Fish in Trees

An Introduction to Boosted Regression Tree Analysis

Charles Bangley
Dalhousie University

Disclaimer

Ecologist ⟷ Statistician

Photo credit: Cecilia Krahforst

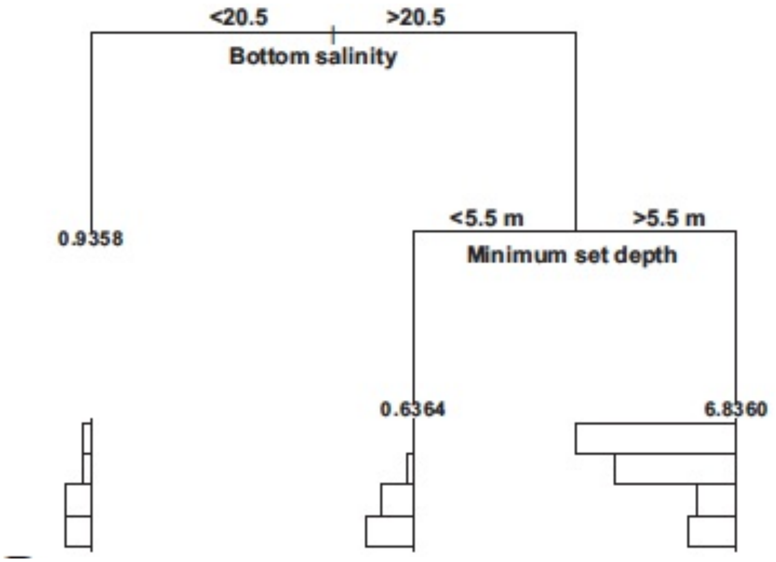
Overview

- BRT Basics
- Using BRTs with acoustic telemetry data
- BRT Analysis in R using *gbm.auto*
- Case Study: Dusky sharks off the US Mid-Atlantic region

BRT Basics

Regression Tree Analysis

- Splits data into branches (or nodes) at cut points in explanatory variables
  - Splits aim to minimize variance in the resulting branches
  - Usually between high and low values of the response variable
- Results provide ranges of explanatory variables associated with high response variable values

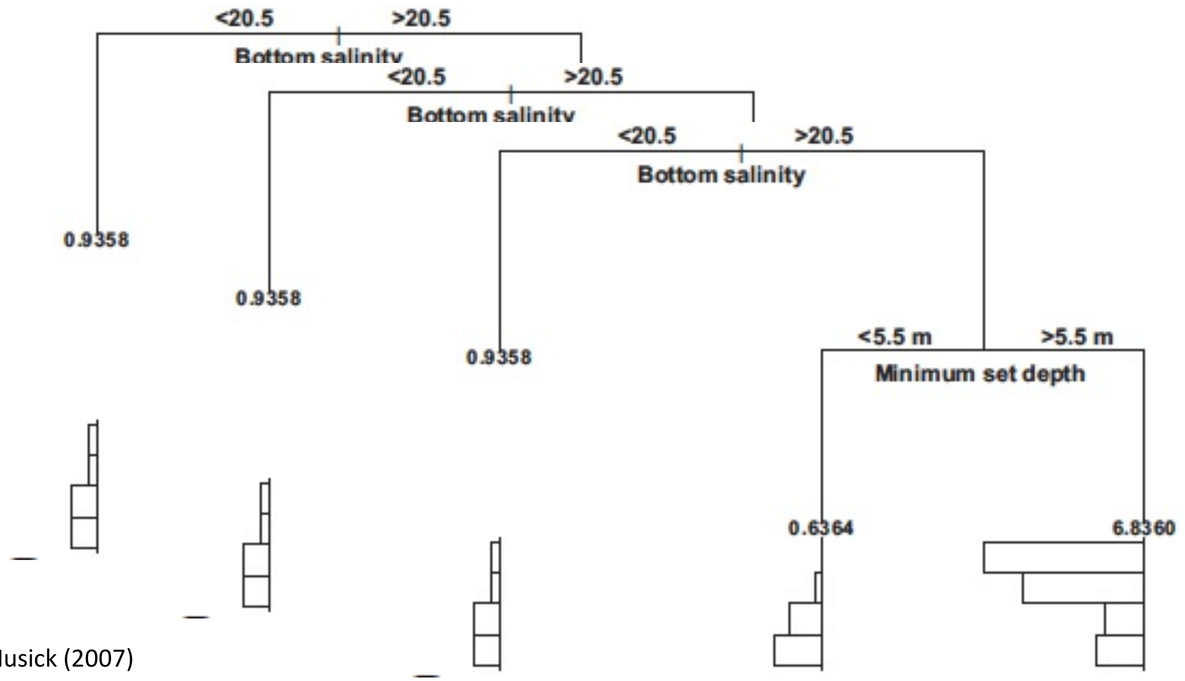


Grubbs and Musick (2007)

BRT Basics

Boosting

- Reduces variance in individual regression tree analysis
- Boosting – repeats analysis over many iterations
    - Machine learning allows each successive tree to “learn from” the last
    - Process repeated over many iterations until deviance between trees is minimized – typically at least 1000



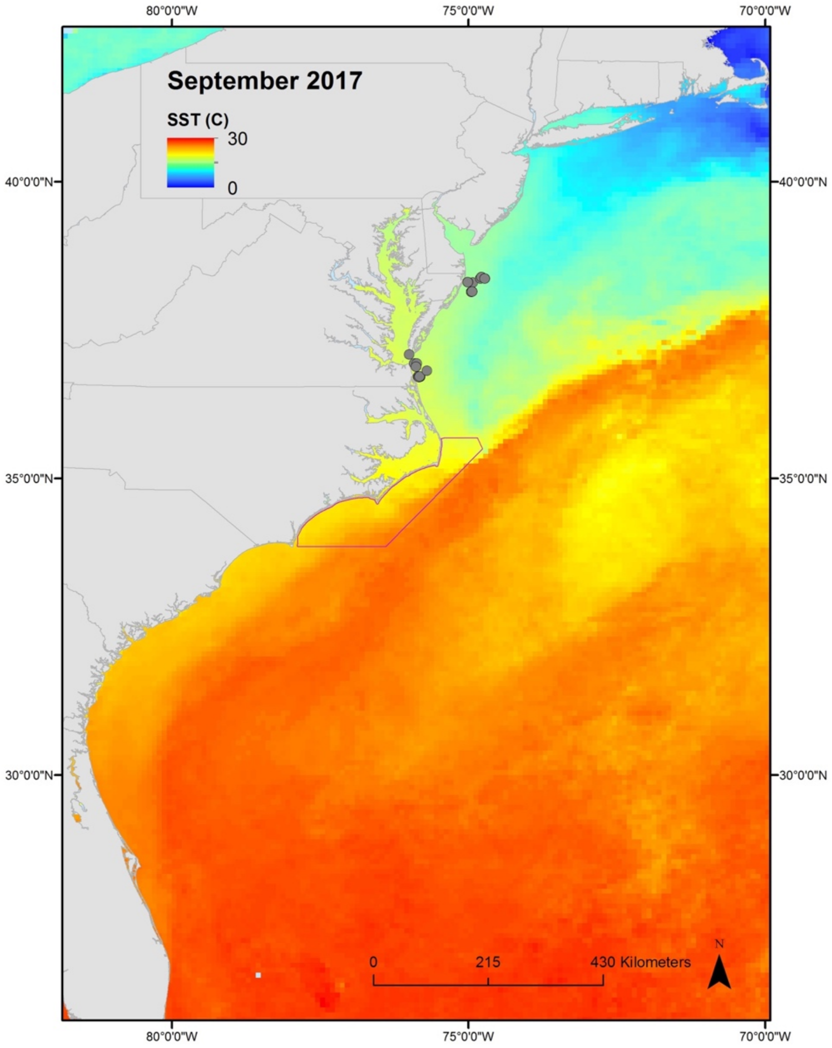
Grubbs and Musick (2007)

