Quantifying linked rare events in fish and environmental Chesapeake Bay time series

Genny Nesslage & Slava Lyubchich

UMCES Chesapeake Biological Laboratory

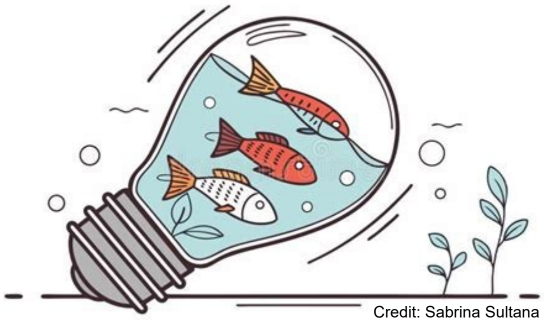
CBL Seminar Series

February 25, 2026

Credit: NASA

NOOA Chesapeake Bay Office
Chesapeake Bay Fisheries Research Program

Call: Investigate impacts of climate change, habitat quality, and/or changing environmental variables on fish and shellfish resources to inform sustainable and ecosystem based fisheries management



Credit: Sabrina Sultana

FAQ 11.2: **Will climate change cause unprecedented extremes?**

Yes, in a changing climate, extreme events may be unprecedented when they occur with...

Larger magnitude

Increased frequency

New locations

Different timing

New combinations (compound)

Questions

- Can we detect rare events that have meaningful impact on fish/shellfish?
- What kind of impact do they have (+/-)?
- Are the impacts the same or different across species and age/sex classes?
- Can we predict impacts of future rare events on stock dynamics?

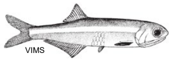

Credit: Stephen Badger MDNR

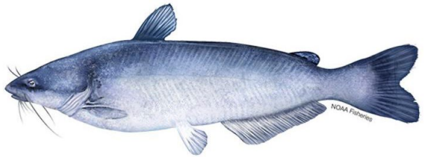

Credit: Patty Brinkmeyer

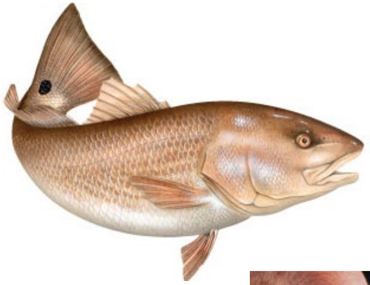
Goals

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





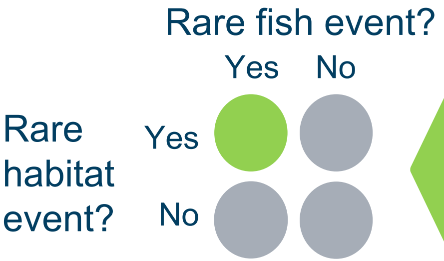


Credit: Chesapeake Bay Program

Overview

- Data and analysis approach
- Methods development
- Example results
- Next steps

