| Anatrutone logan | |
|---|---------------------------|
| Species Distribution Model (SDM) assessment metrics and metadata Common name: Delaware Skipper | fair |
| Date: 17 Nov 2017 | TSS=0.74 |
| 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.

| Dist to mafic rock | 0 |
|---|--|
| Evergreen forest cover 100-cell mean | · · · · · · · · · · · · · · · · · · · |
| Deciduous forest cover 10-cell mean | • |
| Water cover 100-cell mean | • |
| Dist to acidic granitic rock | · · · · · · · · · · · · · · · · · · · |
| Mean temp of wettest quarter | · · · · · · · · · · · · · · · · · · · |
| Dist to ultramafic rock | · · · · · · · · · · · · · · · · · · · |
| Canopy 1-cell mean | |
| Precip of driest quarter | • |
| Topographic postion index 10-cell radius | · · · · · · · · · · · · · · · · · · · |
| Dist to estuary | |
| Impervious surface 10-cell mean | |
| Dist to coastal waters | · · · · · · · · · · · · · · · · · · · |
| Dist to lake or river | • |
| Dist to river | • |
| Flowpath dist to water or wetland | · · · · · · · · · · · · · · · · · · · |
| Precip of coldest quarter | · · · · · · · · · · · · · · · · · · · |
| May precip | |
| Dist to calc rock | |
| Dist to salt marsh | |
| Impervious surface 100-cell mean | · · · · · · · · · · · · · · · · · · · |
| Open cover 100-cell mean | |
| Isothermality | |
| Total annual precip | |
| Dist to acidic shale | |
| Topographic postion index 100-cell radius | |
| Dist to loam | |
| Temp annual range | |
| Canopy 10-cell mean | |
| Forest cover 10-cell mean | |
| Dist to lake | |
| Elevation | |
| Slope | |
| Dist to fresh marsh | |
| Roughness 1–cell square | |
| Dist to woody wetland | |
| Dist to sand | |
| Precip of warmest quarter | |
| Mean temp of driest quarter | |
| Dist to silt/clay | |
| July precip | |
| June precip | · · · · · · · 0 ⁻ · · · · · · · · · · · · · · · · |
| Mean diurnal range | |
| Water cover 10-cell mean | |
| Slope length | |
| Max temp of warmest month | 0 |
| Normalized dispersion of precip | |
| Dist to pond | • |
| Solar radiation winter solstice | · · · · 0 · · · · · · · · · · · · · · · |
| Dist to inland waters | · · · · 0 · · · · · · · · · · · · · · · |
| Forest cover 100-cell mean | |
| Canopy 100-cell mean | 0 |
| Roughness 10-cell circle | 0 |
| Wetland cover 100-cell mean | 0 |
| Shrub cover 100-cell mean | 0 |
| Annual mean temp | 0 |
| Mean temp of coldest quarter | 0 |
| Dist to moderately calc rock | 0 |
| Dist to acidic sedimentary rock | 0 |
| Deciduous forest cover 100-cell mean | 0 |
| | L |
| | $\begin{array}{cccc} 18 & 20 & 22 & 24 \\ lower \rightarrow greater \end{array}$ |

importance





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.

| Threshold | Value | EOs | Polys | Pts | Description |
|------------------------------------|--------|---------|---------|------|--|
| Equal sensitivity and specificity | 0.588 | 100(43) | 100(46) | 99.5 | The probability at which the absolute |
| | 0.0000 | 100(10) | 100(10) | 00.0 | 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 (classifying presence points as |
| | | | | | suitable habitat). |
| Maximum of sensitivity plus | 0.562 | 100(43) | 100(46) | 100 | The probability at which the sum |
| specificity | | | | | of sensitivity (true positive rate) and |
| | | | | | imized. |
| 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 clas- |
| Minimum Training Presence by | 0.836 | 100(43) | 100(46) | 71 | The lowest probability value assigned |
| Element Occurrence | 0.000 | 100(10) | 100(10) | | 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 | 0.836 | 100(43) | 100(46) | 71 | The lowest probability value assigned |
| Polygon | | | . , | | to any of the input presence polygons. |
| Tenth percentile of training pres- | 0.715 | 100(43) | 100(46) | 90 | The probability at which 90% of the |
| ence | | | | | input presence points are classified as 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.

- 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.

- [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 pres-
- [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 TSS=0.85 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.

| Dist to fresh marsh | |
|--|---|
| Dist to woody wetland | • |
| Roughness 1–cell square | ••••••••••••••••••••••••••••••••••••••• |
| Dist to ocean | ••••••••••••••••••••••••••••••••••••••• |
| Slope | • • • • • • • • • • • • • • • • • • • |
| Slope length | • |
| Flowpath dist to water or wetland | • |
| Max temp of warmest month | · · · · · · · · · · · · · · · · · · · |
| Wetland cover 10-cell mean | · · · · · · · · · · · · O · · · · · · · |
| Canopy 1-cell mean | · · · · · · · · · · · · O · · · · · · |
| Roughness 10-cell circle | • |
| Evergreen forest cover 100-cell mean | · · · · · · · · · · · O · · · · · · · · |
| Dist to coastal waters | • |
| Growing degree days | • |
| Topographic postion index 10-cell radius | • |
| Open cover 100-cell mean | • |
| Precip of driest month | · · · · · · · · · O · · · · · · · · |
| Dist to estuary | · · · · · · · · · O · · · · · · · · |
| Canopy 10-cell mean | · · · · · · · · · O · · · · · · · · |
| Dist to salt marsh | · · · · · · · · · O · · · · · · · · · |
| Mean temp of wettest guarter | · · · · · · · · · O · · · · · · · · · |
| Annual mean temp | • |
| Water cover 100-cell mean | · · · · · · · · O · · · · · · · · · · |
| Precip of wettest quarter | · · · · · · · O · · · · · · · · · · |
| Canopy 100-cell mean | · · · · · · · O · · · · · · · · · · · |
| Precip of coldest quarter | • |
| Wetland cover 100-cell mean | • |
| Mean temp of coldest quarter | •••••••••••• |
| Impervious surface 100-cell mean | •••••• |
| Topographic moisture | • |
| Dist to moderately calc rock | · · · · · · · · · · · · · · · · · · · |
| July precip | · · · · · · · · · · · · · · · · · · · |
| Elevation | • |
| Dist to inland waters | • |
| Solar radiation summer solstice | • |
| Dist to lake | • |
| Dist to calc rock | • |
| May precip | ••••• |
| Dist to sand | •••••• |
| Dist to silt/clay | • |
| Forest cover 100–cell mean | • |
| Roughness 100–cell circle | • |
| Open cover 10–cell mean | • |
| Normalized dispersion of precip | • |
| Deciduous forest cover 100-cell mean | • |
| Dist to lake or river | • |
| Temp annual range | 0 |
| Total annual precip | 0 |
| Dist to acidic granitic rock | |
| Dist to matic rock | 0 |
| Deciduous forest cover 10-cell mean | 0 |
| lopographic postion index 1-cell square | |
| Impervious surface 10-cell mean | - 0 |
| Mean diurnal range | 0 |
| Prome curvature | 2 |
| Dist to acidic shale | |
| Dist to loan | l õ |
| Everyieen lorest cover 10-cell mean | |
| Dist to poidio sodimontory rock | Ğ |
| Dist to actuic securiterilary TUCK | Ľ |
| | 20 22 24 26 28 lower → greater |

