EcoSHEDS Northeast Brook Trout Occupancy Model

Jeffrey D. Walker, Benjamin Letcher

v2.0.0 (July 8, 2022)

Contents

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.7     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

## Linking to GEOS 3.9.1, GDAL 3.2.3, PROJ 7.2.1; sf_use_s2() is TRUE

## Loading required package: lattice

##
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':
##
##     lift

##
## Attaching package: 'janitor'

## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

Chapter 1

Introduction

The northeast brook trout occupancy model predicts the occupancy probability of eastern brook trout based on estimated stream temperatures across a high resolution catchment delineation for the northeast U.S. (Maine to Virginia). This model is developed and maintained as part of the EcoSHEDS project.

1.1 Model Summary

The occupancy model is a logistic linear mixed effects model for predicting occupancy probability based on estimated stream temperature. The primary model input is the estimated mean July stream temperature generated by the EcoSHEDS Northeast Stream Temperature Model and averaged over all years from 1980 to present. The model also includes a varying intercept (random effect) based on the HUC8[1] basin containing each catchment to account for regional variations in occupancy not explained by stream temperature. Like the EcoSHEDS stream temperature model, the occupancy model domain is based on the EcoSHEDS Northeast Catchment Delineation.

The occupancy model was calibrated using presence/absence data compiled from multiple state agencies across the region.

The model was validating by splitting the observed presence/absence data into independent training and testing subsets, calibrating the model to the training data, and then comparing the model performance between the two subsets. The validation confirmed that the model structure is not susceptible to over-fitting. Following validation, the model was re-calibrated using all available observation data in order to maximize the amount of information used to generate the final predictions. The following table summarizes the model performance for each dataset.

[1] 8-digit Hydrologic Unit Code (HUC)

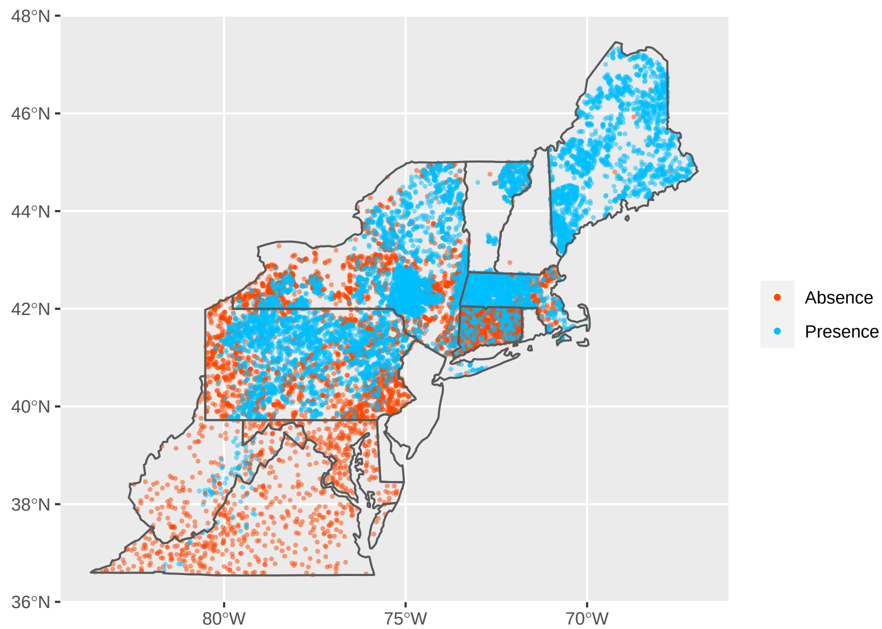


Figure 1.1: Map of Presence/Absence by Catchment

Metric	Calibration	Validation	Full (Prediction)
Accuracy	0.832	0.809	0.832
Sensitivity	0.897	0.844	0.894
Specificity	0.712	0.748	0.717
Pos Pred Value	0.852	0.857	0.853
Neg Pred Value	0.790	0.727	0.786
Precision	0.852	0.857	0.853
Recall	0.897	0.844	0.894
F1	0.874	0.850	0.873
Prevalence	0.649	0.642	0.648
Detection Rate	0.582	0.542	0.579
Detection Prevalence	0.684	0.632	0.679
Balanced Accuracy	0.805	0.796	0.806

Predicted occupancy probabilities were generated for each catchment under both historical conditions as well as for a series of simple climate change scenarios (e.g., +2, +4, +6 degC air temperature). Additional metrics were also computed to evaluate the vulnerability of brook trout occupancy within each catchment to climate change (e.g., the maximum air temperature increase that would result in 30, 50, or 70% occupancy). Model predictions were only generated for catchments with cumulative drainage areas less than 200 km^2 (i.e., lower order streams) due to the complexity of predicting stream temperatures in larger rivers.

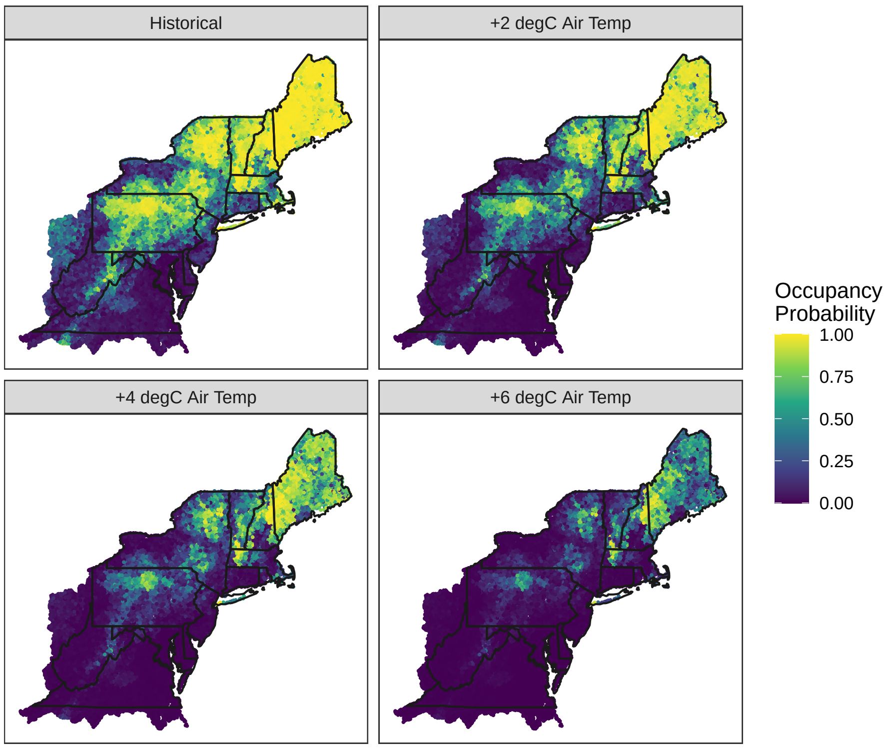


Figure 1.2: Predicted Occupancy Probabilities under Historical and Future Climate Change Scenarios

