

# Applications of statistics and machine learning in environmental science

Vyacheslav Lyubchich

Chesapeake Biological Laboratory, UMCES

2025-01-15

# Stats

## **Research 51%**

60+ peer-reviewed articles

4 R packages

80+ pubs total

\$16M+ research funding

## **Teaching and student advising 49%**

43 credit hours taught (61 with co-instructors)

400+ students

20+ undergraduates supervised

15 graduate committees

# Outline

Scientific discovery

Integration

Application

Teaching

Future directions

# Scientific discovery: **Quantifying linked rare events in fish and environmental time series**

with

Genny Nessler (CBL, UMCES), Eric Durell (MD DNR),  
Troy Tuckey (VIMS) & Mary C. Fabrizio (VIMS)



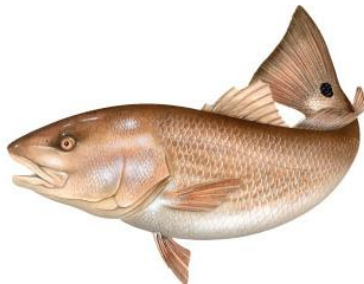
**NOAA**  
FISHERIES

# FY23 Chesapeake Bay Fisheries Research Program





## NOAA Chesapeake Bay Office

### Goals:

- 1) Identify linkages between fish catch-per-unit-effort and environmental **rare events** in Chesapeake Bay using time series analysis and machine learning techniques
- 2) Explore performance of predictive models that quantify the impact of rare environmental events on fish/shellfish stocks



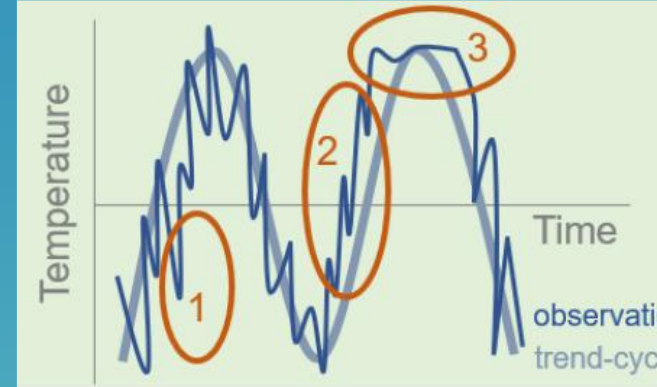
# Methods

|                   | Rare for fishery?   |   |
|-------------------|---|---|
|                   | Yes   | No  |
| Yes               |  |  |
| Rare for habitat? |   |   |
| No                |  |  |

Identify rare events

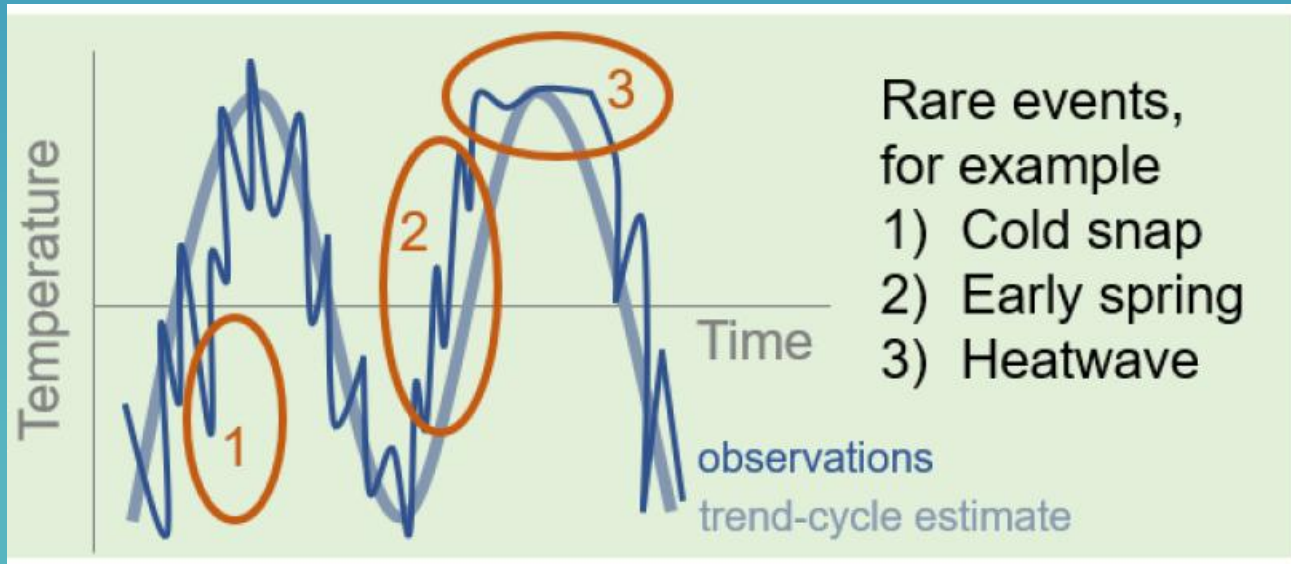
Identify linkages

Develop predictive models



# Identify rare events

Seldom occurring and inherently happening in the time domain (Carreño et al. 2020)



# Time series decomposition

Trend-seasonal model for environmental data (no seasonality for fish time series)

$$Y_t = M_t + S_t + \epsilon_t$$

- Non-parametric estimates of trend and seasonality using loess

Compare seasonality estimates

- Parametric quadratic trends and two pairs of Fourier series with periods 12 and 6 months