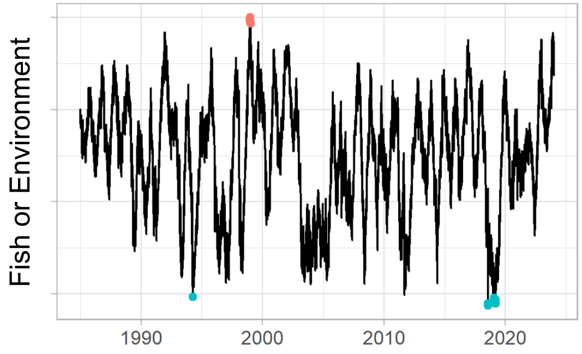
Data-driven approach

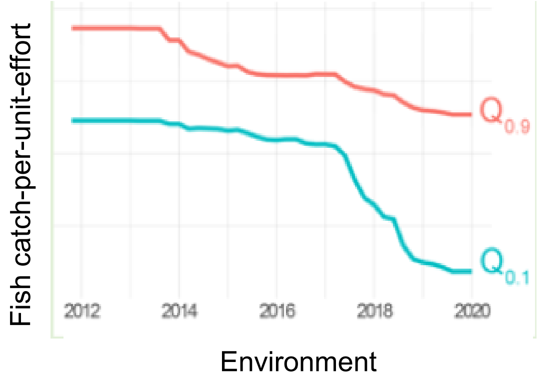


Identify rare events

Identify linkages

Develop predictive models



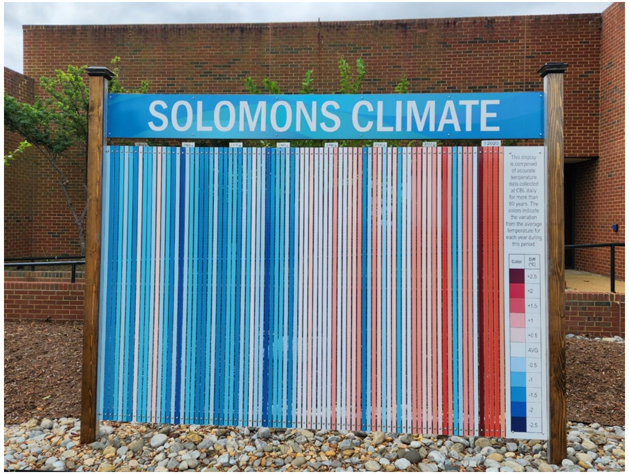


Advantages of data-driven approach

1) Reduce need for predetermined assumptions about
   -how to define environmental anomalies
   -metabolic/ecological thresholds

2) Account for climate change impacts across time series
   -directional shifts in environment response
   -adaptations/changes in fish/shellfish response to environment

3) Explore impacts of assumptions we do have to make (e.g., time lags)



Collaborators

Julie Reichert-Nguyen

NOAA FISHERIES

Eric Durrell
Beth Versak
Glenn Davis

MARYLAND DEPARTMENT OF NATURAL RESOURCES

Troy Tuckey & Mary Fabrizio
Marjy Friedrichs & Pierre St-Laurent
Romuald Lipcius
James Gartland & Robert Latour

VIMS | WILLIAM & MARY
VIRGINIA INSTITUTE OF MARINE SCIENCE

Fish/Shellfish Data Sources

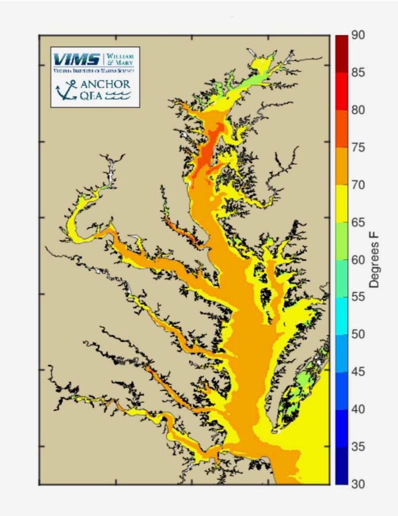
- Maryland DNR
  - -Juvenile Fish Seine Survey
  - -Striped Bass Spawning Stock Survey
  - -Blue Crab Trawl Survey
- VIMS
  - -Seine Survey
  - -Juvenile Finfish Trawl Survey
  - -ChesMMAP
- Maryland DNR & VIMS
  - -Blue Crab Winter Dredge Survey