BRT Basics

Response Variables
- Presence – binary/probability
- Abundance – Gaussian
- Can model each separately or both
    - Abundance corrected for presence probability

Explanatory Variables
- Environmental data
- Presence/abundance of other species

Pairwise Interactions – optional but useful
- Linear modeling between pairs of variables
- Interaction strength = residual variance of pairwise linear models



BRT Basics

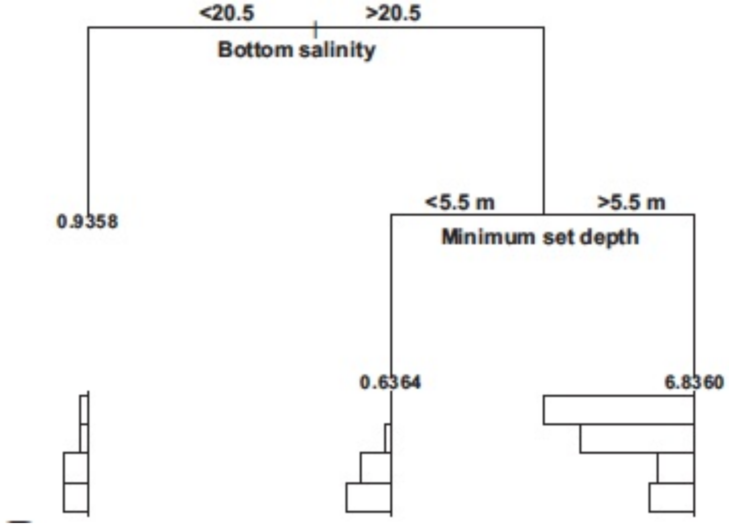
Model Parameters

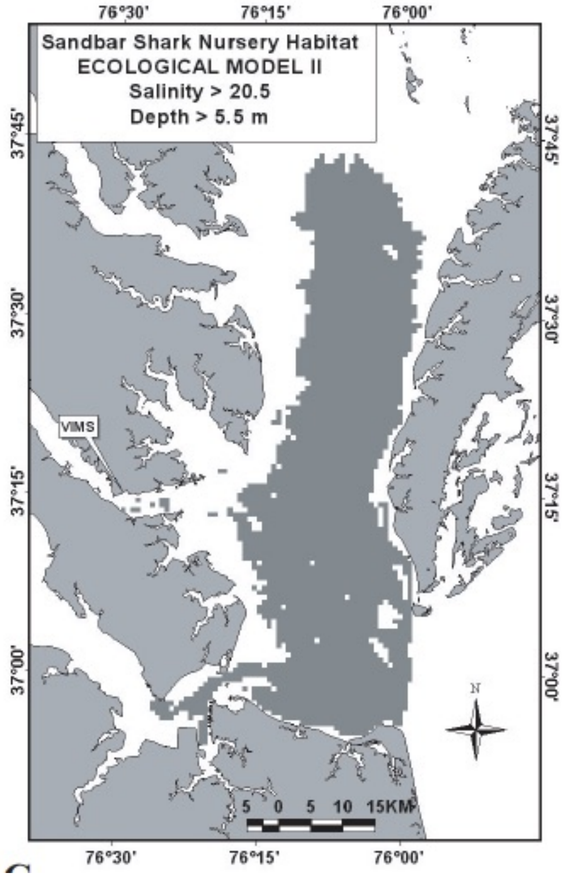
- Tree Complexity (*tc*) – number of nodes at each split
  - Typically either two or number of explanatory variables (within reason)
- Learning Rate (*lr*) – contribution of each tree to reducing deviance of the next
- Bag Fraction (*bf*) – proportion of data randomly selected and used to cross-validate the rest (the training data)
  - Usually 0.4-0.7
  - Data are randomly selected for each tree iteration

BRT Basics

Spatial Analysis – mapped BRT results are very intuitive (and look really cool)

- Applies model results to grid of environmental data





Grubbs and Musick (2007)

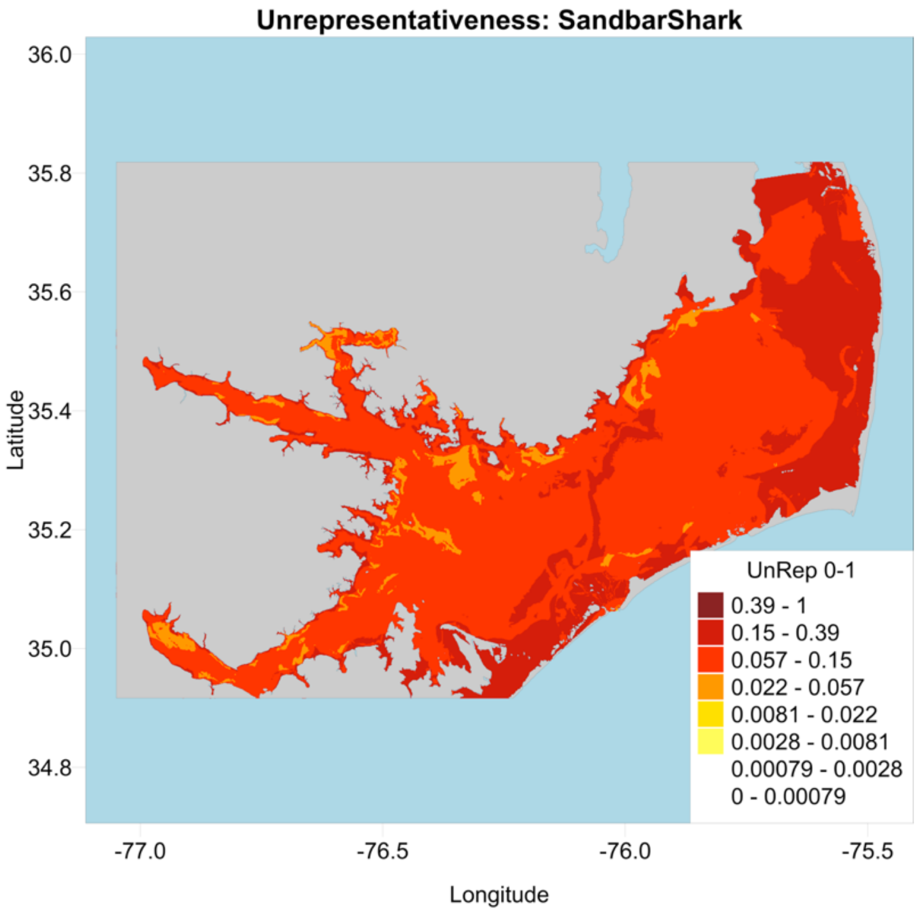
BRT Basics

Model diagnostics – no *p*-value, so how do you (and/or reviewers) know how well it performed?