importance

good

0

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, $\max = \max \min, \min = \min \max$.

30



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.

| Threshold | Valuo | FOr | Polys | Dte | Description |
|--|-------|---------|--------------------------|------|---|
| | Value | 100(57) | $\frac{101ys}{005(010)}$ | 1 15 | |
| Equal sensitivity and specificity | 0.603 | 100(57) | 99.5(210) | 98.9 | 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 max- imized. |
| 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 clas- sified 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 pres- ence | 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.

- 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.

- [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 pres-
- [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] Liù, Ć., P. M. Berry, T. P. Dawson, and R. G. Pearson. 2005. Selecting thresholds of occurrence in the prediction of species distributions. Ecography 28:3857393.
- [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 good Common name: Arctic Skipper TSS=0.98 Date: 18 Nov 2017 ability to find new sites Code: cartpala

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.

| Wetland cover 10–cell mean Mean temp of coldest quarter | |
|--|---|
| Mean temp of wettest quarter | |
| Annual mean temp | |
| Max temp of warmest month | |
| Mean temp of driest quarter | |
| Dist to lake | |
| June precip | |
| Growing degree days | |
| Wetland cover 100-cell mean | |
| Flevation | ····· |
| Temp seasonality | ····· |
| Total annual precip | ······ |
| Impervious surface 100-cell mean | 0 |
| Flowpath dist to water or wetland | 0 |
| May precip | |
| Precip of warmest quarter | |
| Water cover 100-cell mean | 0 |
| Dist to inland waters | |
| Dist to fresh marsh | 0 |
| Isothermality | · · · · · · · · · · · · · · · · · · · |
| Forest cover 10-cell mean | 0 |
| Wetland cover 1-cell mean | 0 |
| Topographic postion index 10-cell radius | |
| Deciduous forest cover 100-cell mean | 0 |
| Dist to woody wetland | |
| Roughness 10-cell circle | · · · · · · · 0 · · · · · · |
| Mean diurnal range | |
| Topographic postion index 100-cell radius | |
| Forest cover 100-cell mean | |
| Roughness 1-cell square | |
| Dist to lake or river | |
| Deciduous forest cover 10-cell mean | · · · · · 0 · · · · · · · · |
| July precip | · · · · · 0 · · · · · · · · |
| Canopy 100-cell mean | · · · · · o · · · · · · · · |
| Slope | · · · · · o · · · · · · · · |
| Dist to silt/clav | |
| Slope length | |
| Evergreen forest cover 10-cell mean | 0 |
| Dist to acidic sedimentary rock | 0 |
| Open cover 100-cell mean | 0 |
| Shrub cover 100-cell mean | |
| Precip of coldest guarter | |
| Solar radiation winter solstice | |
| Topographic moisture | |
| Dist to river | |
| Dist to pond | 0. |
| Canopy 10-cell mean | |
| Normalized dispersion of precip | |
| Dist to moderately calc rock | • |
| Deciduous forest cover 1-cell mean | |
| Canopy 1-cell mean | 0 |
| Precip of driest month | 0 |
| Dist to stream | 0 |
| | 8 10 12 lower \rightarrow greater |

reater importance

14

16

0

0

o 0

0 0

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, $\max = \max \min, \min = \min \max$.



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.

| 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 max- imized. |
| 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 clas- sified 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 pres- ence | 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.

- 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.

- [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.
- 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.
- 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 | good |
| Date: 30 Jan 2018 | TSS=0.81 |
| 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.

| Growing degree days | |
|---|--|
| Canopy 10–cell mean | |
| Annual mean temp | 0 |
| Roughness 1-cell square | · · · · · · · · · · · · · · · · · · · |
| Evergreen forest cover 100-cell mean | |
| Canopy 1-cell mean | |
| Dist to silt/clay | |
| Dist to silivitay | |
| max temp of warmest month | |
| Dist to woody wetland | 0 |
| Dist to fresh marsh | • • • • • • • • • • • • • • • • • • • |
| Mean temp of wettest quarter | 0 |
| Dist to lake | · · · · · · · · · · · · · · · · O · · · · |
| Mean temp of coldest quarter | · · · · · · · · · · · · · · · · · · · |
| Slope | o |
| Open cover 100-cell mean | |
| | , in the second se |
| | |
| Dist to acidic granitic rock | |
| Topographic postion index 10–cell radius | 0 |
| Wetland cover 100–cell mean | ····· |
| Roughness 100–cell circle | 0 |
| Precip of coldest guarter | · · · · · · · · · · · · · O · · · · · · |
| Dist to mafic rock | 0 |
| Topographic postion index 100-cell radius | · · · · · · · · · · · · · · · · · · · |
| Canopy 100-cell mean | |
| Nermelized dispersion of presin | , , , , , , , , , , , , , , , , , , , |
| | ő |
| Temp seasonality | |
| Dist to lake or river | O |
| Roughness 10–cell circle | · · · · · · · · · · · O · · · · · · · · |
| Precip of driest month | 0 |
| Dist to pond | · · · · · · · · · O · · · · · · · · · |
| Dist to sand | · · · · · · · · · O · · · · · · · · · |
| Isothermality | • • • • • • • • • • • • • • • • • • • |
| Dist to loam | |
| Dist to river | · · · · · · · · · · · · · · · · · · · |
| Dist to moderately calc rock | |
| May provin | ŏ |
| | |
| Precip of wettest month | |
| Mean diurnal range | 0 |
| Total annual precip | 0 |
| Water cover 100–cell mean | · · · · · · O · · · · · · · · · · · · |
| Open cover 10–cell mean | · · · · · · · O · · · · · · · · · · · |
| Wetland cover 10-cell mean | · · · · · · O · · · · · · · · · · · · |
| Impervious surface 100-cell mean | 0 |
| Forest cover 100–cell mean | 0 |
| | |
| July procip | ŏ |
| Diat to cale reals | 0 |
| | |
| Deciduous forest cover 10-cell mean | |
| Evergreen forest cover 10–cell mean | ····· |
| Dist to acidic shale | ••••• |
| Forest cover 10–cell mean | · · · · O · · · · · · · · · · · · · · · |
| Shrub cover 100–cell mean | ···· 0 · · · · · · · · · · · · · · · · |
| Deciduous forest cover 100-cell mean | · · · · o · · · · · · · · · · · · · · · |
| Mean temp of driest quarter | |
| Solar radiation summer solstice | |
| Impervious surface 10 coll moon | ام - <u>-</u> |
| Dist to acidic sodimentary rock | č |
| Dist to aciulo seulmentary took | Ľ |
| | |
| | 20 22 24 26 |