Credit: Brittnee Barris


Credit: Maryland DNR


Credit: Maryland DNR

Environmental Data Sources

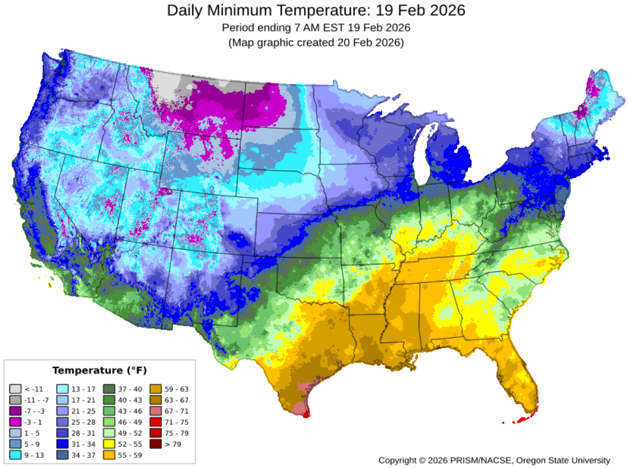


- Chesapeake Bay Environmental Forecast System (CBEFS) hindcasts of daily water condition estimates at bottom and surface: salinity, temp/DO, pH, diffuse attenuation

VIMS | William & Mary
Virginia Institute of Marine Science

- Parameter-elevation Regressions on Independent Slopes Model (PRISM) daily weather pattern estimates of min/max/avg: air temp, precipitation

Oregon State University



Data preparation

For each fish survey:

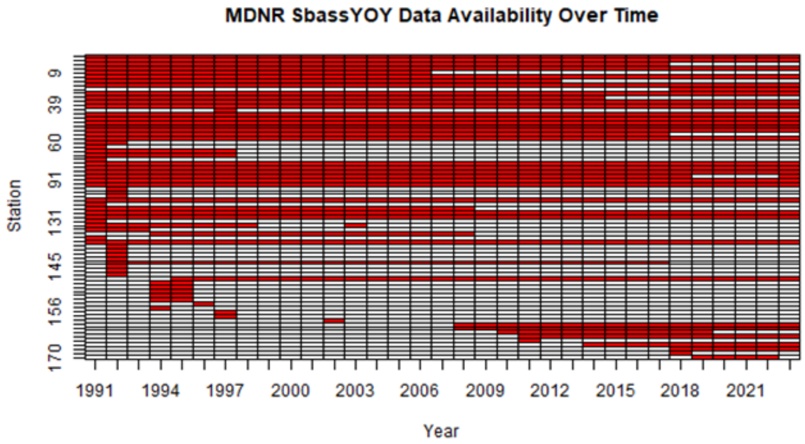
1. Identify stations/regions with adequate time series for each species (and age/sex class, when available)

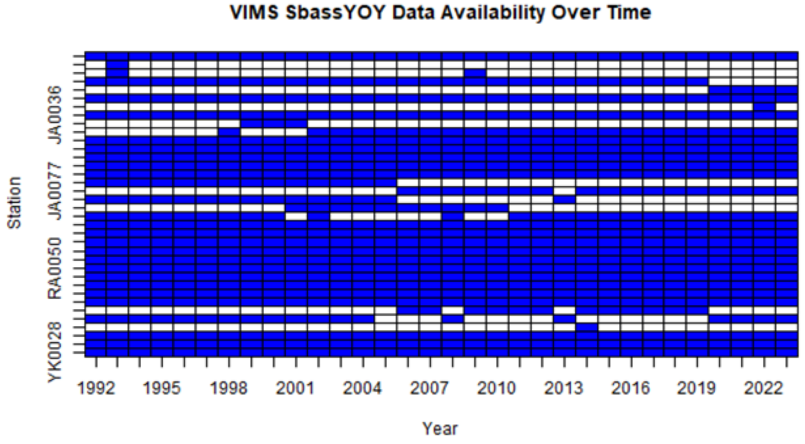
2. Average seasonal catch-per-unit-effort (CPUE) for each species (and age/sex class, when available)

3. Associate survey stations (fixed) or sampling locations (random) to subregions to align with environmental data in space and time