- Cross-validation score (CV score) – the greater, the better
  - 0.6 or greater considered "good"
- Area Under Curve (AUC) – the greater, the better
- Mean Deviance – the lesser, the better
- Cross-validated AUC vs. training data AUC
  - Used to measure model overfitting – not significant if values are similar

BRT Basics

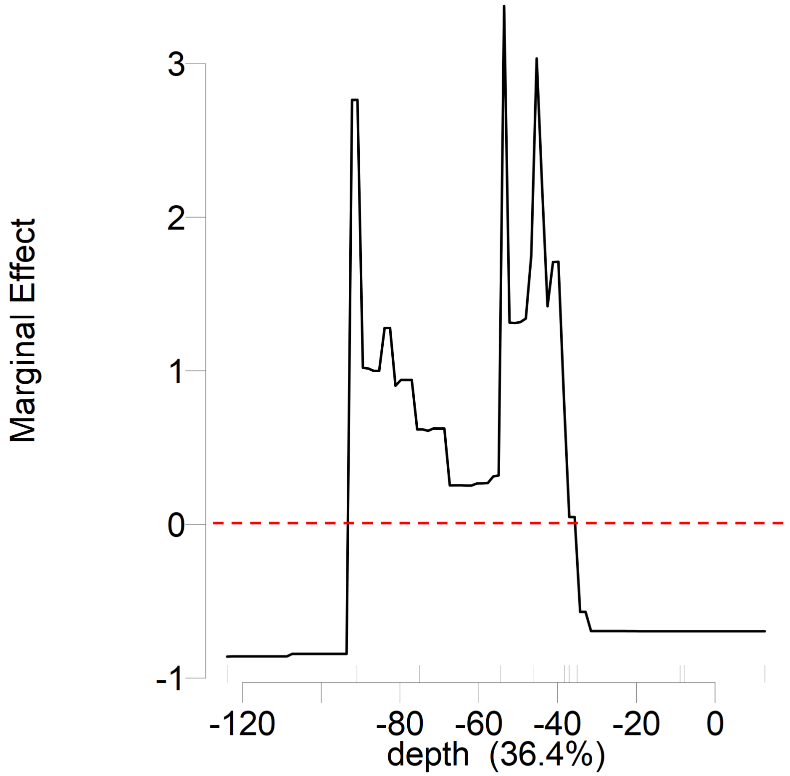
More ways to measure model performance:
- Unrepresentativeness Maps (for spatial analysis)
  - Shows how well data used in the model overlap with the actual range of explanatory variables in environmental grid
- Measures of accuracy
  - True/false positives and negatives
  - Follow-up studies with new data



BRT Basics

Marginal Effect Plots – Provide information on:

- Relationship with each explanatory variable
    - Line height shows positive/negative effect
    - Y-axis is relative probability for binary models, deviation from mean for Gaussian
- Relative importance/influence of each explanatory variable
    - Measured as % of tree splits attributed to that variable



BRT Analysis and Acoustic Telemetry

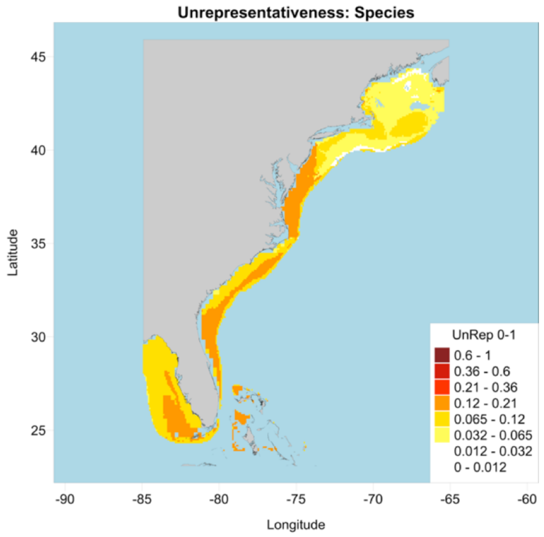
Variables from acoustic telemetry data

- Response Variables
  - Presence – detection at a given receiver over a certain time
    - Presence/absence per day, hour, etc.
  - Abundance – usually requires a fair number of detections
    - Number of individuals detected over certain time
    - Amount of time spent at receiver
- Explanatory Variables
  - Environmental data
    - Large-scale – extracted from satellite/model data
    - Smaller-scale – recorded by instruments at/near receivers
  - Detections of other animals

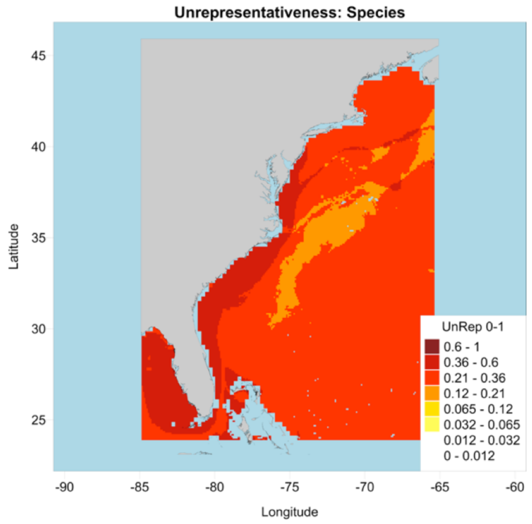
BRT Analysis and Acoustic Telemetry

Practical considerations
- Receiver coverage
  - Limit modeling to general area of receiver coverage and similar environments
- Temporal coverage
  - Limit modeling to time frames (months, seasons) during which tagged animals were actually detected
- Summarizing data
  - Summarize to an appropriate time scale
  - Match to temporal resolution of explanatory variables
    - For satellite data, usually daily