The predicted occupancy probabilities along with stream temperature predictions and various catchment characteristics can be found in the Northeast Interactive Catchment Explorer (ICE). Data files containing the predictions are also available for download (see Downloads).

The model is periodically updated when either new observation data are available, or following each update of the northeast stream temperature model. Currently, updates are being performed annually, usually in the fall.

1.2 Documentation

The documentation is divided into the following sections:

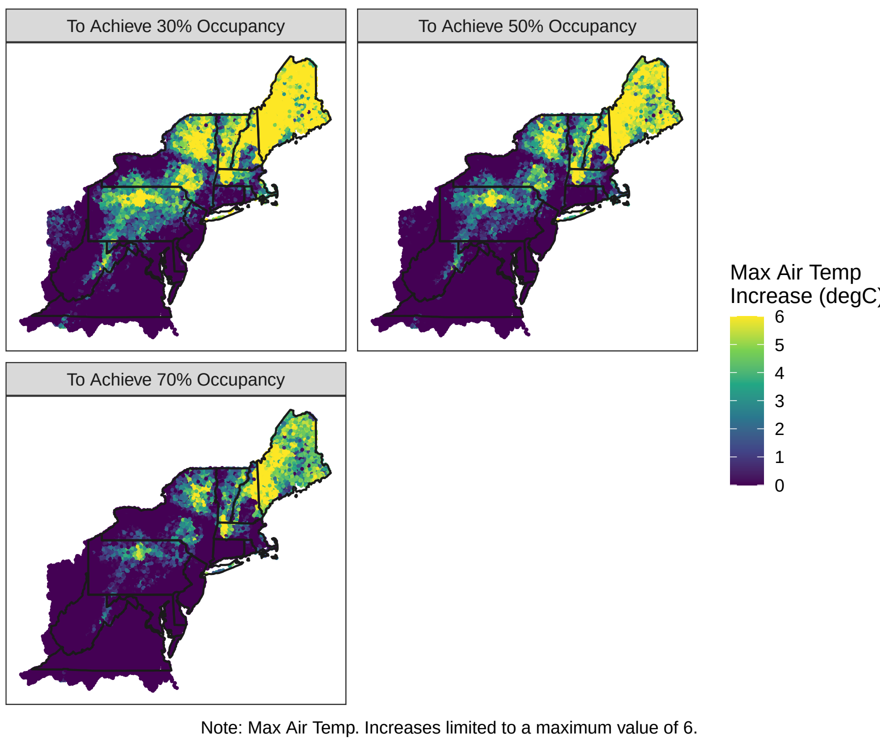


Figure 1.3: Predicted Max. Air Temperature Increases to Achieve Varying Occupancy Probabilities

1. Introduction provides an overview the model and its documentation
2. Data Sources describes the data sources and data processing steps used to generate the model inputs and observed data
3. Model Development describes the model structure and implementation
4. Calibration and Validation summarizes the model calibration and validation
5. Predictions describes the model prediction scenarios and metrics
6. Downloads provides links to download the model predictions
7. History contains a list of changes to the model over time

1.3 Source Code

The source code for the model itself and this documentation is available in the Github repository walkerjeffd/sheds-bto-model. Each version of the model will be included under the list of Releases.

Chapter 2

Data Sources

2.1 Catchment Delineation

The model domain was based on the EcoSHEDS Northeast Catchment Delineation, which is also used for the EcoSHEDS Northeast Stream Temperature Model. This delineation was created in 2014-2015 to provide a high resolution catchment delineation across the northeast that was spatially consistent over the entire region. The EcoSHEDS delineation was created prior to the completion of the USGS NHDPlus High Resolution dataset. Unfortunately, the two delineations are not directly relatable.

Figure 2.1 shows the extent of this delineation, which covers the northeast and mid-Atlantic portions of the eastern U.S. from Maine to Virginia. The delineation includes approximately 450,000 individual catchments, with an mean surface area of 1.6 km^2.

2.2 HUC8 Basins

The 8-digit Hydrologic Unit Code (HUC8) basins were used to account for regional variations in brook trout occupancy that was not explained by stream temperature data (see [Model Structure]). Figure 2.2 shows the boundary of each HUC8 basin.

2.3 Observation Data

Observed presence/absence data of eastern brook trout was gathered from multiple state agencies.