$$M_t$$

$$=$$

$$\alpha_0 + \alpha_1 t + \alpha_2 t^2$$

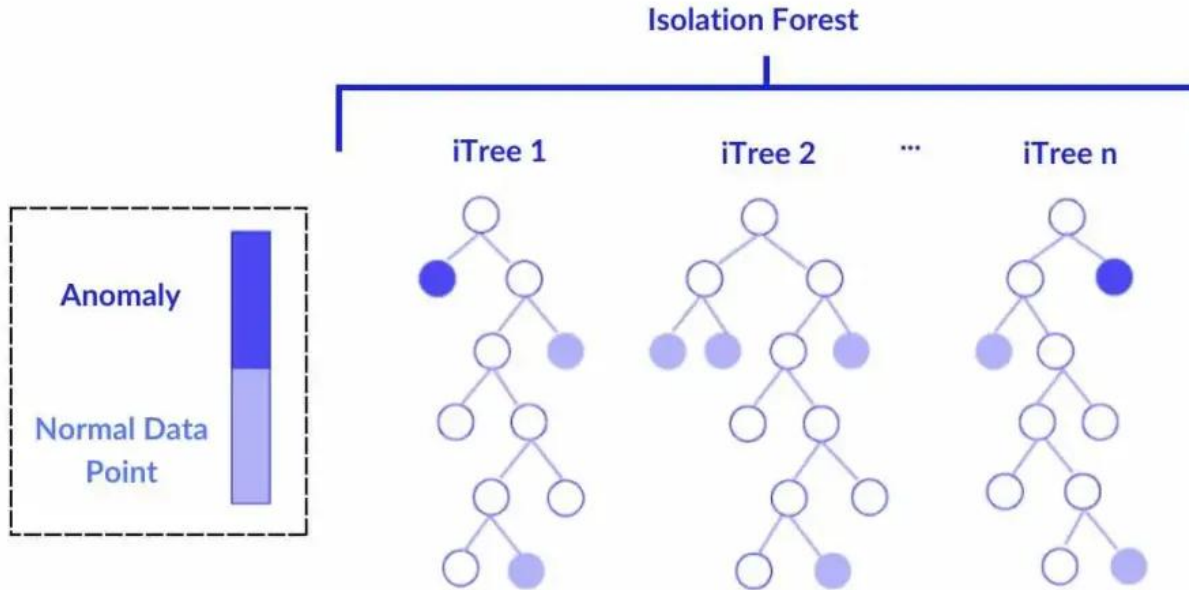
$$S_t$$

$$=$$

$$\beta_1 \cos_{1,t} + \beta_2 \sin_{1,t} + \beta_3 \cos_{2,t} + \beta_4 \sin_{2,t}$$



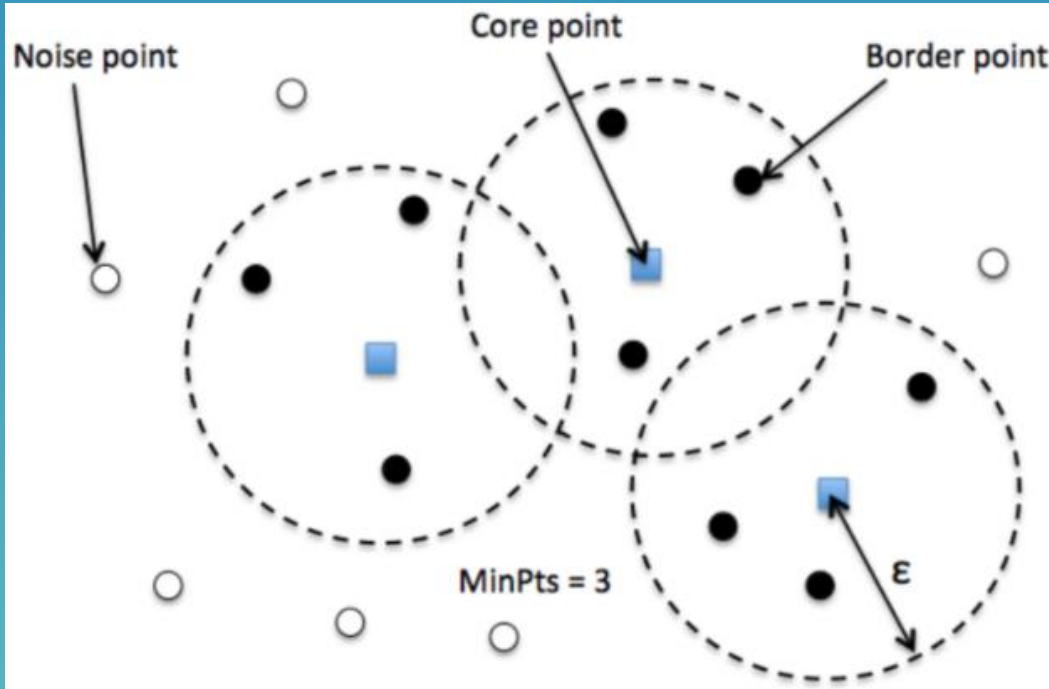
# Individual event identification: RF



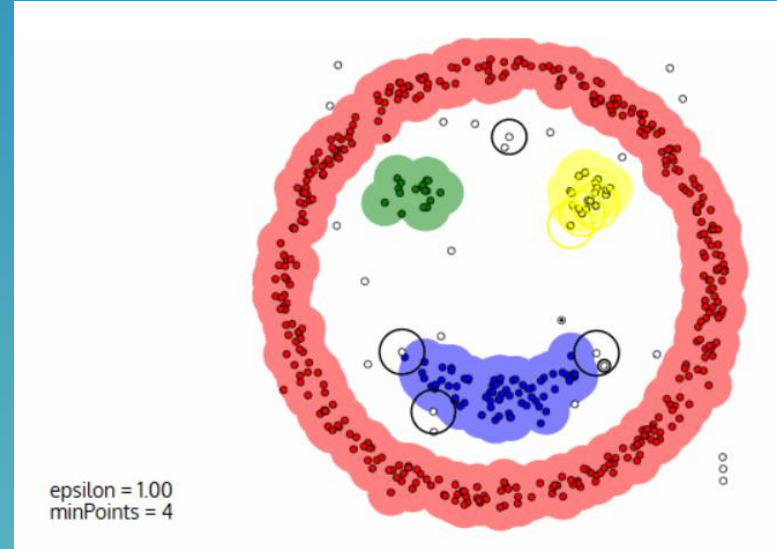
<https://spotintelligence.com/2024/05/21/isolation-forest/>

Average path length (number of splits to isolate an observation within a tree) is the anomaly score

# Individual event identification: DBSCAN



<https://www.kdnuggets.com/2020/04/dbscan-clustering-algorithm-machine-learning.html>



<https://www.digitalvidya.com/blog/the-top-5-clustering-algorithms-data-scientists-should-know/>

Select the noise points

# Identify linkages

Time series lags used to consider delayed effects (during fish spawning) of environmental variables on corresponding fisheries.