lower → greater importance

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.

0

0



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.

| Equal sensitivity and specificity0.669100(55)100(78)99.4The probability at which the absolute value of sensitivity minus specificity is minimized.F-measure with alpha set to 0.010.529100(55)100(78)100The harmonic average of precision and recall, with strong weighting towards recall (classifying presence points as suitable habitat). | Threshold | Value | EOs | Polys | Pts | Description |
|---|------------------------------------|-------|---------|---------------------------|------|--|
| F-measure with alpha set to 0.01 0.529 100(55) 100(78) 100(78) 100 The harmonic average of precision and recall, with strong weighting towards recall (classifying presence points as suitable habitat). | Equal sensitivity and specificity | 0.660 | 100(55) | $\frac{100(78)}{100(78)}$ | 99.4 | The probability at which the absolute |
| 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). | Equal sensitivity and specificity | 0.003 | 100(00) | 100(10) | 33.4 | value of sensitivity minus specificity is |
| 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). | | | | | | minimized. |
| recall, with strong weighting towards recall (classifying presence points as suitable habitat). | F-measure with alpha set to 0.01 | 0.529 | 100(55) | 100(78) | 100 | The harmonic average of precision and |
| recall (classifying presence points as suitable habitat). | | | | | | recall, with strong weighting towards |
| suitable habitat). | | | | | | recall (classifying presence points as |
| M_{2} | | 0.695 | 100(55) | 100(79) | 00.2 | suitable habitat). |
| Maximum of sensitivity plus $0.085 + 100(55) + 100(78) + 99.3$ The probability at which the sum specificity of sensitivity (true positive rate) and | specificity | 0.085 | 100(55) | 100(78) | 99.3 | of sensitivity (true positive rate) and |
| specificity (true positive rate) is max- | specificity | | | | | specificity (true pegative rate) is max- |
| imized. | | | | | | imized. |
| Minimum Training Presence $0.529 100(55) 100(78) 100$ The lowest probability value assigned | Minimum Training Presence | 0.529 | 100(55) | 100(78) | 100 | The lowest probability value assigned |
| to any of the input presence points. | | | | | | to any of the input presence points. |
| 100% of input presence points are clas- | | | | | | 100% of input presence points are clas- |
| sified as suitable habitat. | | 0.000 | 100/55) | 00 5(00) | cc 7 | sified as suitable habitat. |
| Minimum Training Presence by $0.929 = 100(55) = 88.5(69) = 66.7$ The lowest probability value assigned Element Occurrence | Minimum Training Presence by | 0.929 | 100(55) | 88.5(69) | 66.7 | The lowest probability value assigned to any of the input presence element |
| concurrences This calculation first | Element Occurrence | | | | | occurrences This calculation first |
| summarizes EOs by their maximum | | | | | | summarizes EOs by their maximum |
| and then finds the minimum of these | | | | | | and then finds the minimum of these |
| values. | | | | | | values. |
| Minimum Training Presence by 0.697 100(55) 100(78) 99.1 The lowest probability value assigned | Minimum Training Presence by | 0.697 | 100(55) | 100(78) | 99.1 | The lowest probability value assigned |
| Polygon to any of the input presence polygons. | Polygon | | | | | to any of the input presence polygons. |
| Tenth percentile of training pres- $0.864 \ 100(55) \ 96.2(75) \ 90$ The probability at which 90% of the | Tenth percentile of training pres- | 0.864 | 100(55) | 96.2(75) | 90 | The probability at which 90% of the |
| ence input presence points are classified as | ence | | | | | suitable babitat and 10% are classified |
| as unsuitable. | | | | | | 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.

- 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 har-risii*). Created on 30 Jan 2018. Western Pennsylvania Conservancy, Pittsburgh, PA.

- [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 pres-
- [6] Fielding, A. H. 2002. What are the appropriate characteristics of an accuracy measure? Pages 271-280 in Predicting
- [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] Liù, Ć., 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 | good |
| Date: 30 Jan 2018 | TSS=0.93 |
| 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.