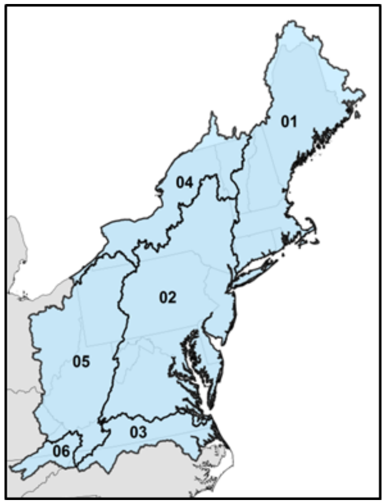


Figure 2.1: Hydrologic Regions of EcoSHEDS Northeast Catchment Delineation

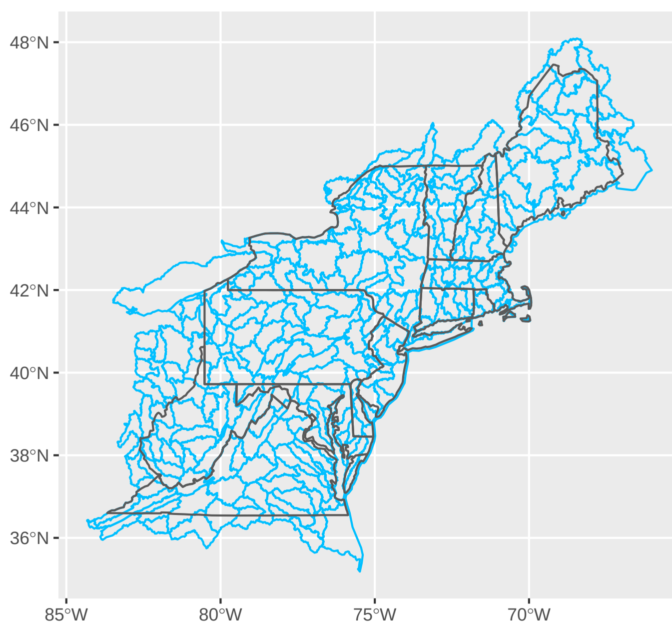


Figure 2.2: HUC8 Basins

The raw presence/absence observations were aggregated by catchment for input to the model. If there were multiple observations in a single catchment (either at different locations or at different times), then the catchment was assigned to the `presence` category if there was at least presence among those observations. If all observations were absences, then the catchment was assigned to `absence`.

The follow map, figures, and table provide a summary of the observed catchment presence/absence by state.

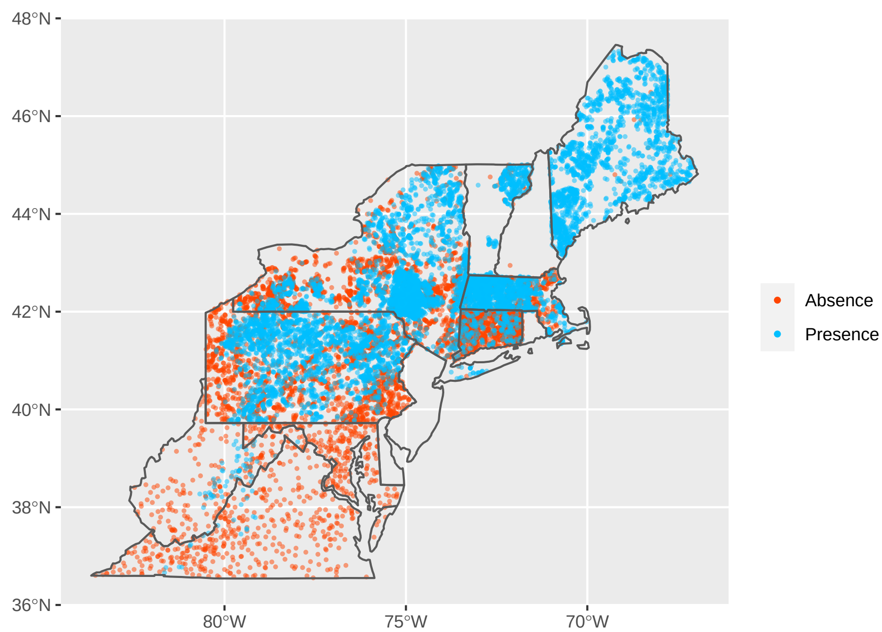


Figure 2.3: Map of Presence/Absence by Catchment

State	Presence	Absence	Total	% Presence
ME	1,874	10	1,884	99.5%
NH	8	5	13	61.5%
VT	283	45	328	86.3%
MA	1,585	294	1,879	84.4%
RI	3	4	7	42.9%
CT	206	1,066	1,272	16.2%
NY	2,722	1,633	4,355	62.5%
NJ	2	8	10	20.0%
PA	2,050	1,747	3,797	54.0%
MD	0	226	226	0.0%
WV	55	178	233	23.6%
VA	18	404	422	4.3%
Total	8,806	5,620	14,426	61.0%

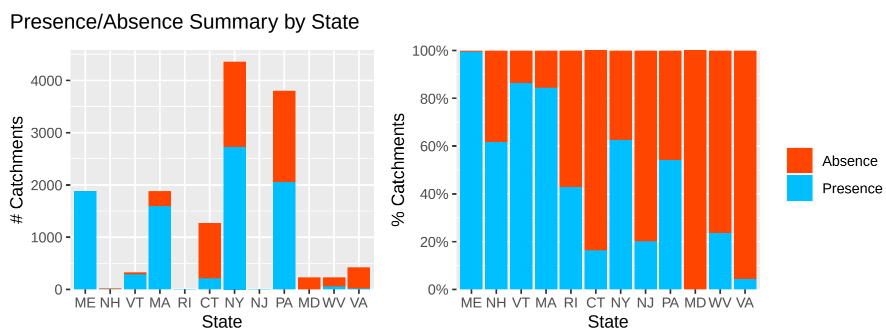


Figure 2.4: Presence/Absence Catchment Totals by State

2.4 Stream Temperature

The primary input to the brook trout occupancy model is the estimated mean July stream temperature of each catchment. These estimates were generated using the EcoSHEDS Northeast Stream Temperature Model. The estimated value for each catchment represents the long-term mean July stream temperature computed over all years since 1980. Estimates are only generated for catchments with total cumulative drainage areas less than 200 km^2 due to the great complexity, anthropogenic impacts, and non-linearities associated with temperature dynamics in larger rivers.

The stream temperature model is based on a linear mixed effects framework that accounts for spatial and temporal correlations using a hierachical Bayesian structure. The primary input variables include air temperature, precipitation, land use (forest, agriculture, high development), impounded drainage area, and total drainage area. See the stream temperature model documentation for more information.

Figure 2.5 shows the estimated mean July stream temperature for each catchment.