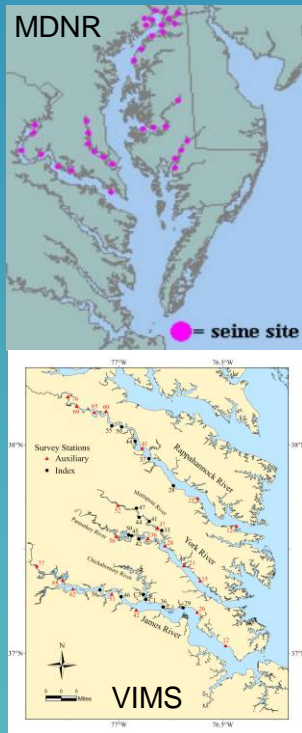
$$X(\text{April}) + X(\text{May}) + X(\text{June}) \sim Y$$

Run a chi-square test on positive and negative anomalies detected by each algorithm.

# Data sources

## Fish/shellfish

- Maryland DNR
  - Seine Survey
  - Striped Bass Spawning Stock Survey
- VIMS
  - Seine Survey
  - Juvenile Trawl Survey
  - ChesMMAP
- Maryland DNR & VIMS
  - Blue crab Winter Dredge Survey
  - Blue crab Trawl Survey
- CBSAC & ASMFC landings

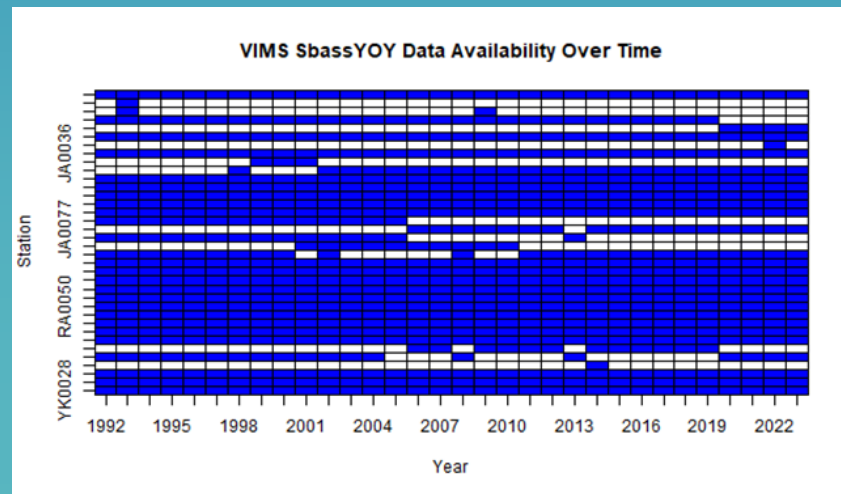
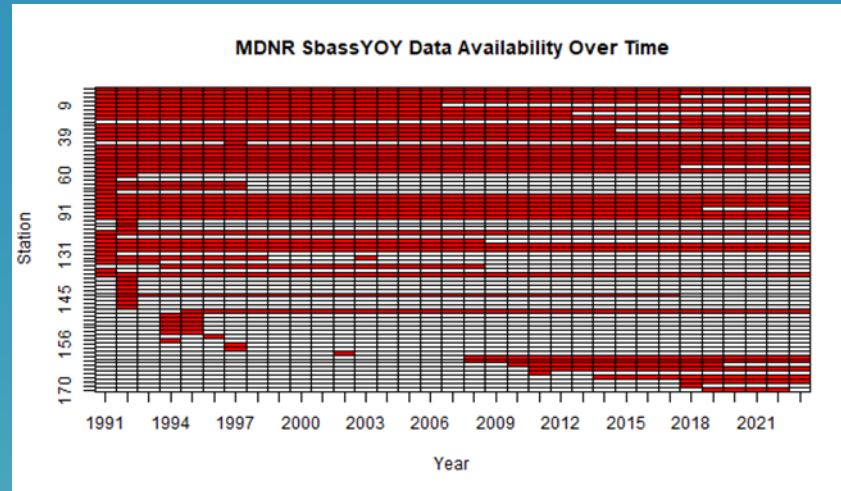


## Environmental

- NASA Daymet – min/max temp & precipitation
- Chesapeake Bay Environmental Forecast System (CBEFS)
- Chesapeake Bay Interpretive Buoy System (CBIBS)
- VIMS Submerged Aquatic Vegetation Program

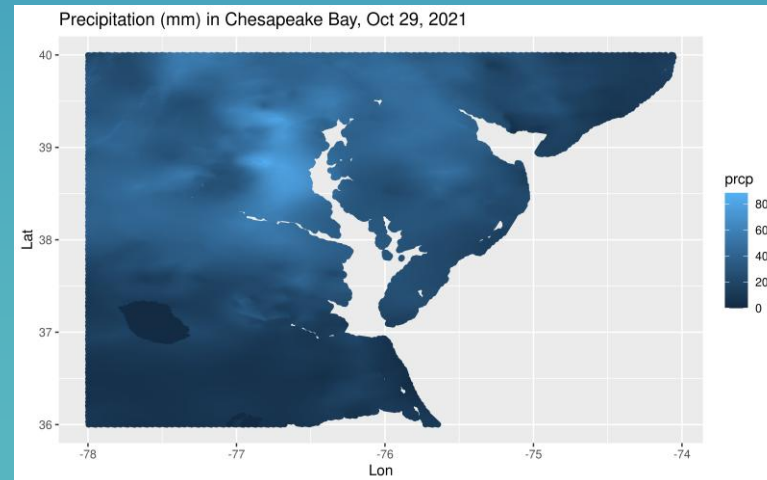
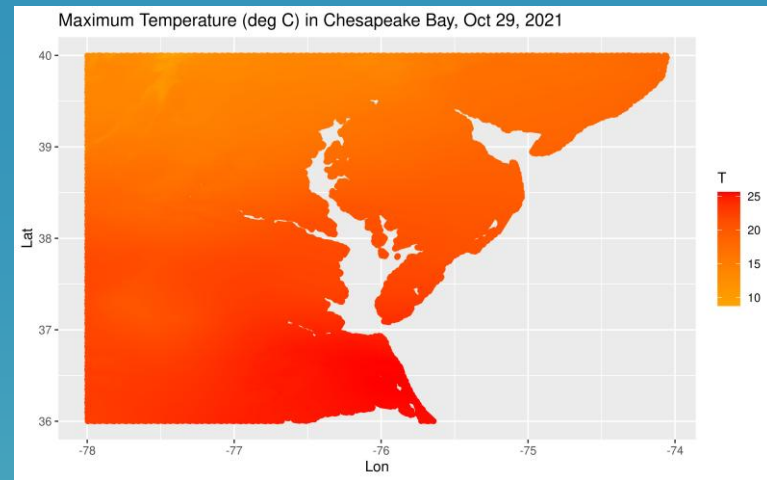
# Data treatment 1