Clipped to receiver coverage



Not clipped



BRT Analysis in R

Commonly-used packages

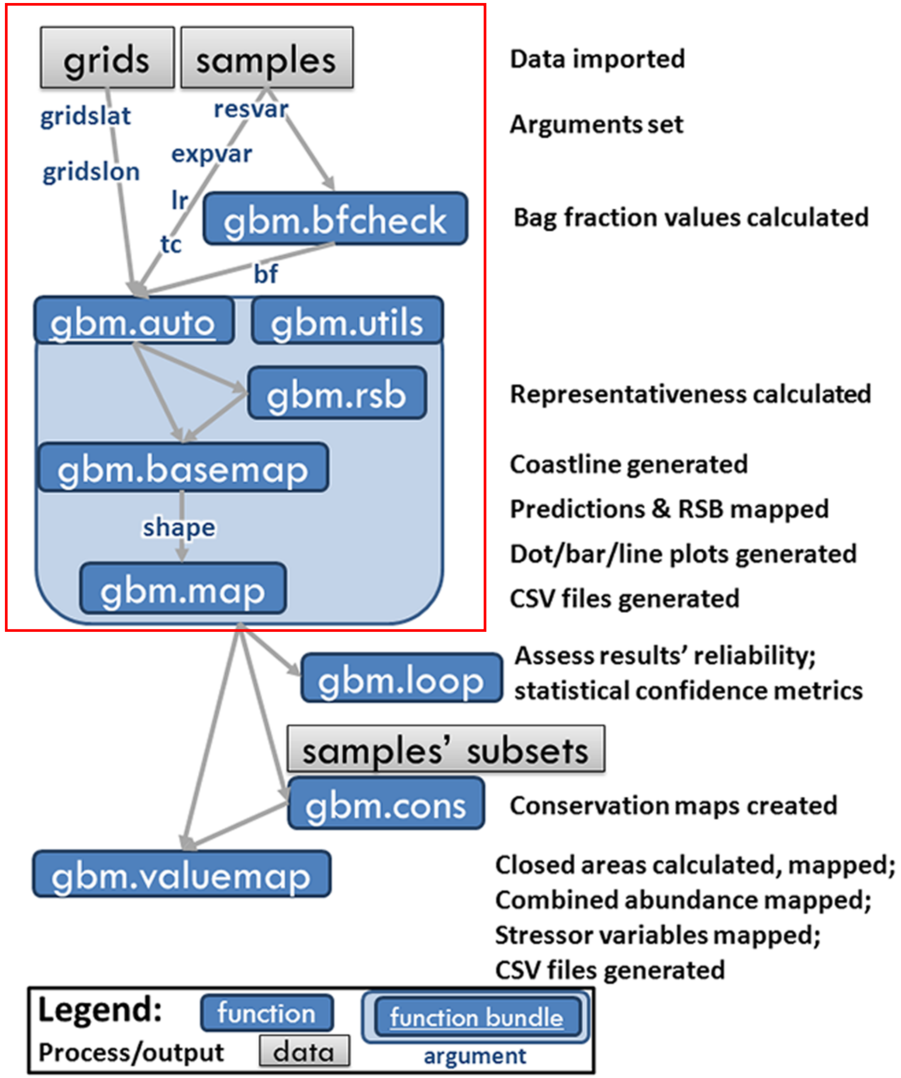
- *Dismo* (Elith and Leathwick 2011)
  - Functions for running BRTs, mapping, diagnostic metrics, etc.
  - Each step is a separate function
- *Gbm.auto* (Dedman et al. 2017)
  - Automates model running, mapping, marginal effect plots, diagnostics...
  - Very handy, but can sometimes be tough to find sources of errors
  - Includes functions for running steps individually
  - Runs a binary BRT for presence/absence, and a Gaussian BRT for abundance (using only data where species was present)

BRT Analysis in R

BRT analysis using *gbm.auto* and the "gbm.auto" function

Highly recommended reading: Dedman, S., R. Officer, M. Clarke, D. G. Reid, and D. Brophy. 2017. Gbm.auto: a software tool to simplify spatial modeling and Marine Protected Area planning. PLOS One 12: e0188955.

- Especially the supplementary material
- Available at simondedman.com



Dedman et al. (2017)

BRT Analysis in R

Prepping data for a run through the "gbm.auto" function

- What you'll need:
  - Presence/abundance data including explanatory variables – "samples"
  - Gridded data of latitude, longitude, and explanatory variables – "grids"
    - Only needed if mapping results
  - Can be helpful to already have a shapefile of your study area, but one can be generated during the "gbm.auto" function

BRT Analysis in R

What's in the code?

- The basics

Explanatory variable columns

Response variable column

Grids data frame

Samples data frame

```
gbm.auto(expvar=c(5,7,8,9), resvar=10, grids=NovShelfGrid,
samples=DuskyNov17, lr=0.005, bf=0.6, tc=4,
gridslat=3, gridslon=4, map=TRUE, shape=BaseMap,
gaus=FALSE, varint=TRUE, simp=FALSE, max.trees=10000)
```

Learning rate

Bag fraction

Tree complexity

Latitude and longitude columns

- Must match in both grids and samples data

BRT Analysis in R

What's in the code?

- Advanced options – there are more available than shown here

Do you want to map results?

- If not, grids, gridslat, gridslon can be removed

Specify shoreline shapefile

```
gbm.auto(expvar=c(5,7,8,9), resvar=10, grids=NovShelfGrid,
samples=DuskyNov17, lr=0.005, bf=0.6, tc=4,
gridslat=3, gridslon=4, map=TRUE, shape=BaseMap,
gaus=FALSE, varint=TRUE, simp=FALSE, max.trees=10000)
```

Do you want to run the Gaussian model?

- If not, only the binary presence/absence model is run

Tests whether model performs better with some variables removed

Calculate interactions between explanatory variables

Set maximum number of tree iterations to run

BRT Analysis in R

Choosing and testing model parameters

- Can run multiple combinations of lr, bf, tc
  - e.g. `lr=c(0.005,0.001)` will run separate models for lr = 0.005 and 0.001
  - “gbm.auto” automatically selects best-performing combination of parameters based on CV score
- General advice for choosing starting parameters
  - “gbm.bfcheck” function returns minimum bf values for binary and Gaussian models that will allow them to run
  - Lower lr values generally provide less variability in model results, but can also take much longer to run
  - High tc values in cases where many explanatory variables are used may be impractical

BRT Analysis in R

Interpreting reports of model results

Variables

Zero-inflation test

Models run
- Parameters
- Model metrics

Best model metrics

Variables dropped/kept
- If simp=TRUE

Variable influence

Two strongest interactions

Explanatory Variables	Response Variables	Zero Inflated?	Bin_BRT.tc4.lr0.0075.bf0.6	Bin_BRT.tc4.lr0.005.bf0.6	Bin_BRT.tc4.lr0.0025.bf0.6	Best Binary BRT	Bin_BRT_simp predictors dropped	Bin_BRT_simp predictors kept	Simplified Binary BRT stats	Best Binary BRT variables	Relative Influence (Bin)	Biggest Interactions (Bin)
Depth	Species	TRUE	trees: 1000	trees: 1200	trees: 2400	Model combo: Bin_BRT.tc4.lr0.0075.bf0.6	simp turned off	simp turned off	simp turned off	SST	47.99867482	SST and Sal. Size: 49.57
Chla			Training Data Correlation: 0.737218111416522	Training Data Correlation: 0.705479212023333	Training Data Correlation: 0.705977125037348	Model CV score: 0.737218111416522				Chla	35.90218358	SST and Chla. Size: 8.56
Sal			CV Mean Deviance: 0.0818320371585656	CV Mean Deviance: 0.0794615826954027	CV Mean Deviance: 0.0822913091736796	Training data AUC score: 0.9719				Depth	10.70610764	
SST			CV Deviance SE: 0.00599084392027356	CV Deviance SE: 0.00614505740669313	CV Deviance SE: 0.00486113069207683	CV AUC score: 0.7929				Sal	5.393033962	
			CV Mean Correlation: 0.284953925487672	CV Mean Correlation: 0.29083894202811	CV Mean Correlation: 0.273500088714658	CV AUC se: 0.0490395713231218						
			CV Correlation SE: 0.0793090352045057	CV Correlation SE: 0.0740988037184598	CV Correlation SE: 0.0692233078297898							