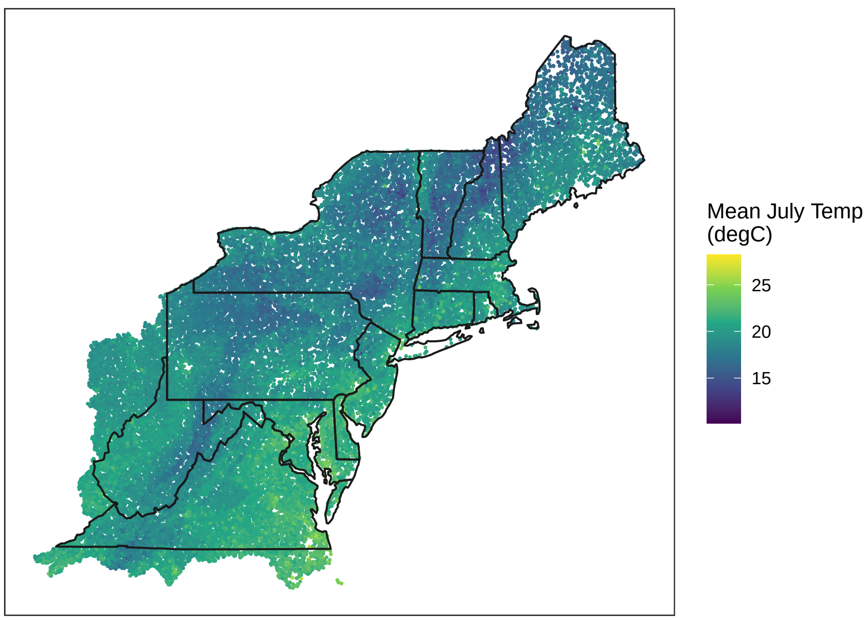


Figure 2.5: Estimated Mean July Stream Temperature by Catchment

Chapter 3

Model Development

3.1 Model Definition

The occupancy model uses a logistic linear mixed effects model framework for estimating the probability of occupancy in each catchment (Bolker et al., 2009; Zuur et al., 2009).

The model includes one fixed effect representing the estimated mean July stream temperature (see Stream Temperature), and a random effect intercept that varies by HUC8. The random effect is included to account for spatial variations in observed occupancy that are not explained by the mean July stream temperature.

The model is fit using the `glmer()` function of the `lme4` R package (Bates et al., 2015) using the following formula and parameters.

```
lme4::glmer(
  presence ~ mean_jul_temp + (1 | huc8),
  family = binomial(link = "logit"),
  data = model_data,
  control = glmerControl(optimizer = "bobyqa")
)
```

3.2 Model History

The original development of this model was based on a similar brook trout occupancy model that was developed using a Bayesian hierarchical framework to evaluate associations between various catchment and riparian characteristics as well as climate inputs and the occupancy probability in catchments within the state of Connecticut (Kanno et al., 2015).

Prior to version 2.0.0, the EcoSHEDS northeast brook trout occupancy model included a number of independent variables (i.e., covariates) representing land use (forest, agriculture, high intensity development), climate (summer precipitation), and drainage area. The earlier model versions also included the mean July stream temperature as estimated by the EcoSHEDS northeast stream temperature model. However, because the stream temperature model depended on a number of the same independent variables, the estimated effects in the occupancy model were often counter-intuitive due to cross-correlations between some covariates and the estimated stream temperature.

Therefore, beginning with version 2.0.0, the brook trout occupancy model uses only the mean July stream temperature as the sole fixed effect. This change resulted in a small decrease in model accuracy, but provides more intuitive results and can be more easily applied for evaluating alternative climate or land use change scenarios.

Chapter 4

Calibration and Validation

4.1 Dataset Split

To calibrate and validate the model, the observed dataset of presence/absence by catchment (see Observation Data) was randomly split using 80% of the catchments for calibration and 20% for validation.

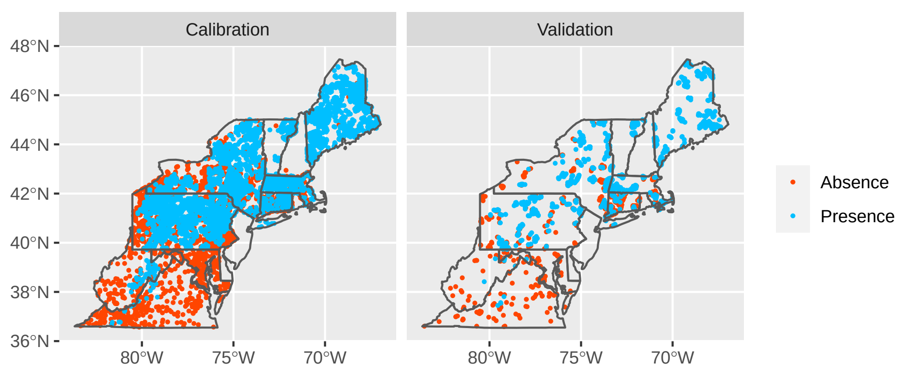


Figure 4.1: Calibration and Validation Splits

Partition	Presence	Absense	Total	% Presence
Calibration	6,690	3,612	10,302	64.9%
Validation	1,822	1,017	2,839	64.2%
Total	8,512	4,629	13,141	64.8%

4.2 Calibration

The following output summarizes the fitted model using the calibration subset.

```
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: presence ~ mean_jul_temp + (1 | huc8)
   Data: model_data
Control: glmerControl(optimizer = "bobyqa")

     AIC      BIC   logLik deviance df.resid
  8592.8   8614.7  -4293.4   8586.8    10934

Scaled residuals:
    Min      1Q  Median      3Q     Max
-32.628  -0.371   0.135   0.427  11.133

Random effects:
 Groups Name        Variance Std.Dev.
 huc8   (Intercept) 4.999    2.236
Number of obs: 10937, groups:  huc8, 202

Fixed effects:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   19.54035    0.68083   28.70   <2e-16 ***
mean_jul_temp -1.03582    0.03444  -30.07   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
            (Intr)
mean_jl_tmp -0.965
```

The estimated fixed effect for mean July temp (`mean_jul_temp`) was -1.04. Because the estimated value is negative, the occupancy probability is higher at lower stream temperatures. Figure @**??**fig:calib-fixed) contains a marginal effects plot showing the predicted probability over varying mean July stream temperatures (excluding random effects).