- Fish seine survey effort data collection consistent from 1991+, so time series spanned 1991-2023
- Stations with sampling in >80% of years were used in analyses
- 27 stations in Maryland and 22 stations in Virginia



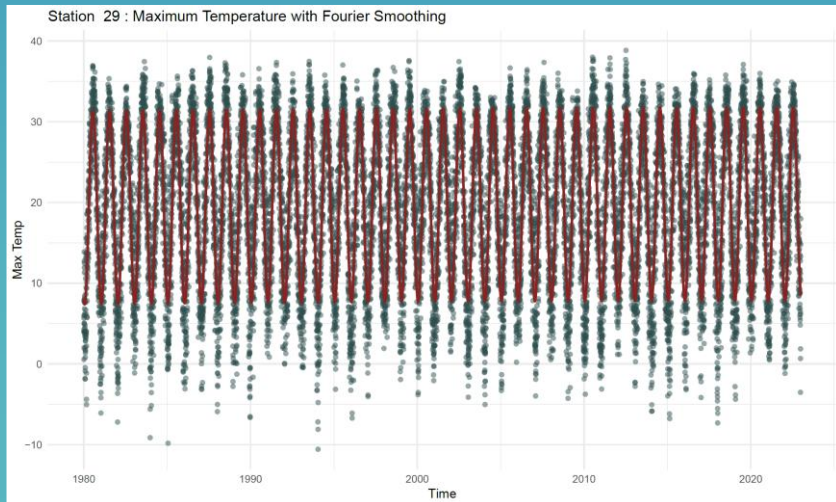
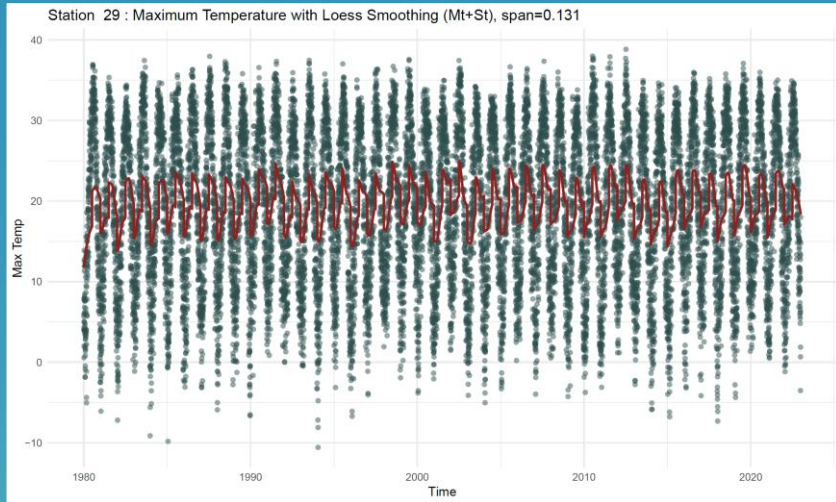
# Data treatment 2

- Selected Daymet cell nearest to each fish sampling station to represent nearby weather conditions
- At each individual fish sampling station, both environmental and fish CPUE rare events were tallied across the time series
- To identify relationships between environmental and fish CPUE, we tallied Daymet rare events March-June annually and all sampling events in a year for fish seine surveys



# Implementation

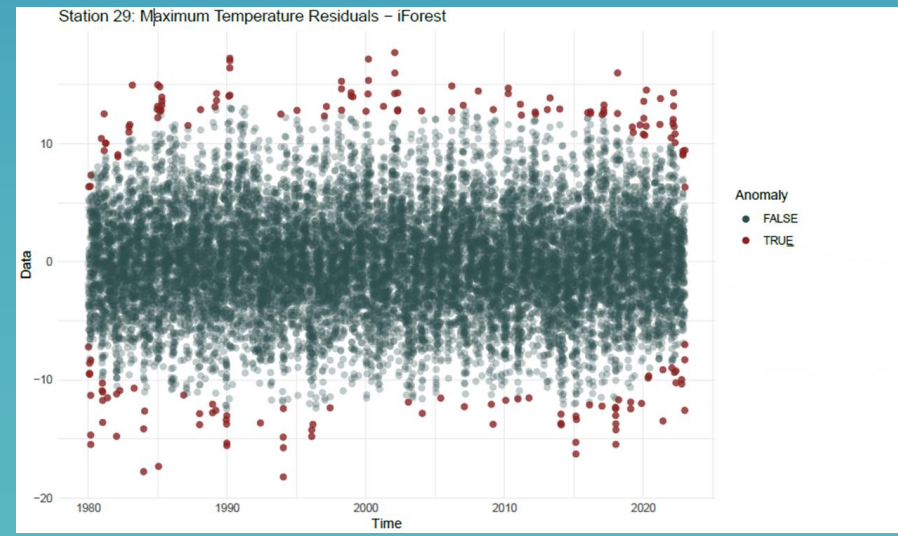
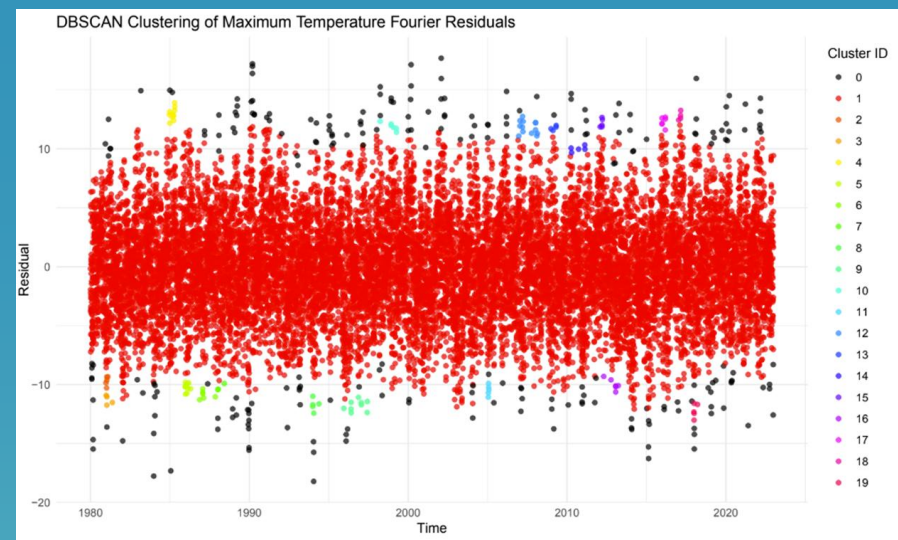
1. Smoothed all fish and environmental time series using locally weighted smoothing (loess) or trend + Fourier series