| Evergreen forest cover 100-cell mean | |
|---|---|
| Roughness 10-cell circle | |
| Dist to fresh marsh | ·····o |
| Slope | · · · · · · · · · · · · · · · · · · · |
| Mean temp of wettest quarter | · · · · · · · · · · · · · · · · · · · |
| Canopy 1-cell mean | · · · · · · · · · · · · · · · · · · · |
| Roughness 1-cell square | ····· |
| Open cover 100-cell mean | · · · · · · · · · · · · · · · · · · · |
| Topographic postion index 10-cell radius | · · · · · · · · · · · · · · · · · · · |
| Impervious surface 100-cell mean | · · · · · · · · · · · · · · · · · · • |
| Mean temp of coldest guarter | · · · · · · · · · · · · · · · · · · · |
| Annual mean temp | · · · · · · · · · · · · · · · · · · · |
| Wetland cover 10-cell mean | · · · · · · · · · · · · · • • • • • • • |
| Growing degree days | ····· |
| Elevation | · · · · · · · · · · · · · · · · · · · |
| Mean temp of warmest guarter | · · · · · · · · · · · · · · · · · · · |
| Dist to coastal waters | · · · · · · · · · · · · · · · · · · · |
| Dist to woody wetland | · · · · · · · · · · · · • • • • • • • • |
| Temp annual range | · · · · · · · · · · · · · · · · · · · |
| Slope length | · · · · · · · · · · · · o · · · · · · |
| Dist to silt/clay | • • • • • • • • • • • • • • • • • • • |
| Canopy 10-cell mean | 0 |
| Isothermality | · · · · · · · · · · · · · · · · · · · |
| Flowpath dist to water or wetland | · · · · · · · · · · · · · · · · · · · |
| June precip | · · · · · · · · · · · O · · · · · · · · |
| Canopy 100-cell mean | · · · · · · · · · · · · · · · · · · · |
| Dist to moderately calc rock | 0 |
| Dist to inland waters | 0 |
| Solar radiation summer solstice | · · · · · · · · · · · · · · · · · · · |
| Dist to salt marsh | 0 |
| Dist to lake | · · · · · · · · · · · · · · · · · · · |
| Dist to estuary | •••••• |
| Forest cover 100-cell mean | · · · · · · · · · 0 · · · · · · · · |
| Dist to stream | · · · · · · · · · · · · · · · · · · · |
| Topographic moisture | · · · · · · · · O · · · · · · · · |
| Mean diurnal range | · · · · · · · · O · · · · · · · · · |
| Total annual precip | 0 |
| Normalized dispersion of precip | •••••• |
| Dist to calc rock | · · · · · · O · · · · · · · · · · |
| Precip of warmest quarter | •••••• |
| Dist to acidic shale | •••••• |
| Open cover 10-cell mean | •••••• |
| Precip of driest quarter | 0 |
| Mean temp of driest quarter | ••••• |
| Dist to sand | · · · · · O · · · · · · · · · · · |
| July precip | •••••• |
| May precip | ••••• |
| Topographic postion index 100–cell radius | ····· |
| Topographic postion index 1–cell square | ····· |
| Deciduous forest cover 10-cell mean | ···· • • • • • • • • • • • • • • • • • |
| Deciduous forest cover 100-cell mean | 0 |
| Dist to acidic sedimentary rock | ····• |
| Dist to lake or river | · · · · O· · · · · · · · · · · · · · |
| Deciduous forest cover 1–cell mean | • • • • • • • • • • • • • • • • • • • |
| Roughness 100-cell circle | •••••••••••••••••••••••••••••••••••••• |
| Precip of coldest quarter | - O |
| Dist to pond | 0 |
| Profile curvature | 0 |
| | 14 16 18 20 2 lower → greater |
| | |

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.

22

0



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.

| Threshold | Value | EOs | Polvs | 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 max- imized. |
| 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 clas- sified 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 pres- ence | 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.

- Maryland Natural Heritage Program, Maryland Department of Natural Resources, Wildlife and Heritage Service
- New Jersev Department of Environmental Protection, Division of Fish and Wildlife, New Jersev 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.

- 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.
- 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 pres-
- ence/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.
- 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 | good |
| Date: 09 Dec 2017 | TSS=0.86 |
| 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.

| Slope Roughness 1–cell square | |
|---|---|
| Dist to silt/clay | • |
| Roughness 10–cell circle | • • • • • • • • • • • • • • • • • • • |
| Dist to woody wetland | • • • • • • • • • • • • • • • • • • • |
| Topographic postion index 10–cell radius | 0 |
| Canopy 1-cell mean | • • • • • • • • • • • • • • • • • • • |
| Elevation | • • • • • • • • • • • • • • • • • • • |
| May precip | 0 |
| Temp seasonality | 0 |
| Dist to acidic shale | 0 |
| Dist to salt marsh | 0 |
| Dist to fresh marsh | ••••• |
| Canopy 10-cell mean | - |
| Precip of driest month | Θ |
| Mean temp of driest quarter | 0 |
| Precip of coldest quarter | 0 |
| Topographic postion index 100-cell radius | |
| Forest cover 100-cell mean | 0 |
| Roughness 100-cell circle | 0 |
| Slene length | |
| Slope length | 0 |
| Concert 100, coll macon | , in the second s |
| Evergroop forget gover 100, cell mean | 0 |
| Deciduous forest cover 100-cell mean | ő |
| Deciduous lotest cover 100-cell mean | 0 |
| Dist to coostal waters | ő |
| Dist to estuary | |
| Dist to lake | |
| Isothermality | |
| Normalized dispersion of precip | |
| Solar radiation summer solstice | |
| Total annual precip | |
| Open cover 10-cell mean | |
| Impervious surface 100-cell mean | · · · · · · · · · · · · · · · · · · · |
| Deciduous forest cover 10-cell mean | • |
| Mean temp of coldest guarter | |
| June precip | • • • • • • • • • • • • • • • • • • • |
| Forest cover 10-cell mean | • |
| Growing degree days | • |
| Dist to acidic sedimentary rock | • |
| Max temp of warmest month | • |
| Topographic postion index 1–cell square | 0 |
| Mean temp of wettest quarter | ••••• |
| Dist to moderately calc rock | |
| Impervious surface 10–cell mean | 0 |
| Wetland cover 100-cell mean | 0 |
| Dist to acidic granitic rock | |
| Dist to lake or river | |
| Dist to inland waters | |
| Dist to malic fock | 0 |
| Diet to river | 0 |
| Water cover 100, coll mean | 0 |
| Dist to loam | |
| Dist to ultramafic rock | õ |
| Slope curvature | õ |
| Dist to stream | o |
| Profile curvature | o |
| | Ĺ,,_, |
| | 22 24 26 28 30 lower \rightarrow greater |

importance

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.

| Threshold | Value | EOs | Polvs | Pts | Description |
|--|-------|----------|-----------|------|---|
| Equal sensitivity and specificity | 0.601 | 100(76) | 100(113) | 99 | The probability at which the absolute |
| | | <i>.</i> | | | 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 max- imized. |
| 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 clas- sified 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 pres- ence | 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.

- 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.

- [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.
- 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 pres-
- ence/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.
- 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 | - <mark>-</mark> - |
|--|---------------------------|
| Species Distribution Model (SDM) assessment metrics and metadata | |
| Common name: Dion Skipper | fair |
| Date: 19 Nov 2017 | TSS=0.77 |
| 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.