The random effect intercept varies by HUC8 basin. Basins with higher values tend to have higher occupancy probabilities for a given mean July stream temperature. Some HUC8 basins do not have an estimated value because there was no observations with the calibration dataset.

The model accuracy and performance is summarized by a series of metrics computed from the confusion matrix, which contains the total number of true

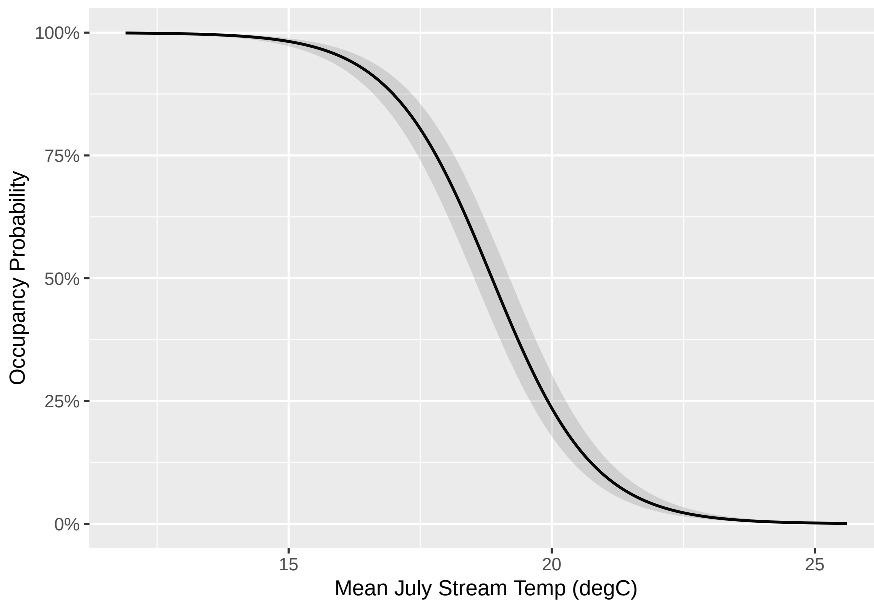


Figure 4.2: Marginal Effects Plot for Mean July Stream Temperature.

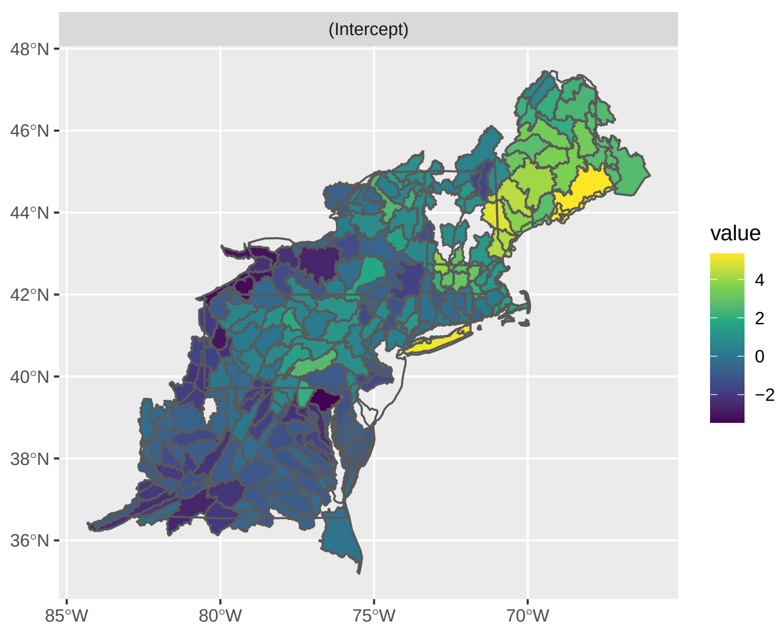


Figure 4.3: Random Effect Intercept by HUC8 Basin

positives, true negatives, false positives, and false negatives. In the 2x2 table at the top of the following output, the columns (`Reference`) refer to the observed condition in each catchment (`1` = presence, `0` = absence), while the rows (`Prediction`) refer to the predicted condition. The predicted probabilities were converted to presence/absence using a 50% cutoff. The remaining output provides a series of performance metrics computed from the confusion matrix using the `confusionMatrix()` of the `caret` package (Kuhn, 2022). See the help page for that function, as well as this Wikipedia article, for definitions of each metric.

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 2733  728
         1 1107 6369

               Accuracy : 0.8322
                 95% CI : (0.8251, 0.8392)
    No Information Rate : 0.6489
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6233

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.8974
            Specificity : 0.7117
         Pos Pred Value : 0.8519
         Neg Pred Value : 0.7897
             Prevalence : 0.6489
         Detection Rate : 0.5823
   Detection Prevalence : 0.6836
      Balanced Accuracy : 0.8046

       'Positive' Class : 1
```

4.3 Validation

Using the calibrated model, predicted probabilities were computed using the indendent validation dataset.

The confusion matrix for the validation dataset indicates slightly lower accuracy (0.83 vs 0.81), but overall comparable performance. These results suggest that the model does not suffer from overfitting.

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0  590  221
         1  199 1194

               Accuracy : 0.8094
                 95% CI : (0.7924, 0.8256)
    No Information Rate : 0.642
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.588

 Mcnemar's Test P-Value : 0.3055

            Sensitivity : 0.8438
            Specificity : 0.7478
         Pos Pred Value : 0.8571
         Neg Pred Value : 0.7275
             Prevalence : 0.6420
         Detection Rate : 0.5417
   Detection Prevalence : 0.6320
      Balanced Accuracy : 0.7958

       'Positive' Class : 1
```

The following table compares the performance metrics between the two subsets.

Metric	Calibration	Validation
Accuracy	0.832	0.809
Sensitivity	0.897	0.844
Specificity	0.712	0.748
Pos Pred Value	0.852	0.857
Neg Pred Value	0.790	0.727
Precision	0.852	0.857
Recall	0.897	0.844
F1	0.874	0.850
Prevalence	0.649	0.642
Detection Rate	0.582	0.542
Detection Prevalence	0.684	0.632
Balanced Accuracy	0.805	0.796

Lastly, Receiver Operator Characteristic (ROC) curves and Area Under the Curve (AUC) values also shows comparable performance between the calibration and validation subsets.