Case Study: Northwest Atlantic Dusky Sharks

Bangley, C. W., T. H. Curtis, D. H. Secor, R. J. Latour, and M. B. Ogburn. 2020. Identifying important juvenile dusky shark habitat in the Northwest Atlantic Ocean using acoustic telemetry and spatial modeling. Marine and Coastal Fisheries 12: 348-363

Danielle Hall

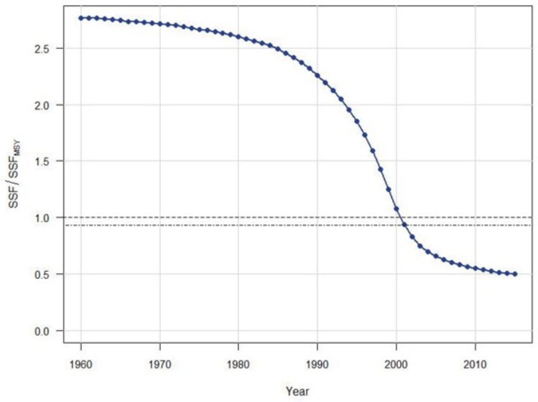
Case Study: Northwest Atlantic Dusky Sharks

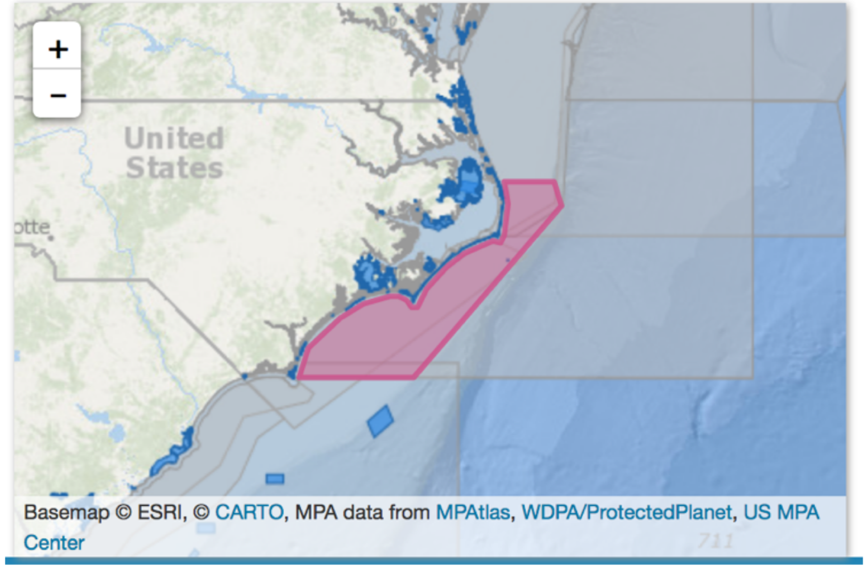
Dusky Shark (*Carcharhinus obscurus*)

VULNERABLE — Worldwide

ENDANGERED — Northwest Atlantic

NMFS – overfished with overfishing occurring (SEDAR 2016)





Case Study: Northwest Atlantic Dusky Sharks

Objectives

Develop spatial models of Dusky Shark presence probability based on telemetry detections and environmental data.

Account for seasonal/migratory changes in distribution.

Use spatial models to predict distribution during periods of low/no tag detection.

Case Study: Northwest Atlantic Dusky Sharks

Danielle Hall

Methods - Telemetry

23 Dusky Sharks

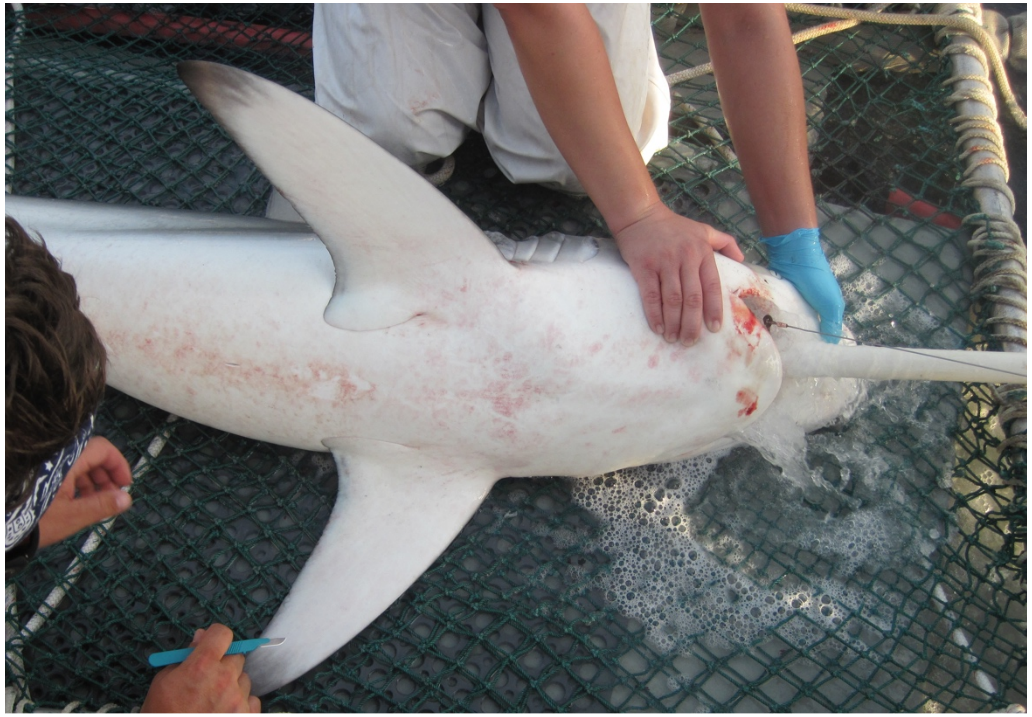
- 5 by VIMS off VA – Sept 2016, Aug 2017
- 3 by Tobey Curtis/OCEARCH off NY Bight – Sept 2016
- 15 off Ocean City, MD – Sept 2017
- 1067-2200 mm total length

The Atlantic Cooperative Telemetry Network

FACT Fish Move.