Wetland cover 10-cell mean

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.

| Wetland cover 100-cell mean | |
|---|---------------------------------------|
| Shrub cover 100-cell mean | |
| Dist to fresh marsh | |
| Water cover 100-cell mean | |
| Wetland cover 1-cell mean | |
| Flowpath dist to water or wetland | |
| Topographic moisture | |
| Dist to loam | |
| Annual mean temp | |
| Dist to woody wetland | |
| Dist to lake | |
| Isothermality | |
| Temp annual range | |
| Dist to calc rock | |
| Dist to mafic rock | |
| Slope length | |
| Dist to silt/clay | |
| Canopy 10-cell mean | |
| Topographic postion index 10-cell radius | |
| lune precip | |
| Dist to lake or river | ດັ |
| Impervious surface 100-cell mean | |
| Max temp of warmest month | |
| Mean diurnal range | |
| Dist to acidic granitic rock | |
| Growing degree days | |
| Forest cover 10_cell mean | |
| Mean temp of coldest quarter | |
| Precip of warmest quarter | |
| Dist to estuary | |
| Mean temp of wettest quarter | |
| Total annual precip | |
| Canony 1_cell mean | |
| Roughness 100-cell circle | |
| Impervious surface 10–cell mean | |
| Dist to coastal waters | |
| Slope | |
| Evergreen forest cover 100–cell mean | · · · · · o · · · · · |
| Roughness 1-cell square | |
| Canopy 100-cell mean | |
| Dist to sand | · · · · · · · · · · · · · · · · · · · |
| Dist to inland waters | · · · · o · · · · · |
| Forest cover 1–cell mean | · · · · o · · · · · · |
| Open cover 10-cell mean | · · · · o · · · · · · |
| Dist to acidic shale | 0 |
| Open cover 100-cell mean | 0 |
| Precip of coldest guarter | |
| Dist to moderately calc rock | |
| Normalized dispersion of precip | |
| Mean temp of driest guarter | |
| Deciduous forest cover 10-cell mean | |
| Roughness 10-cell circle | |
| Dist to salt marsh | 0 |
| Precip of driest quarter | |
| Deciduous forest cover 1-cell mean | 0 |
| Topographic postion index 100-cell radius | |
| Dist to river | |
| Solar radiation equinox | 0 |
| Open cover 1-cell mean | 0 |
| | L, |
| | 12 14 Iower → |

→ greater importance

16 18

0

o

0

0

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, $\max = \max \min, \min = \min \max$.



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.

| Threshold | Value | EOs | Polys | Pts | Description |
|--|-------|---------|----------|------|---|
| Equal sensitivity and specificity | 0.654 | 100(17) | 100(22) | 98.9 | The probability at which the absolute |
| | | | | | 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 max- imized. |
| 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 clas- sified 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 pres- ence | 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.

- 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.

- [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.
- 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 pres-
- ence/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.

| air |
|------------------------|
| =0.78 ind 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.

| Deciduous forest cover 100-cell mean | |
|---|--|
| Canopy 100-cell mean | |
| Elevation | |
| Forest cover 100-cell mean | |
| Dist to moderately calc rock | 0 |
| Dist to lake | 0 |
| Impervious surface 10-cell mean | 0 |
| Mean temp of coldest guarter | 0 |
| Isothermality | |
| Evergreen forest cover 100-cell mean | |
| Dist to sand | |
| Wetland cover 100-cell mean | 0 |
| Max temp of warmest month | |
| Topographic postion index 10, coll radius | , and the second s |
| Dist to acidic sedimentary rock | |
| Dist to ultramatic rock | , i i i i i i i i i i i i i i i i i i i |
| Earoat aguar 10, coll magn | ő |
| | |
| Diet to pand | |
| | 0 |
| Canopy 10-cell mean | |
| Precip of wettest quarter | 0 |
| Solar radiation winter solstice | |
| Dist to fresh marsh | 0 |
| Roughness 10–cell circle | 0 |
| June precip | 0 |
| Roughness 100–cell circle | 0 |
| Mean temp of driest quarter | 0 |
| Dist to woody wetland | 0 |
| Dist to mafic rock | 0 |
| Dist to acidic shale | 0 |
| Slope | ····· |
| Dist to loam | O |
| Dist to estuary | 0 |
| Dist to coastal waters | 0 |
| Dist to lake or river | 0 |
| Dist to silt/clay | •••••••••••••••••••••••••••••••••••••• |
| Impervious surface 100–cell mean | 0 |
| July precip | • • • • • • • • • • • • • • • • • • • |
| May precip | ····· |
| Roughness 1–cell square | •••••••••••••••••••••••••••••••••••••• |
| Mean temp of wettest quarter | 0 |
| Annual mean temp | 0 |
| Precip of driest month | ····· |
| Dist to salt marsh | 0 |
| Open cover 10–cell mean | 0 |
| Dist to acidic granitic rock | 0 |
| Canopy 1-cell mean | 0 |
| Growing degree days | 0 |
| Mean diurnal range | 0 |
| Dist to stream | 0 |
| Flowpath dist to water or wetland | 0 |
| Precip of coldest quarter | 0 |
| Dist to calc rock | · · · · O · · · · · · · · · · · · · · · |
| Normalized dispersion of precip | |
| Open cover 100-cell mean | |
| Total annual precip | 0 |
| Dist to inland waters | 0 |
| Topographic postion index 100-cell radius | 0 |
| Slope length | 0 |
| Evergreen forest cover 10-cell mean | 0 |
| - | $ \sqsubseteq , \ldots , \ldots ,$ |
| | $\begin{array}{cccccccccccccccccccccccccccccccccccc$ |

ater importance

34 36

0

0

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, $\max = \max \min, \min = \min \max$.



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.

| 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 |
| 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 max- imized. |
| 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 clas- sified 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 pres- ence | 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.

- 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.