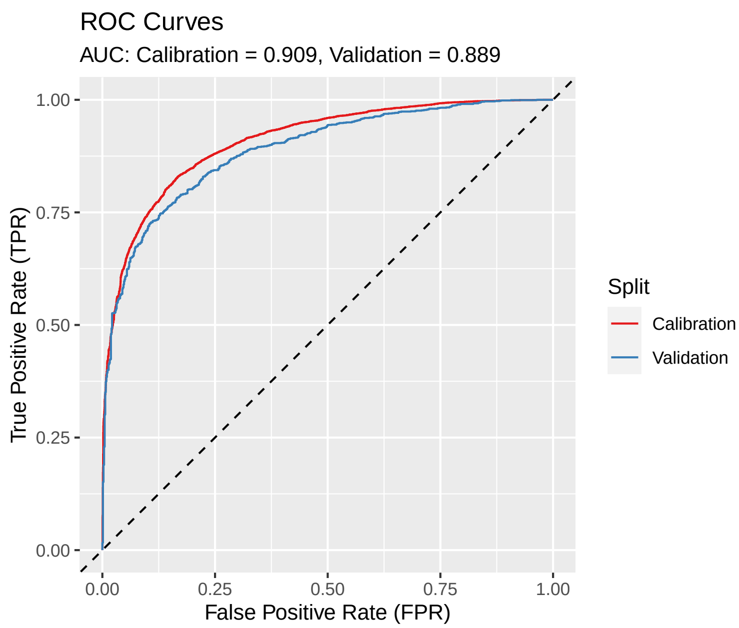


Figure 4.4: ROC Curves for Calibration and Validation

Chapter 5

Predictions

5.1 Full Calibration

To utilize the maximum amount of information available from the observation data, the model was re-calibrated using the full observation dataset. The following output shows that the estimated effects were comparable to those from the initial calibration based on 80% split of the full dataset (see Calibration).

```
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: presence ~ mean_jul_temp + (1 | huc8)
   Data: predict_inp
Control: glmerControl(optimizer = "bobyqa")

     AIC      BIC   logLik deviance df.resid
 10286.4  10308.9  -5140.2  10280.4    13138

Scaled residuals:
     Min       1Q   Median       3Q      Max
-29.1602  -0.3739   0.1199   0.4331  10.7152

Random effects:
 Groups Name        Variance Std.Dev.
 huc8   (Intercept) 5.148    2.269
Number of obs: 13141, groups:  huc8, 209

Fixed effects:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   18.43567    0.61595   29.93   <2e-16 ***
```

```
mean_jul_temp -0.97728    0.03084  -31.69   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
            (Intr)
mean_jl_tmp -0.958
```

The estimated fixed effect for mean July temp (`mean_jul_temp`) was -0.98. Figure @ref(fig:Calibration and Validation) contains a marginal effects plot showing the predicted probability over varying mean July stream temperatures (excluding random effects) showing higher predicted probabilities at lower stream temperatures.

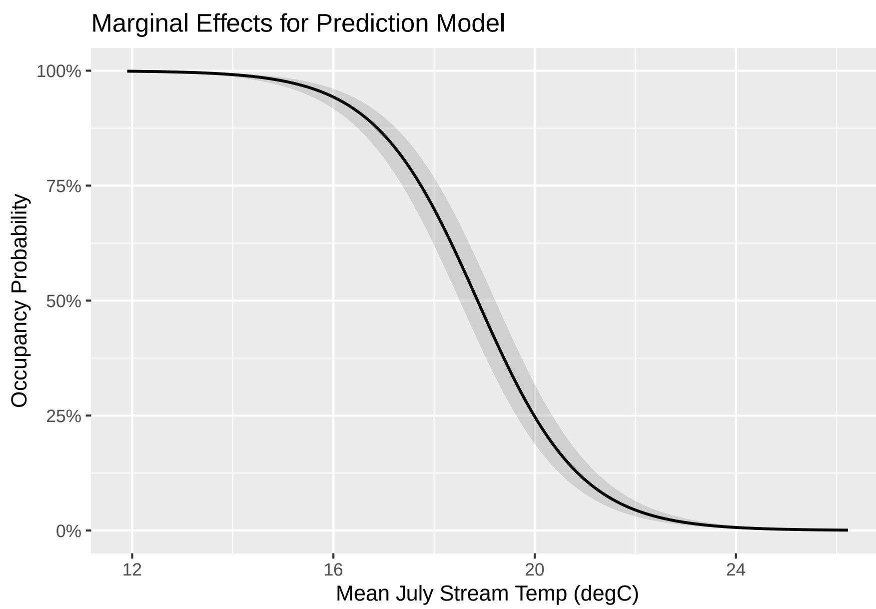


Figure 5.1: Marginal Effects Plot for Mean July Stream Temperature (Prediction Model)

The HUC8 random effect intercepts were also relatively similar to those from the initial calibration model.

The following output summarizes the prediction model's accuracy and performance (see Calibration for explanation of this output).

```
Confusion Matrix and Statistics
```