Case Study: Northwest Atlantic Dusky Sharks

Methods - Mapping, Modeling, and Mapping



VIMS

Matrix of daily presence/absence of tagged dusky sharks at each receiver

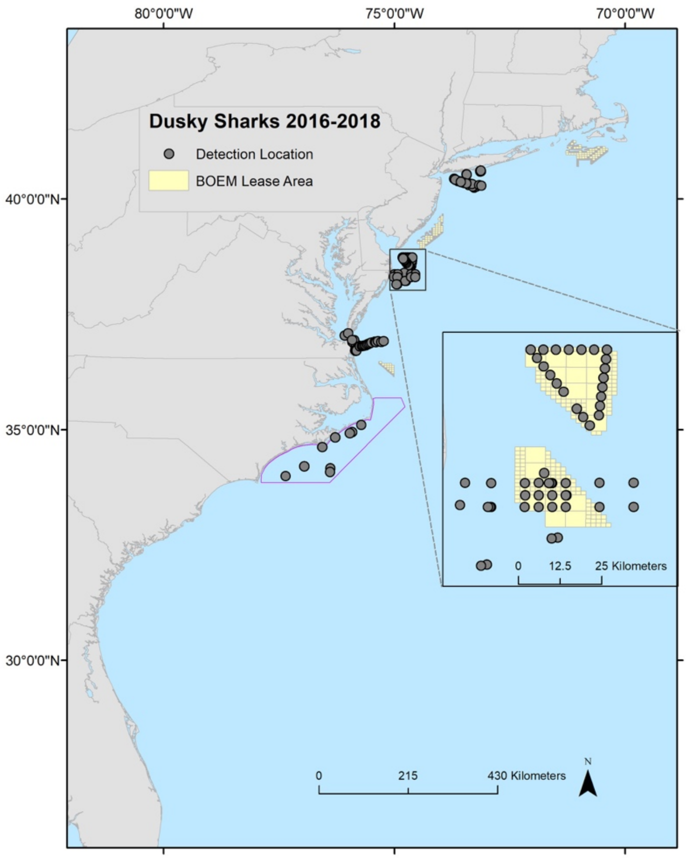
Daily environmental data extracted at receiver locations from ERDDAP products:
Depth (m) – ETOPO1
SST (°C) – MODIS Aqua
Chl a (mg/m^3) – MODIS Aqua
Sal (psu) - SMAP

