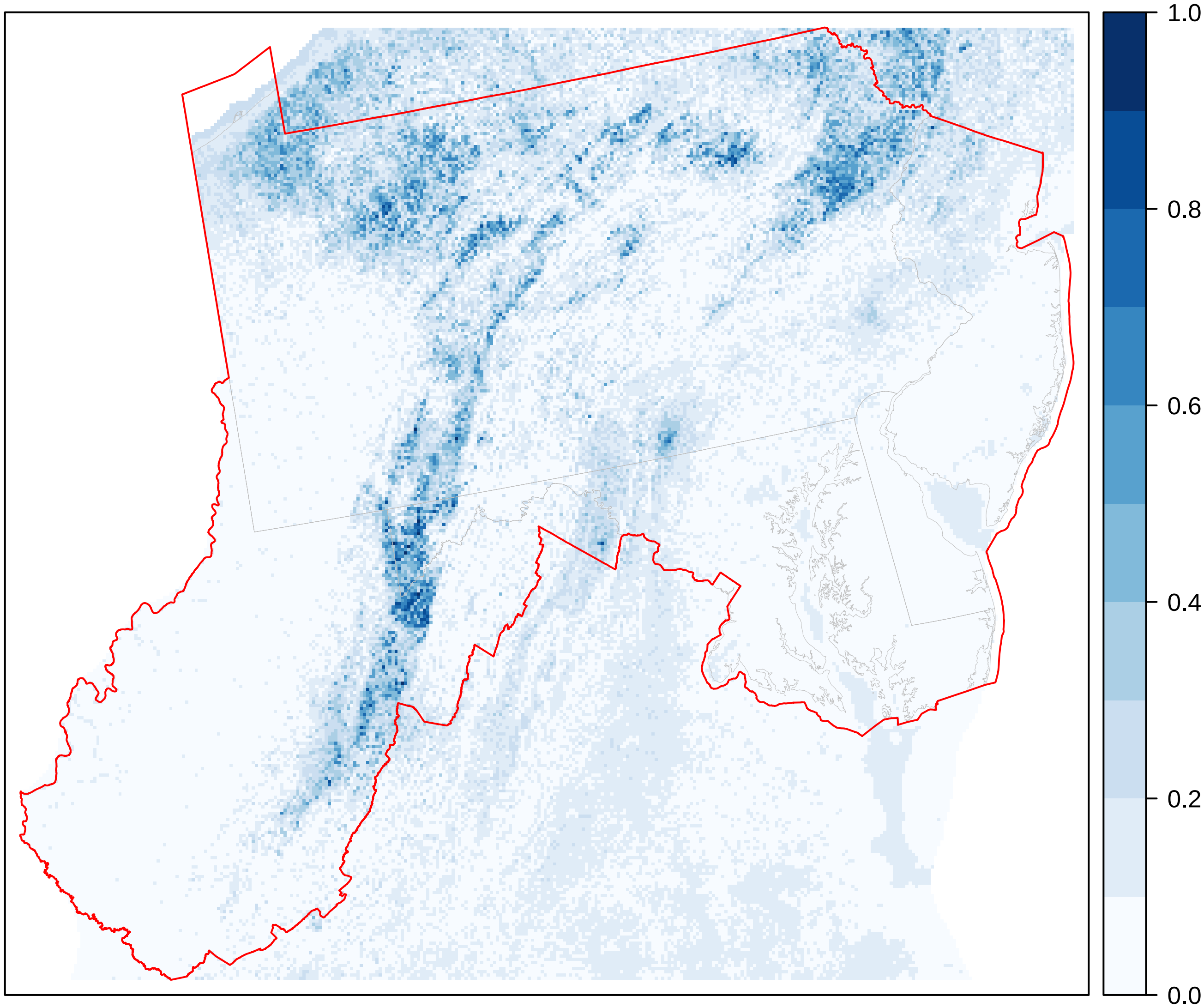


Figure 5. A generalized view of the model predictions throughout the study area. State boundaries are shown in black. The study area is outlined in red.

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

- Maryland Natural Heritage Program, Maryland Department of Natural Resources, Wildlife and Heritage Service
- New Jersey Department of Environmental Protection, Division of Fish and Wildlife, New Jersey Endangered & Nongame Species Program
- Pennsylvania Natural Heritage Program
- West 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:
Pennsylvania Natural Heritage Program. 2018. Species distribution model for Long Dash (*Polites mystic*). Created on 01 Feb 2018. Western Pennsylvania Conservancy, Pittsburgh, PA.

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 3.4.3 (2017-11-30).
[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.