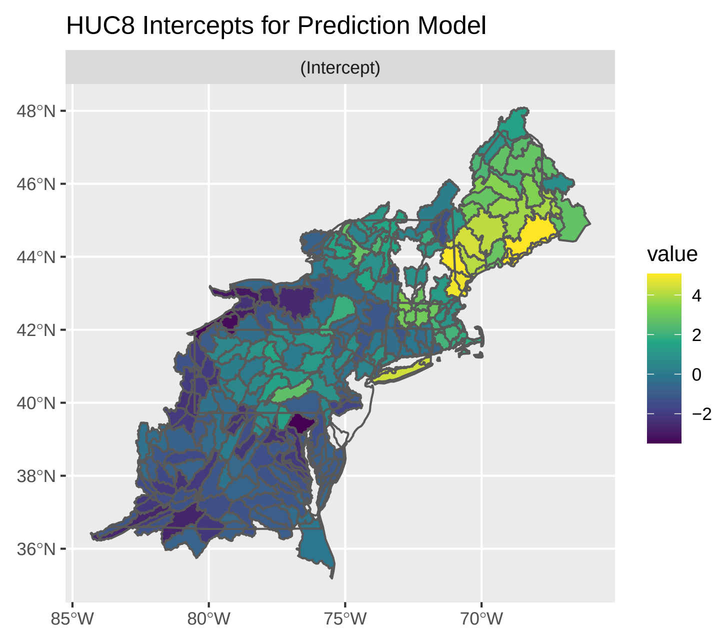


Figure 5.2: Random Effect Intercept by HUC8 Basin (Prediction Model)

```
          Reference
Prediction    0    1
         0 3319  902
         1 1310 7610

               Accuracy : 0.8317
                 95% CI : (0.8252, 0.838)
    No Information Rate : 0.6477
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6236

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.8940
            Specificity : 0.7170
         Pos Pred Value : 0.8531
         Neg Pred Value : 0.7863
             Prevalence : 0.6477
         Detection Rate : 0.5791
   Detection Prevalence : 0.6788
      Balanced Accuracy : 0.8055
```

```
'Positive' Class : 1
```

The following table compares the performance of the prediction model using all available data to that of the calibration and validation subsets (see Calibration and Validation). Overall, the full prediction model performs very similarly to the initial calibration model.

Metric	Calibration	Validation	Full (Prediction)
Accuracy	0.832	0.809	0.832
Sensitivity	0.897	0.844	0.894
Specificity	0.712	0.748	0.717
Pos Pred Value	0.852	0.857	0.853
Neg Pred Value	0.790	0.727	0.786
Precision	0.852	0.857	0.853
Recall	0.897	0.844	0.894
F1	0.874	0.850	0.873
Prevalence	0.649	0.642	0.648
Detection Rate	0.582	0.542	0.579
Detection Prevalence	0.684	0.632	0.679
Balanced Accuracy	0.805	0.796	0.806

5.2 Prediction Metrics

Using the prediction model that was fitted to the full observation dataset, a series of occupancy metrics were computed for each catchment over the region (excluding those with cumulative drainage areas > 200 km^2).

The prediction metrics include:

1. Occupancy probability under historical conditions as well as air temperature increases of +2, +4, and +6 degC, which represents a series of simple climate change scenarios.

2. The maximum increase in air temperature such that the predicted occupancy would be 30, 50, or 70%. These metrics indicate how resistant each catchment is to future climate change. Catchments with higher values can tolerate a larger increase in air temperature and still achieve each target occupancy probability.

A dataset containing the predicted values for these metrics can be downloaded in the Downloads section.

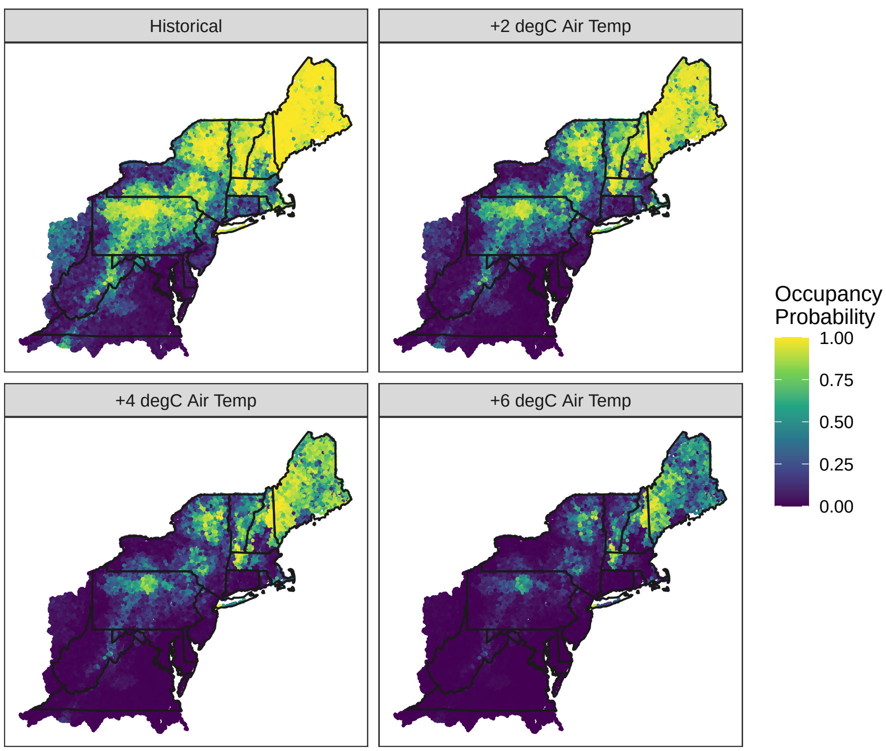


Figure 5.3: Predicted Occupancy Probabilities under Historical and Future Climate Change Scenarios