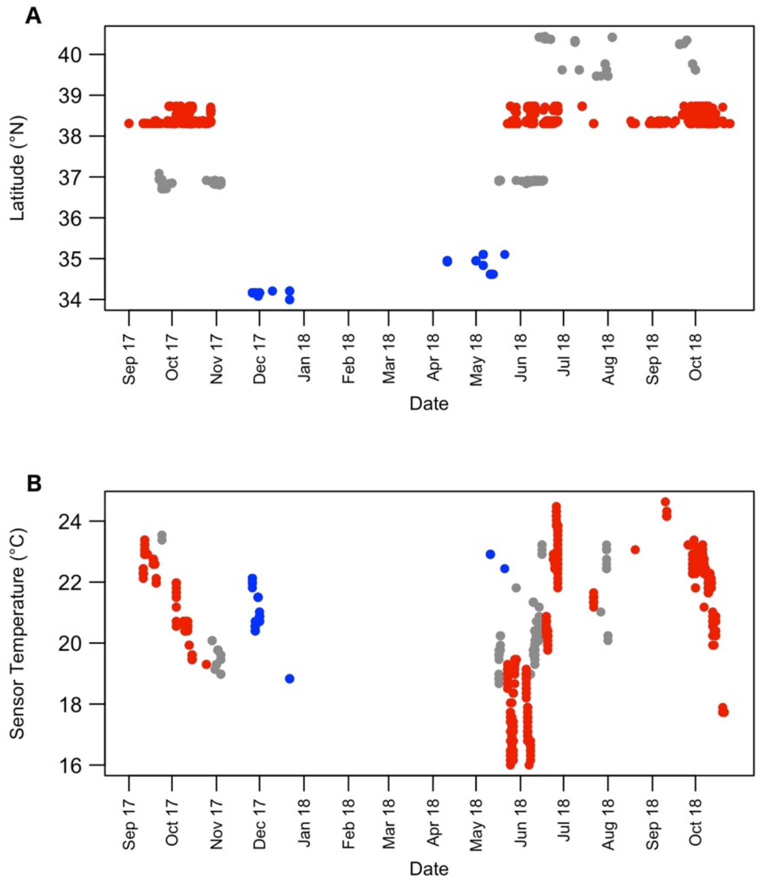
Seasonal and monthly (fall 2017) models

R Packages:
rerddapXtracto – data extraction
gbm.auto – BRT modeling

Case Study: Northwest Atlantic Dusky Sharks

Results – Tag Detections

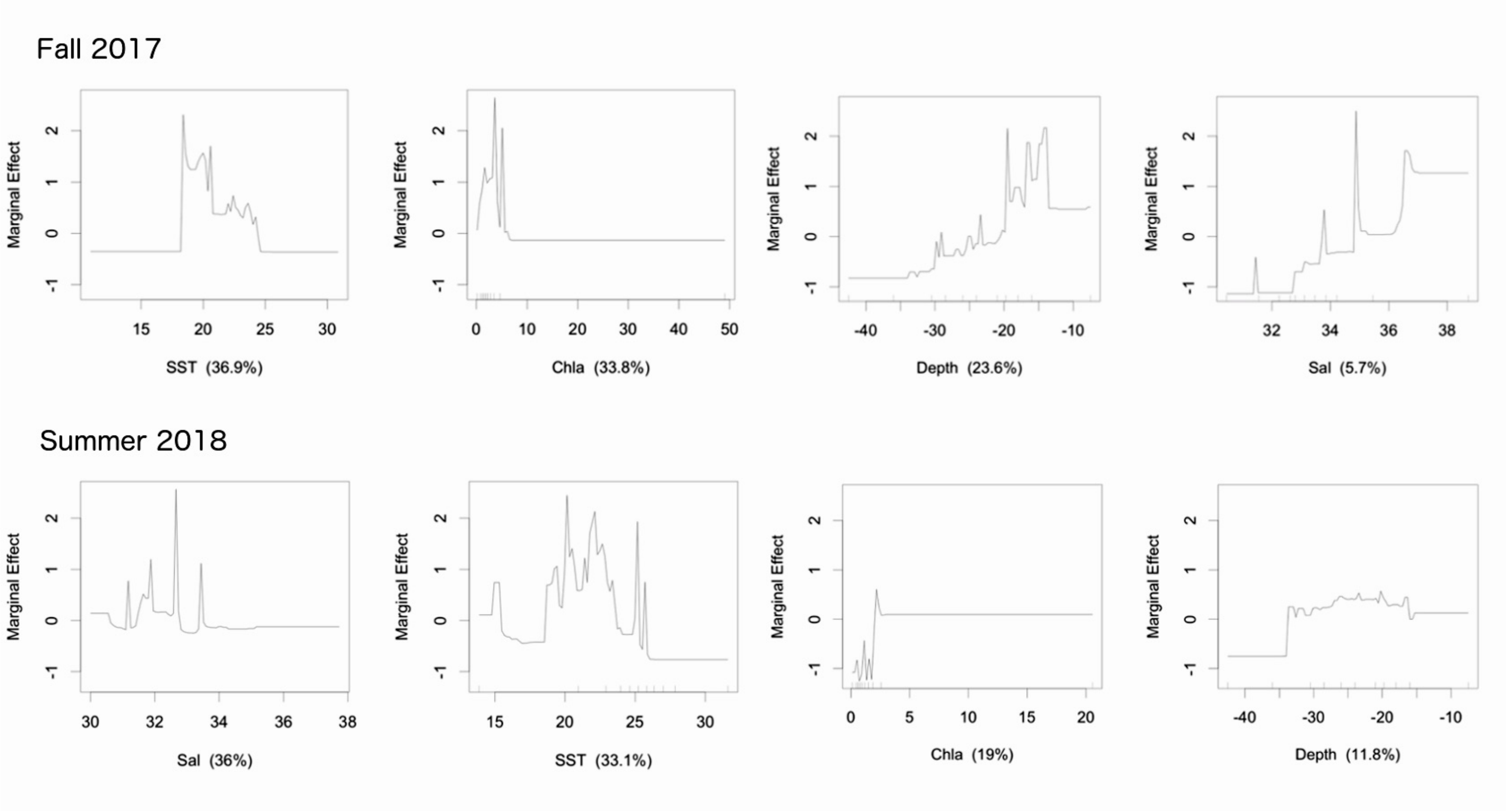




Within BOEM Lease Areas
Within Shark Closure

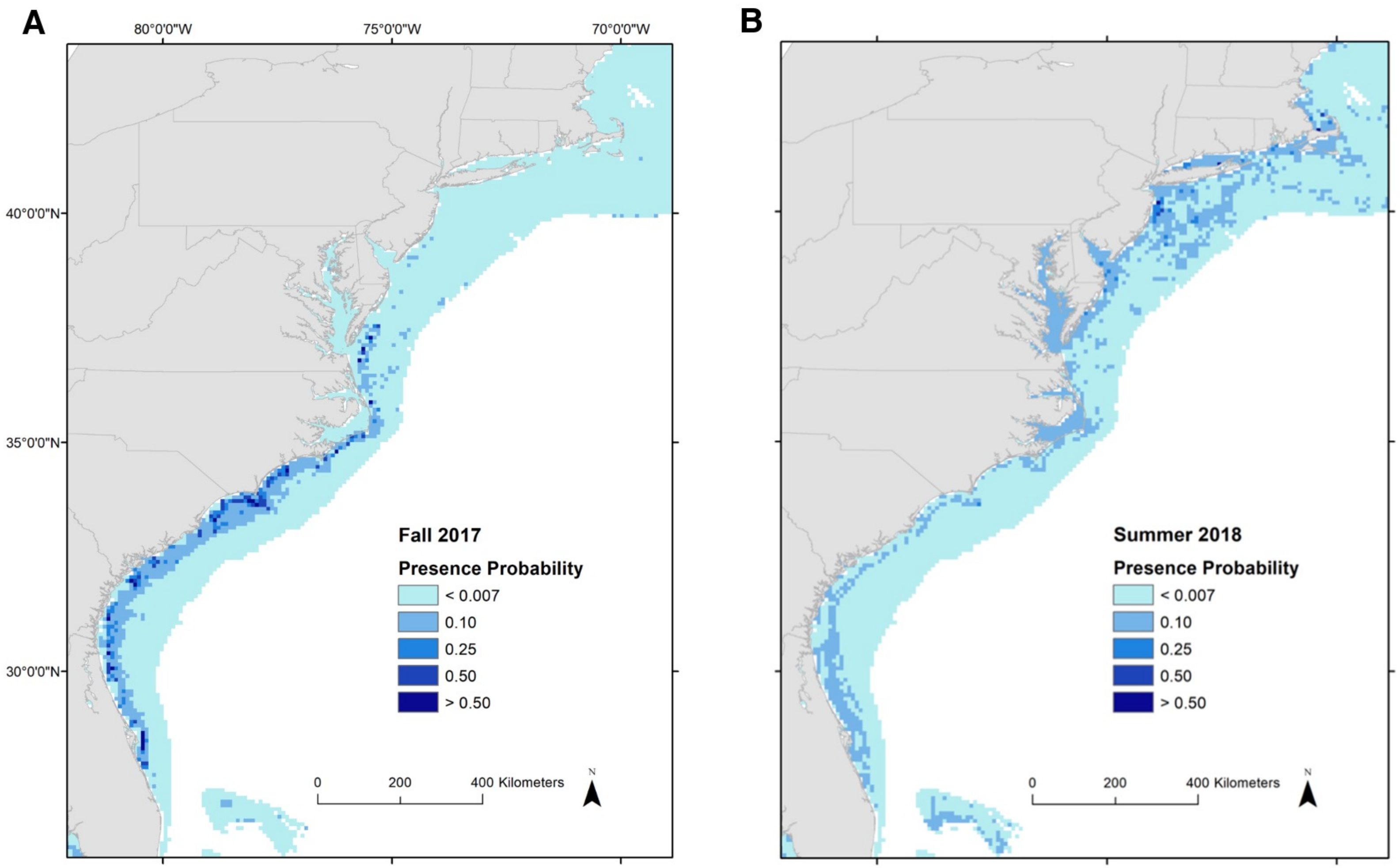
Case Study: Northwest Atlantic Dusky Sharks

Marginal effect plots – seasonal models



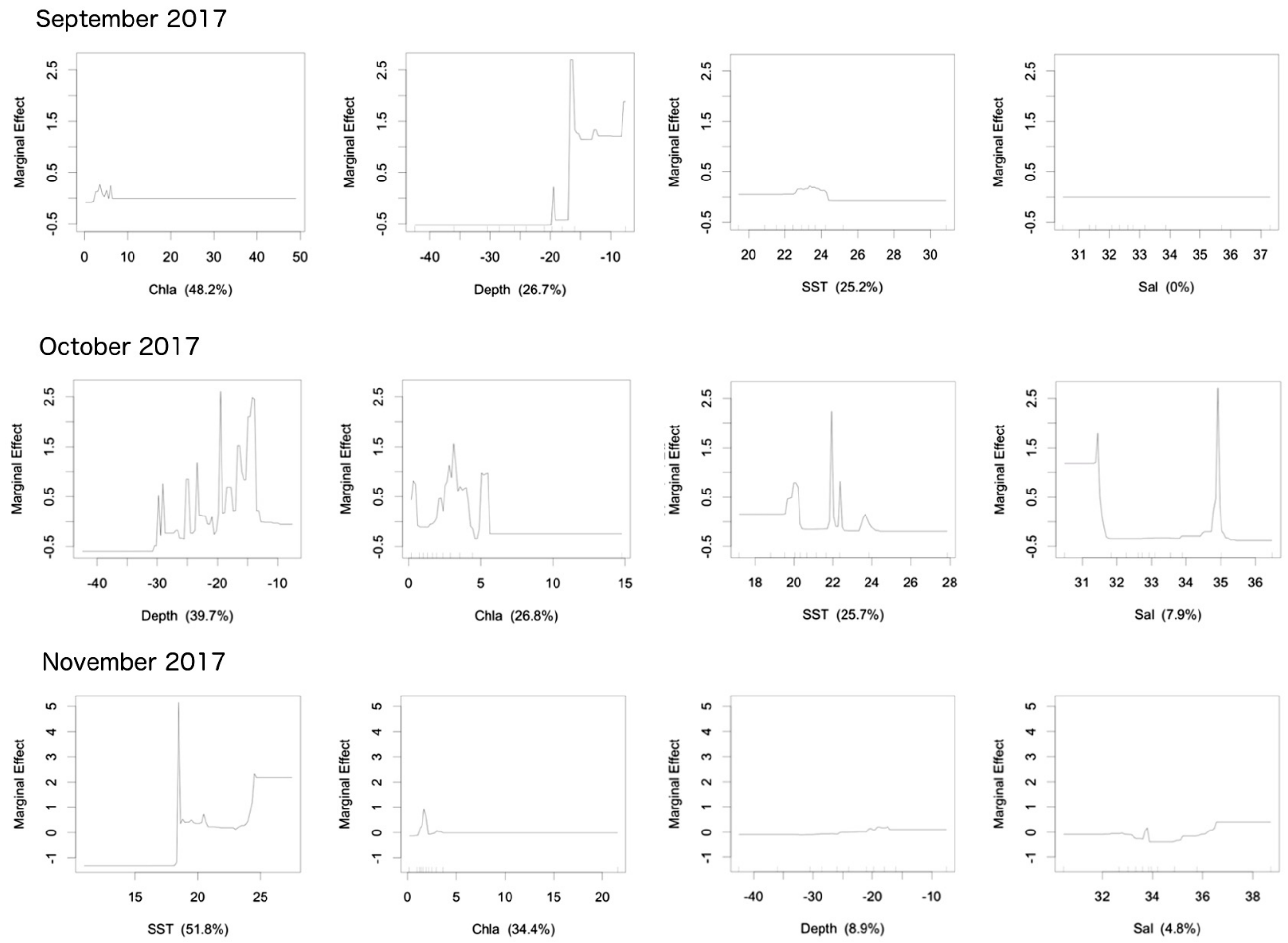
Case Study: Northwest Atlantic Dusky Sharks