- [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 pres-
- [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.

| T . 1 | |
|--|---------------------------|
| Lethe eurydice | |
| Species Distribution Model (SDM) assessment metrics and metadata | |
| Common name: Eyed Brown | good |
| Date: 01 Feb 2018 | TSS=0.91 |
| Code: letheury | 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.

| Wetland cover 10-cell mean | |
|---|---|
| Shrub cover 100-cell mean | |
| Mean temp of driest guarter | |
| Wetland cover 100-cell mean | ····· |
| Max temp of warmest month | · · · · · · · · · · · · · · · · · · · |
| Dist to fresh marsh | • |
| Growing degree days | |
| June precip | • |
| Annual mean temp | ••••••••••••••••••••••••••••••••••••••• |
| Water cover 100-cell mean | • |
| Topographic postion index 100-cell radius | · · · · · · · · · · · · · · · · · · · |
| Evergreen forest cover 100-cell mean | ••••••••••••••••••••••••••••••••••••••• |
| July precip | 0 |
| Min temp of coldest month | 0 |
| Wetland cover 1-cell mean | 0 |
| Precip of wettest quarter | 0 |
| Flowpath dist to water or wetland | 0 |
| Solar radiation summer solstice | 0 |
| Topographic moisture | |
| Roughness TU-cell circle | |
| Slope longth | 0 |
| Deciduous forest cover 1_cell mean | |
| | |
| Topographic postion index 10-cell radius | |
| Dist to mafic rock | |
| Canopy 100-cell mean | |
| Dist to river | |
| Roughness 1-cell square | · · · · · · · · · · · · · · · · · · · |
| Dist to acidic granitic rock | · · · · · · · · · · · · · · · · · · · |
| Impervious surface 100-cell mean | •••••• |
| Dist to silt/clay | · · · · · · · · · · · · · · · · · · · |
| Dist to lake or river | •••••• |
| Normalized dispersion of precip | · · · · · · · · · · · · · · · · · · · |
| Dist to acidic shale | • |
| Forest cover 10-cell mean | • |
| Water cover 10-cell mean | •••••• |
| May precip | ••••• |
| Slope | •••••• |
| Mean diurnal range | 0 |
| Beers aspect | 0 |
| Forest cover 100-cell mean | ····· |
| Porest cover 1–cell mean | <u> </u> |
| Deciduous forest cover 100-cell mean | ě. |
| Dist to lake | ě. |
| Open cover 10_cell mean | |
| Canopy 1-cell mean | |
| Deciduous forest cover 10-cell mean | · · · · · ō. · · · · · · · · · · · · |
| Dist to calc rock | · · · · o · · · · · · · · · · · · |
| Dist to moderately calc rock | · · · · o · · · · · · · · · · · · |
| Dist to acidic sedimentary rock | · · · · · · · · · · · · · · · · · · · |
| Roughness 100-cell circle | |
| Profile curvature | ••• |
| Dist to inland waters | · o · · · · · · · · · · · · · · · · · · |
| Topographic postion index 1-cell square | 0 |
| Canopy 10-cell mean | 0 |
| | L, , , , , , |
| | 9 10 11 12 13 lower \rightarrow greater |

importance

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.

14



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.

| Threshold | Value | EOs | Polvs | 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 max- imized. |
| 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 clas- sified 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 pres- ence | 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.

- 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 Eved Brown (Lethe eurydice). Created on 01 Feb 2018. Western Pennsylvania Conservancy, Pittsburgh, PA.

- [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.
- 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 | good |
| Date: 04 Dec 2017 | TSS=0.97 |
| 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.

| Topographic postion index 100-cell radius | ••••••••••••••••••••••••••••••••••••••• |
|---|---|
| Dist to woody wetland | ••••••••••••••••••••••••••••••••••••••• |
| Canopy 1–cell mean | ••••••••••••••••••••••••••••••••••••••• |
| Roughness 1–cell square | |
| Precip of driest month | ••••••••••••••••••••••••••••••••••••••• |
| Open cover 100-cell mean | • |
| Slope | ••••••••••••••••••••••••••••••••••••••• |
| Roughness 10-cell circle | · · · · · · · · · · · · · · · · · · · |
| Mean diurnal range | · · · · · · · · · · · · · · · · · · · |
| Dist to sand | • • • • • • • • • • • • • • • • • • • |
| Roughness 100-cell circle | · · · · · · · · · · · · · · · · · · · |
| Min temp of coldest month | • |
| Dist to loam | •••••• |
| Annual mean temp | · · · · · · · · · · · · · · · · · · · |
| Max temp of warmest month | |
| Open cover 10-cell mean | · · · · · · · · · · · · · · · · · · · |
| Dist to lake | |
| Dist to river | ō |
| Normalized dispersion of precip | |
| Dist to acidic shale | |
| Isothermality | |
| Dist to moderately calc rock | ō |
| July precip | |
| Wetland cover 10-cell mean | |
| Topographic moisture | |
| Solar radiation winter solstice | |
| Slope length | |
| Dist to calc rock | o |
| Wetland cover 100-cell mean | ō |
| Temp annual range | ····· |
| Forest cover 10-cell mean | 0 |
| Flowpath dist to water or wetland | ····· |
| Precip of coldest quarter | |
| Canopy 10-cell mean | ····· |
| Topographic postion index 10-cell radius | ····· |
| Canopy 100-cell mean | 0 |
| Topographic postion index 1-cell square | |
| Deciduous forest cover 10-cell mean | • |
| Dist to silt/clay | • |
| Total annual precip | • • • • • • • • • • • • • • • • • • • |
| Forest cover 100-cell mean | •••••• |
| Dist to fresh marsh | ····· |
| Deciduous forest cover 100-cell mean | •••••• |
| June precip | •••••• |
| Evergreen forest cover 100-cell mean | •••••• |
| Elevation | •••••• |
| Mean temp of wettest quarter | ····· |
| Impervious surface 100-cell mean | ····· |
| Dist to acidic sedimentary rock | |
| May precip | •••••• |
| Deciduous forest cover 1–cell mean | · · · · • • • • • • • • • • • • • • • • |
| Plan curvature | ···· 0····· |
| Slope curvature | ····0 |
| Precip of wettest month | 0 |
| Dist to lake or river | 0 |
| Beers aspect | o |
| | · · · · · · |
| | 18 20 22 lower → greater importance |

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.

| 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 max- imized. |
| 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 clas- sified 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 pres- ence | 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.

- 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.

- [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 pres-
- [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.