# Implementation

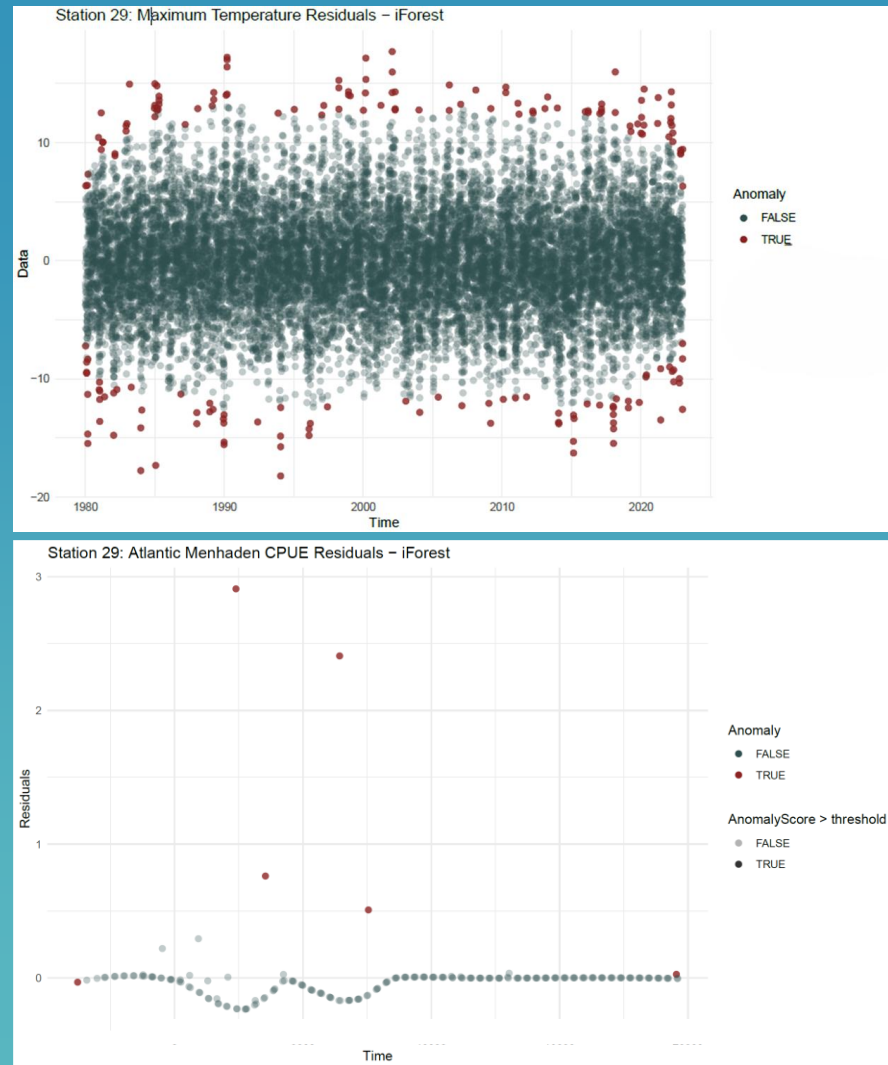
1. Smoothed all fish and environmental time series using locally weighted smoothing (loess) or trend + Fourier series
2. Identified rare events in both fish and environmental residuals using isolation forest and density-based clustering (DBSCAN) algorithms





# Implementation

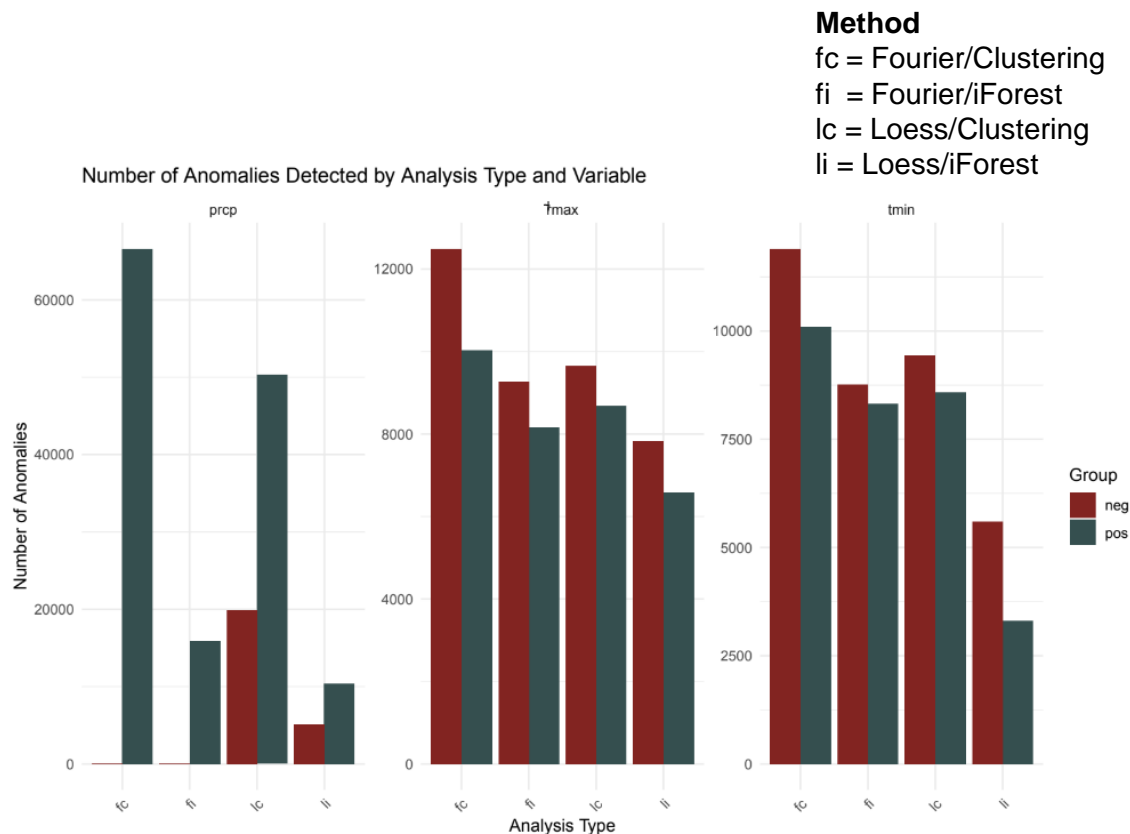
1. Smoothed all fish and environmental time series using locally weighted smoothing (loess) or trend + Fourier series
2. Identified rare events in both fish and environmental residuals using isolation forest and density-based clustering (DBSCAN) algorithms
3. Quantified significance of relationships between rare fish events and rare weather events (March-June) using chi-square tests