Mapped model results – seasonal models



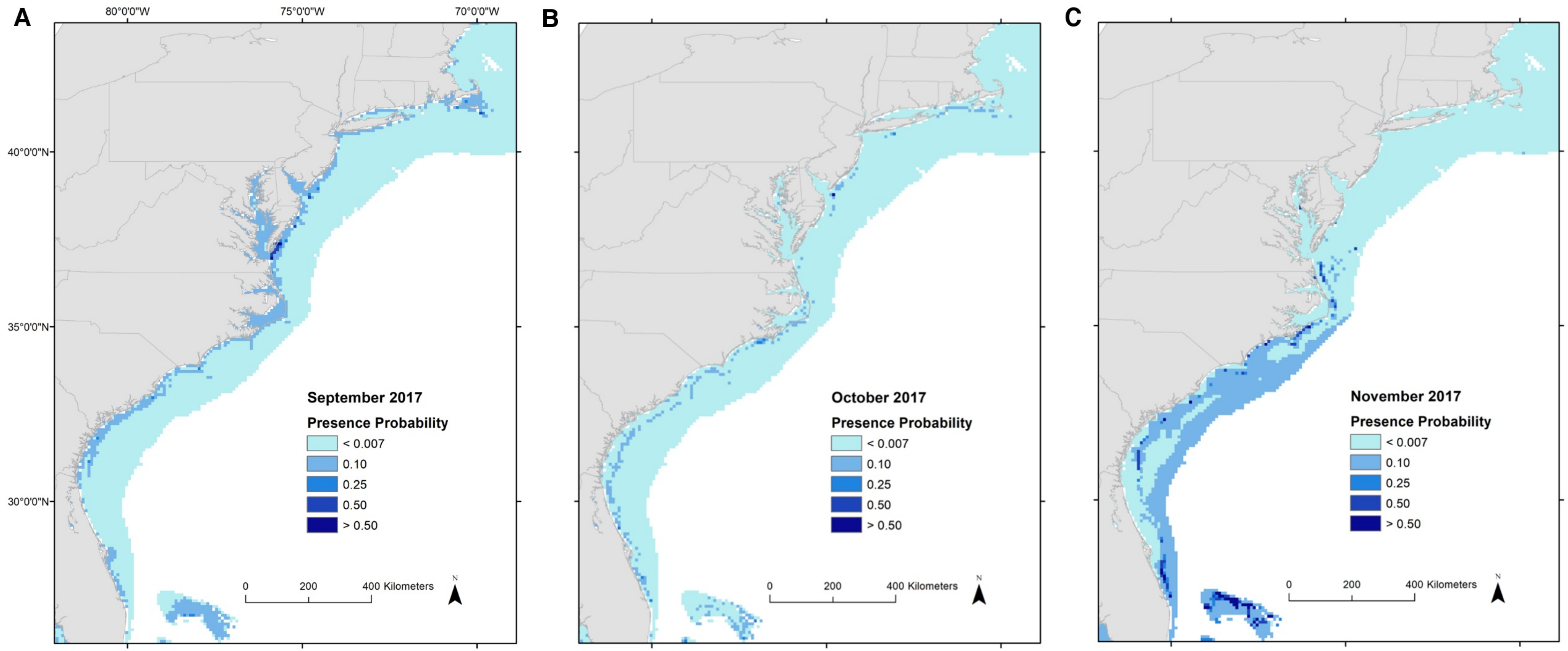
Case Study: Northwest Atlantic Dusky Sharks

Marginal effect plots – monthly models



Case Study: Northwest Atlantic Dusky Sharks

Mapped model results – monthly models



References and Recommended Reading

Bangley, C. W., L. Paramore, S. Dedman, and R. A. Rulifson. 2018. Delineation and mapping of coastal shark habitat within a shallow lagoonal estuary. PLOS ONE 13:e0195221.

Bangley, C. W., T. H. Curtis, D. H. Secor, R. J. Latour, and M. B. Ogburn. 2020. Identifying important juvenile dusky shark habitat in the Northwest Atlantic Ocean using acoustic telemetry and spatial modeling. Marine and Coastal Fisheries 12: 348-363

Elith, J., J. R. Leathwick, and T. Hastie. 2008. A working guide to boosted regression trees. The Journal of Animal Ecology 77:802–13.

Elith, J., and J. R. Leathwick. 2011. Boosted regression trees for ecological modeling. 22pp.

Dedman, S., R. Officer, D. Brophy, M. Clarke, and D. G. Reid. 2015. Modelling abundance hotspots for data-poor Irish Sea rays. Ecological Modelling 312:77–90.

Dedman, S., R. Officer, M. Clarke, D. G. Reid, and D. Brophy. 2017. Gbm.auto: a software tool to simplify spatial modeling and Marine Protected Area planning. PLOS One 12: e0188955.

Drymon, J. M., S. Dedman, J. T. Froeschke, E. A. Seubert, A. E. Jefferson, A. M. Kroetz, J. F. Mareska, and S. P. Powers. 2020. Defining sex-specific habitat suitability for a northern Gulf of Mexico shark assemblage. Frontiers in Marine Science 7: 35.

Froeschke, J., G. Stunz, and M. Wildhaber. 2010. Environmental influences on the occurrence of coastal sharks in estuarine waters. Marine Ecology Progress Series 407:279–292.

Grubbs, R. D., and J. A. Musick. 2007. Spatial delineation of summer nursery areas for juvenile sandbar sharks in Chesapeake Bay, Virginia. Pages 63–85 *in* C. T. McCandless, N. E. Kohler, and H. L. Pratt, Jr., editors. Shark nursery grounds of the Gulf of Mexico and the east coast waters of the United States. American Fisheries Society Symposium 50. American Fisheries Society, Bethesda, MD.

Questions?