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.

| Dist to silt/clay | •••••••••••••••••••••••••••••••••••••• |
|---|--|
| Topographic postion index 100-cell radius | · · · · · · · · · · · · · · · · · · · |
| Dist to lake or river | · · · · · · · · · · · · · · · · · · · |
| Slope | • |
| Precip of driest quarter | · · · · · · · · · · · · · · · · · · · |
| Roughness 1-cell square | |
| Elevation | · · · · · · · · · · · · · · · · · · · |
| Dist to fresh marsh | <u> </u> |
| Dist to calt march | ă IIII |
| Dist to lake | ă I |
| Canony 1 coll moon | ő |
| Drasin of coldect quarter | |
| | |
| Slope length | |
| Flowpath dist to water or wetland | |
| Precip of wettest quarter | 0 |
| Dist to estuary | 0 |
| Dist to inland waters | •••••• |
| May precip | ••••• |
| Water cover 100–cell mean | •••••• |
| Dist to river | · · · · · · · · · · O · · · · · · · · · |
| July precip | · · · · · · · · · · · · · · · · · · · |
| Dist to loam | · · · · · · · · · O · · · · · · · · · · |
| Canopy 10-cell mean | · · · · · · · · · O · · · · · · · · · · |
| Evergreen forest cover 100-cell mean | 0 |
| Topographic postion index 10-cell radius | |
| Impervious surface 100-cell mean | |
| Forest cover 10-cell mean | · · · · · · · · · o · · · · · · · · · · |
| Dist to coastal waters | • |
| Dist to acidic shale | • |
| Mean diurnal range | · · · · · · · O · · · · · · · · · · · · |
| Water cover 10-cell mean | · · · · · · · · · · · · · · · · · · · |
| Open cover 100–cell mean | · · · · · · · · · · · · · · · · · · · |
| Temp annual range | · · · · · · · · · · · · · · · · · · · |
| Dist to acidic granitic rock | · · · · · o |
| Dist to woody wetland | · · · · · o |
| Wetland cover 10-cell mean | |
| Forest cover 100–cell mean | |
| Deciduous forest cover 10-cell mean | |
| Impervious surface 10-cell mean | · · · · · 0 · · · · · · · · · · · · · · |
| June precip | · · · · · o · · · · · · · · · · · · · · |
| Deciduous forest cover 100-cell mean | · · · · · o · · · · · · · · · · · · · · · · · · |
| Isothermality | 0 |
| Roughness 10-cell circle | |
| Dist to matic rock | 0 |
| Solar radiation equinox | |
| Dist to acidic sedimentary rock | · · · · o |
| Growing degree days | |
| Mean temp of wettest quarter | · · · · ō |
| Total annual precip | |
| Dist to ocean | |
| Topographic postion index 1_cell square | |
| Mean temp of driest quarter | |
| Min temp of coldest month | |
| Dist to moderately calc rock | |
| Shrub cover 100-cell mean | 0 |
| Mean temp of warmest quarter | · ō |
| Wetland cover 100-cell mean | 0 |
| Dist to sand | , <u>,</u> |
| Annual mean temp | |
| Normalized dispersion of precip | <u>.</u> |
| Topographic moisture | <u> </u> |
| | Ľ.,, |
| | 22 24 26 20 |
| | 22 24 20 20 30 |
| | importance |
| | |

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.

| 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 max- imized. |
| 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 clas- sified 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 pres- ence | 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.

- Maryland Natural Heritage Program, Maryland Department of Natural Resources, Wildlife and Heritage Service
- New Jersev Department of Environmental Protection, Division of Fish and Wildlife, New Jersev 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.

- 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.
- 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 pres-
- ence/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.
- 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 | good |
| Date: 02 Dec 2017 | TSS=0.86 |
| 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.

| Topographic postion index 10-cell radius | | | | · · · · · • • • • • • • • • • • • • • • |
|---|---------------------------------------|-------------------------|--------------|---|
| Topographic postion index 100-cell radius | | | •••• | |
| Dist to coastal waters | | | •••• | |
| Dist to fresh marsh | | | 0.0 | |
| Dist to silt/clay | | | 0 | |
| Dist to estuary | | • • • • • • • | 6 | |
| Dist to loam | | · · · · · · · • | | |
| Dist to salt marsh | | • • • • • • • | | |
| Dist to inland waters | | | | |
| Dist to acidic granitic rock | | | | |
| Dist to mafic rock | | | | |
| Dist to moderately calc rock | | | | |
| Wetland cover 100-cell mean | | | | |
| Precip of wettest month | | 0 | | |
| Dist to ultramafic rock | | | | |
| Wetland cover 10-cell mean | | | | |
| Dist to calc rock | | | | |
| Impervious surface 100-cell mean | | | | |
| Evergreen forest cover 100-cell mean | | | | |
| Topographic postion index 1-cell square | | | | |
| Dist to acidic shale | | | | |
| Temp annual range | | | | |
| Precip of coldest guarter | | | | |
| Roughness 100-cell circle | | | | |
| Forest cover 100-cell mean | | | | |
| Shrub cover 100-cell mean | | | | |
| Flowpath dist to water or wetland | | | | |
| Deciduous forest cover 100-cell mean | | | | |
| May precip | | | | |
| Total annual precip | | • • • • • | | |
| Mean temp of wettest quarter | | | | |
| Dist to stream | | 0.00 | | |
| Dist to river | | 0.0 | | |
| Roughness 10–cell circle | | 0 | | |
| Elevation | | 0 | | |
| Dist to woody wetland | | 0 | | |
| Dist to lake or river | | 0 | | |
| Mean temp of driest quarter | ••••• | 3 | | |
| Precip of driest quarter | · · · · · · · · · · · · · · · · · · · | | | |
| Water cover 100-cell mean | · · · · · · · · · | 9 | | |
| Dist to acidic sedimentary rock | · · · · · c | | | |
| Growing degree days | · · · · C |) | | |
| Normalized dispersion of precip | • • • • • | | | |
| Mean temp of coldest quarter | | | | |
| July precip | 00 | | | |
| Deciduous forest cover 10-cell mean | | | | |
| Isothermality | | | | |
| Dist to sand | | | | |
| lopographic moisture | | | | |
| Siope length | 0.0 | | | |
| Appuel mean temp | | | | |
| Moon diurnal range | ~ | | | |
| Diet to pond | õ | | | |
| Moon tomp of warmost quarter | ~ | | | |
| Dist to lake | Ň | | | |
| Canony 1_cell mean | Ň | | | |
| Canopy 10_cell mean | | | | |
| lune precip | õ | | | |
| Shrub cover 10–cell mean | o | | | |
| | Ľ, | | | |
| | 10 Iov | 14 wer \rightarrow | 18 greate | ¦ ər |

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.

| 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 max- imized. |
| 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 clas- sified 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 pres- ence | 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.

- 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.