# Preliminary results

## Rare event detection

- More precipitation rare events, more high than low
- Similar number of pos/neg temperature rare events
- Smoothing method (Fourier vs loess) had minimal effect on rare event detection
- DBSCAN identified more rare events than Isolation Forest
- In general, most sensitive combination of methods was FC and least was LI



# Preliminary results

- For both species, more relationships identified between CPUE and min temperature than max temperature or precipitation
- Depending on algorithms used, ~20-39% of relationships between CPUE and min temperature rare events were significant
- Higher proportion of significant relationships identified when # rare events was lower. Focus on more strict methods to yield fewer false positives.

## Atlantic menhaden

Proportion of Significant Relationships and Number of Anomalies Detected by Test - Menhaden

|                     | Fourier-iForest | Fourier-Clustering | Loess-iForest | Loess-Clustering |
|---------------------|-----------------|--------------------|---------------|------------------|
| Precipitation       | 0.027/15953     | 0.14/66662         | 0.013/15444   | 0.127/70237      |
| Maximum Temperature | 0.14/17442      | 0.14/22533         | 0.247/18337   | 0.24/14424       |
| Minimum Temperature | 0.353/17093     | 0.267/22003        | 0.207/18022   | 0.347/8909       |

<sup>1</sup> p significant Relationships/n weather anomaly detected

<sup>2</sup> P<0.05

## Striped bass

Proportion of Significant Relationships and Number of Anomalies Detected by Test - Striped Bass

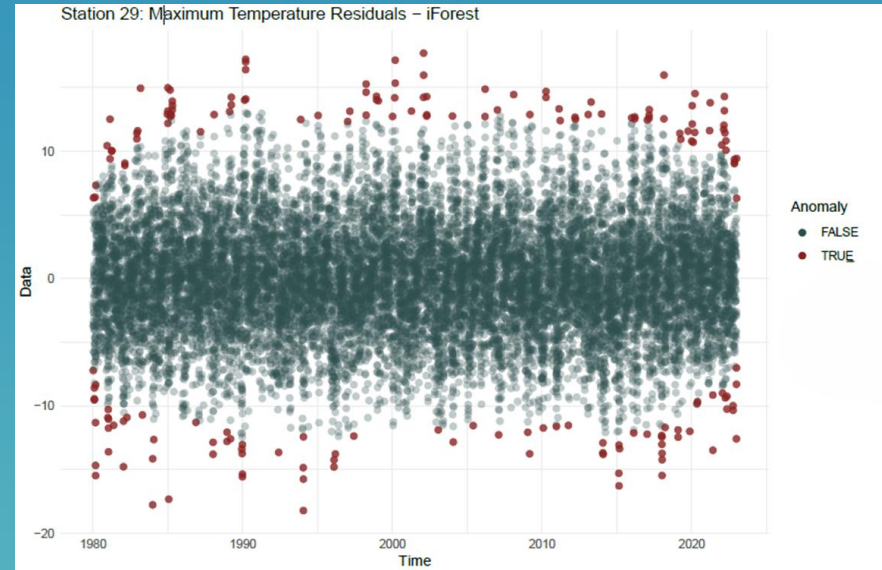
|                     | Fourier-iForest | Fourier-Clustering | Loess-iForest | Loess-Clustering |
|---------------------|-----------------|--------------------|---------------|------------------|
| Precipitation       | 0.1/15953       | 0.12/66662         | 0.107/15444   | 0.113/70237      |
| Maximum Temperature | 0.12/17442      | 0.12/22533         | 0.26/18337    | 0.267/14424      |
| Minimum Temperature | 0.333/17093     | 0.267/22003        | 0.28/18022    | 0.387/8909       |

<sup>1</sup> p significant Relationships/n weather anomaly detected

<sup>2</sup> P<0.05

# Next steps

- Summarize rare event results by pos/neg anomalies
- Identify rare events in CBEFS hindcast estimates of water temp, salinity, DO, attenuation depth, and wave height
- Refine methods and apply to all YOY and adult fish and environmental data sources
- Develop predictive models
- Data analysis tools



# Integration: Ecological monitoring and inference for wind energy development

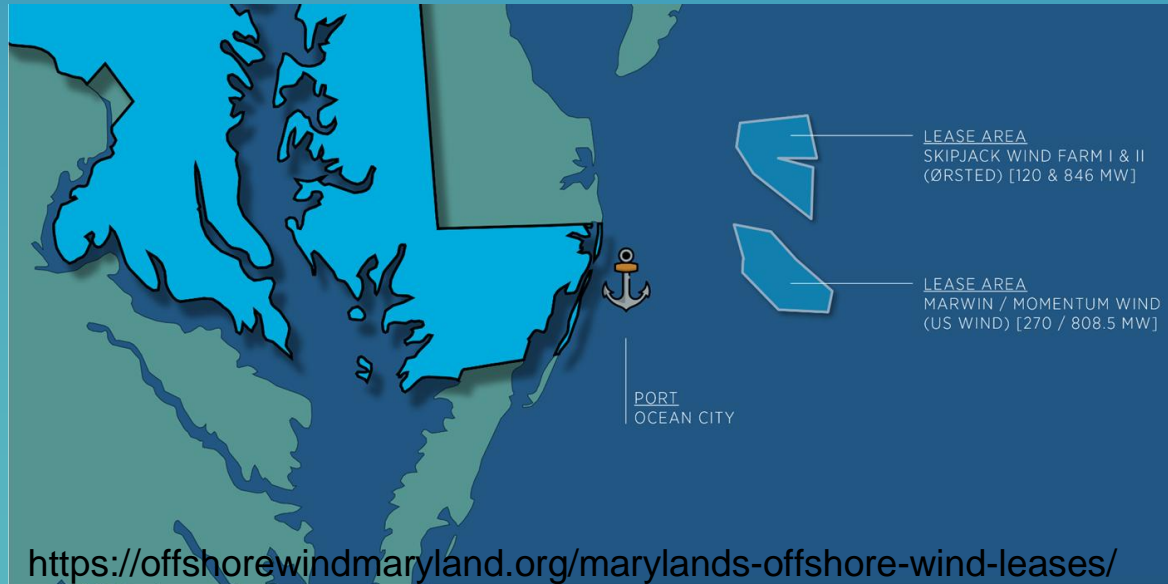
with



Team for **A**ssessing **I**mpacts to **L**iving resources from offshore **WIND**  
turbine**S** (UMCES, <https://tailwinds.umces.edu/>),  
Annamaria DeAngelis (NOAA), Bruna Pagliani (Fed. U. of Rio de Janeiro,  
Brazil)