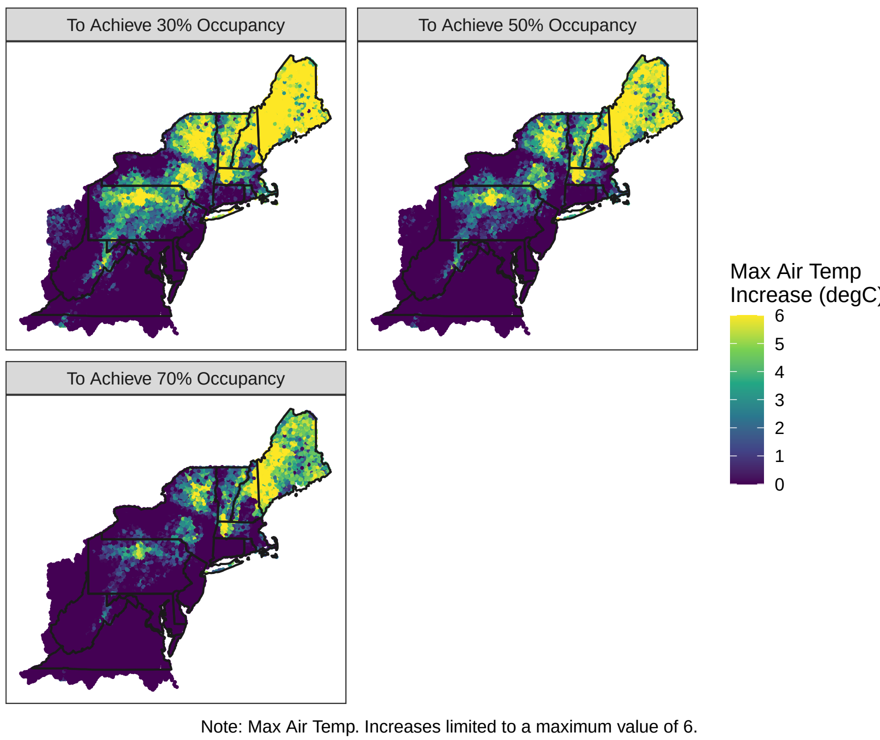


Figure 5.4: Predicted Max. Air Temperature Increases to Achieve Varying Occupancy Probabilities

Chapter 6

Downloads

6.1 Model Predictions

The EcoSHEDS Northeast Brook Trout Occupancy Model predictions can be downloaded as CSV file from the following link.

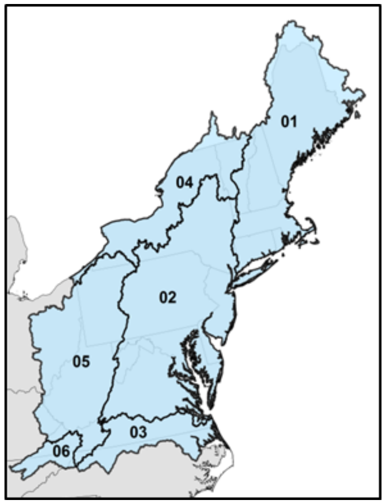
Brook Trout Occupancy Predictions 2.0.0 (csv)

This file contains the following headers:

CSV Column	Description
featureid	Catchment ID
huc8	HUC8 ID
mean_jul_temp	Mean July Stream Temp. (degC)
occ_current	Occupancy Probability (Historical)
occ_air_2	Occupancy Probability (+2.0 degC Air Temp)
occ_air_4	Occupancy Probability (+4.0 degC Air Temp)
occ_air_6	Occupancy Probability (+6.0 degC Air Temp)
max_air_occ30	Max Air Temp Increase (degC) for 30% Occupancy Probability
max_air_occ50	Max Air Temp Increase (degC) for 50% Occupancy Probability
max_air_occ70	Max Air Temp Increase (degC) for 70% Occupancy Probability

6.2 Catchment Delineation

The EcoSHEDS Northeast Catchment Delineation can be downloaded from the following links, which contain shapefiles grouped by HUC2 region. The model predictions dataset can be joined to these shapefiles using the common `featureid` column, which serves as the unique catchment ID.



- Region 01 Catchments (zip)
- Region 02 Catchments (zip)
- Region 03 Catchments (zip)
- Region 04 Catchments (zip)
- Region 05 Catchments (zip)
- Region 06 Catchments (zip)

The documentation for the catchment delineation is also available:

EcoSHEDS Northeast Catchment Delineation (docx)

Chapter 7

History

7.1 Versioning

The model uses semantic versioning of the form: `vX.Y.Z`

- `X` is the **major** version, which will be incremented when there is a new set of results and major changes to the model structure, code, or datasets.
- `Y` is the **minor** version, which will be incremented when there is a new set of results and minor changes in the model code or datasets (but no change in model structure).
- `Z` is the **patch** version, which will be incremented only when there is a change to the documentation or code that *does not* yield different results.

7.2 Change Log

- v2.0.0 | Jul 8, 2022 Removed all fixed effects except for mean July stream temperature due to cross-correlation between stream temperatures and other covariates (e.g., forest cover). Changed random effect levels from HUC10 to HUC8.
- v1.4.0 | Sep 8, 2021
  Re-calibrate model with updated stream temperature model results through 2020 (v1.3.0), removed July stream temperature scenarios from predictions (these have been replaced by the air temperature scenarios)
- v1.3.0 | Jul 13, 2020
  Re-run calibration with updated stream temperature model results through 2019 (v1.2.0)
- v1.2.2 | Jan 17, 2020
  Rename prediction metrics to be more consistent

- v1.2.1 | Jan 16, 2020
  Added predictions for air temperature scenarios (+2, +4, +6 degC), remove 0.5 degC increments from July stream temperature scenarios, revise method for estimating max temperature increases to achieve occupancy thresholds using linear interpolation
- v1.2.0 | Dec 3, 2019
  Re-run calibration with updated stream temperature model results (v1.1)
- v1.1.1 | Mar 26, 2019
  Update documentation, add [Download] section containing links to model predictions, catchment delineation, and covariates.
- v1.1.0 | Mar 25, 2019
  Add observation data from MA DFW. Recalibrate model using v1.0 of stream temperature model.
- v1.0.0 | Oct 25, 2018
  Recalibrate model using v1.0 of stream temperature model.
- v0.9.0 | Aug 16, 2018
  Preliminary release of the new model framework and documentation.
- **Previous Versions (prior to 2018)**
  Previous versions of the brook trout occupancy model can be found here. That website is now deprecated, but will remain available for future reference. Beginning with v1.0.0 of the new framework and codebase, all model changes and results will be tracked and made available.

Bibliography

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., and White, J.-S. S. (2009). Generalized linear mixed models: A practical guide for ecology and evolution. *Trends in Ecology & Evolution*, 24(3):127–135.

Kanno, Y., Letcher, B. H., Rosner, A. L., O'Neil, K. P., and Nislow, K. H. (2015). Environmental Factors Affecting Brook Trout Occurrence in Headwater Stream Segments. *Transactions of the American Fisheries Society*, 144(2):373–382.

Kuhn, M. (2022). *caret: Classification and Regression Training*. R package version 6.0-92.

Zuur, A. F., Ieno, E. N., Walker, N., Saveliev, A. A., and Smith, G. M. (2009). *Mixed Effects Models and Extensions in Ecology with R*. Statistics for Biology and Health. Springer New York, New York, NY.