- [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.
- 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 pres-
- ence/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.
- 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 | - |
|-------------------------------------|---------------------------|
| Common name: Broad-winged Skipper | good |
| 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 |
|------------------|-------|------|------|
| Name | Wiean | 50 | 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.

| Wetland cover 10-cell mean | |
|---|---|
| Wetland cover 100–cell mean | |
| Topographic moisture | •••••• |
| Temp seasonality | •••••••••••••••••••••••••••••••••••••• |
| Wetland cover 1-cell mean | ····· |
| Slope length | ····· |
| Dist to silt/clay | ····· |
| Slope | ······ |
| Solar radiation winter solstice | 0. |
| Roughness 1–cell square | 0. |
| Dist to woody wetland | ···· • • • • • • • • • • • • • • • • • |
| Flowpath dist to water or wetland | ····••••••••••••••••••••••••••••••••• |
| Annual mean temp | · · · · · · · · · · · · · · · · · · · |
| June precip | 0 |
| Roughness 10–cell circle | 0 |
| Min temp of coldest month | ····· |
| Growing degree days | ····o··· |
| Topographic postion index 100-cell radius | 0 |
| Mean temp of driest quarter | ····· |
| Max temp of warmest month | · · · · · · · · · · · · · · · · · · · |
| Open cover 100–cell mean | 0 |
| July precip | 0 |
| Elevation | 0 |
| Deciduous forest cover 100-cell mean | 0 |
| Precip of warmest quarter | 0 |
| Forest cover 10-cell mean | 0 |
| Forest cover 100-cell mean | |
| Normalized dispersion of precip | 0 |
| Shruh sover 100 cell mass | 0 |
| Evergreen forest cover 100-cell mean | |
| Deciduous forest cover 10–cell mean | |
| Dist to sand | |
| Mean temp of wettest guarter | |
| Dist to acidic sedimentary rock | 0 |
| Canopy 100-cell mean | ····· |
| Dist to fresh marsh | 0 |
| Topographic postion index 10-cell radius | 0 |
| Isothermality | 0 |
| Forest cover 1–cell mean | ····· |
| Precip of driest quarter | 0 |
| May precip | ···· |
| Deciduous forest cover 1–cell mean | · · · · 0 · · · · · · · · · · · · |
| Mean diurnal range | · · · • • · · · · · · · · · · · · · · |
| Dist to moderately calc rock | |
| Dist to loam | 0 |
| Water cover 100-cell mean | 0 |
| Iotal annual precip | 0 |
| Topographic position index 1-cell square | |
| Canopy TU-Cell mean | |
| Siope cuivalure | |
| Deers aspect | ă |
| Dist to lake | _ |
| | Ľ |
| | $\begin{array}{ccc} 4 & 6 & 8 & 12 \\ & \text{lower} \rightarrow \text{greate} \end{array}$ |

importance

16

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.

0



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.

| 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 max- imized. |
| 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 clas- sified 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 pres- ence | 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.

- 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.

- [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.
- 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 pres-
- ence/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.
- 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 | fair |
| Date: 01 Feb 2018 | TSS=0.78 |
| Code: polimyst | 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.

| Roughness 1–cell square | ······ |
|---|---|
| Slope | ····· |
| Mean temp of wettest quarter | 0 |
| Evergreen forest cover 100-cell mean | 0, |
| Elevation | •••••••••••••••••••••••••••••••••••••• |
| Dist to woody wetland | 0 |
| Canopy 1–cell mean | O |
| Mean temp of warmest quarter | 0 |
| Canopy 10–cell mean | •••••••••••••••••••••••••••••••••••••• |
| Roughness 10–cell circle | •••••••••••••••••••••••••••••••••••••• |
| Normalized dispersion of precip | 0 · · · · · · · · · · · · · · · · · · · |
| Dist to sand | O |
| Annual mean temp | •••••••••••••••••••••••••••••••••••••• |
| Growing degree days | ····· |
| Precip of driest month | 0 |
| Dist to fresh marsh | 0 |
| Dist to mafic rock | O |
| Min temp of coldest month | O |
| Forest cover 100-cell mean | ····· |
| mpervious surface 100-cell mean | •••••••••••••••••••••••••••••••••••••• |
| July precip | •••••••••••••••••••••••••••••••••••••• |
| May precip | •••••••••••••••••••••••••••••••••••••• |
| Dist to acidic shale | •••••••••••••••••••••••••••••••••••••• |
| Topographic postion index 10-cell radius | •••••••••••••••••••••••••••••••••••••• |
| Dist to silt/clav | · · · · · · · · · · · · · · · · · · · |
| Temp annual range | 0 |
| Precip of coldest quarter | 0 |
| Solar radiation winter solstice | 0 |
| Total annual precip | 0 |
| Mean temp of driest quarter | · · · · · · · · · · · · · · · · · · · |
| Dist to acidic granitic rock | · · · · · · · · · · · · · · · · · · · |
| Open cover 100-cell mean | ····· |
| Canopy 100-cell mean | o |
| Topographic postion index 100-cell radius | 0 |
| Mean diurnal range | 0 · · · · · · · · · · · · · · · · · · · |
| Deciduous forest cover 10-cell mean | •••••••••••••••••••••••••••••••••••••• |
| Precip of wettest guarter | •••••••••••••••••••••••••••••••••••••• |
| Dist to river | •••••••••••••••••••••••••••••••••••••• |
| Wetland cover 100-cell mean | 0 |
| Water cover 100-cell mean | •••••••••••••••••••••••••••••••••••••• |
| Roughness 100-cell circle | •••••••••••••••••••••••••••••••••••••• |
| Dist to calc rock | •••••••••••••••••••••••••••••••••••••• |
| Dist to moderately calc rock | · · · · · · · · · · · · · · · · · · · |
| Topographic postion index 1-cell square | •••••••••••••••••••••••••••••••••••••• |
| Shrub cover 100-cell mean | 0 |
| Deciduous forest cover 100-cell mean | · · · · · · · · · · · · · · · · · · · |
| June precip | 0 |
| Dist to lake | 0 |
| Dist to lake or river | •••••••••••••••••••••••••••••••••••••• |
| Dist to loam | 0 |
| Isothermality | · · · · · O · · · · · · · · · · · · · · |
| Slope curvature | · · · · · O · · · · · · · · · · · · · · |
| Flowpath dist to water or wetland | · · · · · O · · · · · · · · · · · · · · |
| Dist to inland waters | ····· |
| Profile curvature | ····· |
| Forest cover 10-cell mean | 0 |
| Dist to pond | 0 |
| | $ \sqsubseteq \dots $ |
| | 18 22 26 lower → greater importance |

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, $\max = \max \min, \min = \min \max$.

0



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.

| 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 max- imized. |
| 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 clas- sified 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 pres- ence | 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.

- 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.

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