# Motivation

- Conduct a passive acoustic monitoring (PAM) study within the area of potential effect for the Maryland Lease Area
- Improve marine mammal acoustic call detectors (distinguish between bottlenose and common dolphins)
- Ecological understanding, biodiversity monitoring



# Data processing: PamGuard + ROCCA

Data were processed in PamGuard using the Real-time Odontocete Call Classification Algorithm (ROCCA), a random forest classifier that measures 50 features of each whistle contour

- ➡ ROCCA groups individual whistles into different encounters
- ➡ Encounters are given unique IDs representing collections of whistles assumed to be from a discrete individual or group of dolphins
- ➡ Whistle features measured by ROCCA, grouped by encounter IDs, were exported and used to train the models

ROCCA was originally trained on whistles of 8 delphinid species in the tropical Pacific Ocean

- ➡ Whistle structure has been shown to vary across region and population
- ➡ Training classification algorithms on data from the region and/or species of interest are most effective

# Data processing: PamGuard + ROCCA

Acoustic files included either confirmed only common dolphin or most probable only bottlenose dolphin whistles in the Atlantic Ocean

Common dolphin whistles from:

- ↓ NW Atlantic (NOAA AMAPPS)
- ↓ SW Atlantic off Brazil (Dr. Bruna Pagliani)
- ↓ Sable Island, Canary Island, and an unknown location (Watkins Marine Mammal Sound Database)

Bottlenose dolphin whistles from subset of data

- ↓ Mid-Atlantic region (previous Maryland DNR/BOEM study)



# Number of whistles available from each source and year along with the event counts

| Source  | Dolphin species | 1958 | 1975 | 1987 | 2014 | 2016 | 2017 | 2018 | Whistle, count | Whistle, % | Event, count | Event, % |
|---------|-----------------|------|------|------|------|------|------|------|----------------|------------|--------------|----------|
| Brazil  | Common          | 0    | 0    | 0    | 3580 | 0    | 0    | 0    | 3580           | 24.153     | 29           | 2.241    |
| NOAA    | Common          | 0    | 0    | 0    | 0    | 2637 | 0    | 0    | 2637           | 17.791     | 9            | 0.696    |
| T1C     | Bottlenose      | 0    | 0    | 0    | 0    | 5251 | 2049 | 1075 | 8375           | 56.504     | 1225         | 94.668   |
| Watkins | Common          | 131  | 97   | 2    | 0    | 0    | 0    | 0    | 230            | 1.552      | 31           | 2.396    |
| Total   | –               | 131  | 97   | 2    | 3580 | 7888 | 2049 | 1075 | 14822          | 100        | 1294         | 100.001  |

VS



# Methods

BANTER (Rankin et al. 2017)

NN + GLM

RF + Boruta (variable selection; Kursa & Rudnicki 2010) + GLM

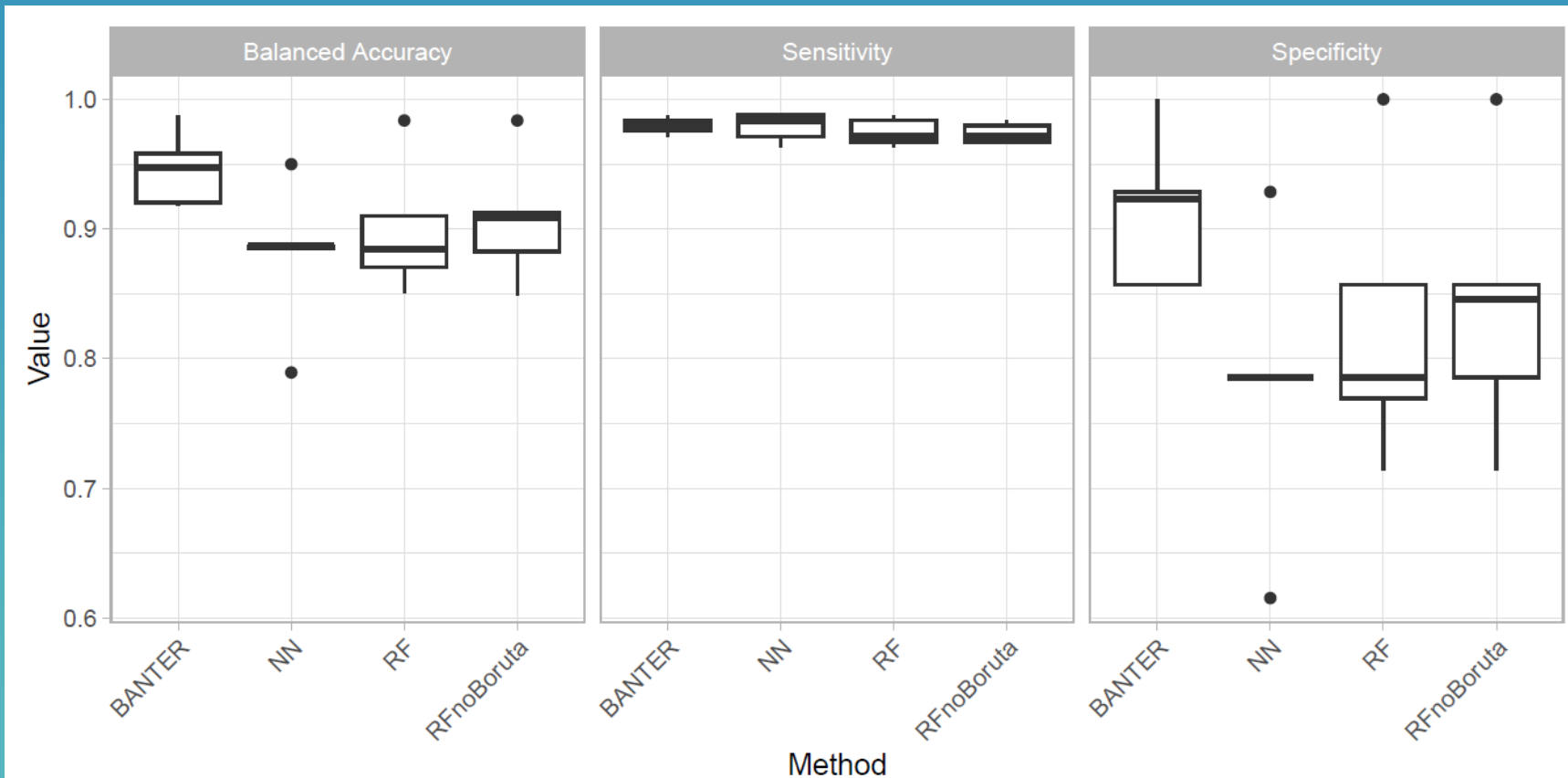
RF + GLM

# Model assessment

5-fold cross-validation applied 5 times

Performance metrics: balanced accuracy, sensitivity, specificity

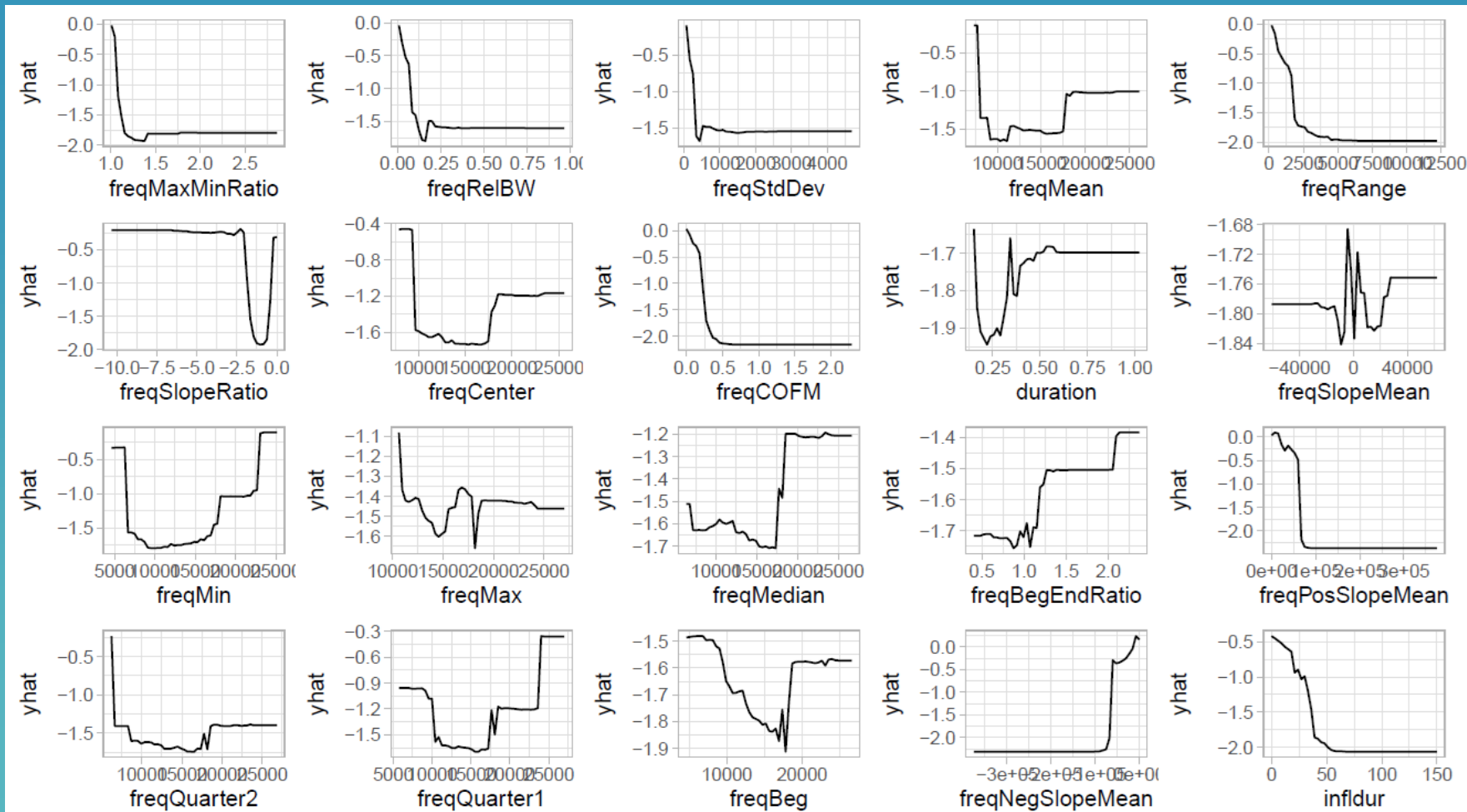
# Classification performance in cross-validation



## Accuracy (%) at different levels of retrained BANTER

| Species    | Whistle | Event |
|------------|---------|-------|
| Bottlenose | 91.5    | 97.7  |
| Common     | 82.6    | 91.3  |
| Overall    | 87.7    | 97.4  |

# Partial dependence plots for the most important variables



# Conclusions

- A retrained BANTER model outperformed other models by classifying events with bottlenose dolphins more accurately
- Partial dependence plots help to identify thresholds for distinguishing the two species of dolphins
- Classification of unlabeled (without ground truth) recordings from the Maryland Wind Energy Area reveals predominantly bottlenose dolphins and a few common dolphins in the winter, mainly December to May



# Application: **Machine learning of factors for improving oyster hatchery production**

with

Mathew Gray (HPL, UMCES), Greg Silsbe (HPL, UMCES)

<https://vlyubchich.github.io/OysterHatcheryYield/>

# Teaching and advisement

# Courses

Environmental statistics 1 (MEES 613, 3 credits)

Environmental statistics 2 (MEES 713, 3 credits)

R programming basic (MEES 602, 1 credit) – self-paced  
with open-source materials (thanks to UMD TLC grant)

R programming advanced (MEES 702, 1 credit)

# Students

400+ students taught

15 graduate committees (10 completed)

20+ undergraduate interns (Maryland Sea Grant NSF REU, College of Southern Maryland, St. Mary's College of Maryland, etc.)

# Future directions

Data-driven solutions to environmental problems

Causality

Machine learning for big data, PAM recordings

High-dimensional inference

# Questions?

lyubchich@umces.edu

# Thank you!