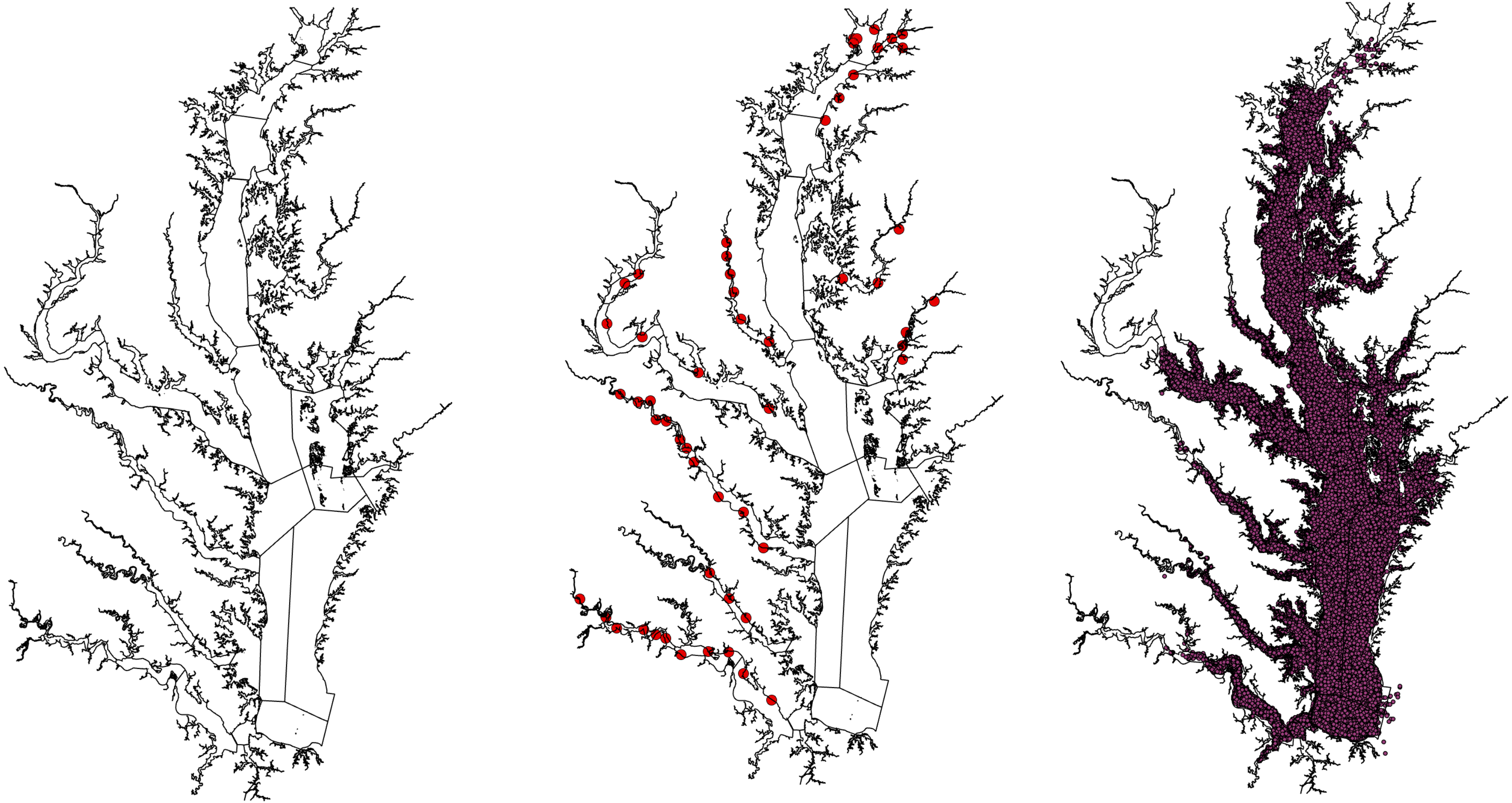
Data preparation

- Maryland DNR and VIMS seine survey time series: 1991/1992-2023
- Selected stations with catch per unit effort sampling in >80% of years
- 27 stations in Maryland and 22 stations in Virginia





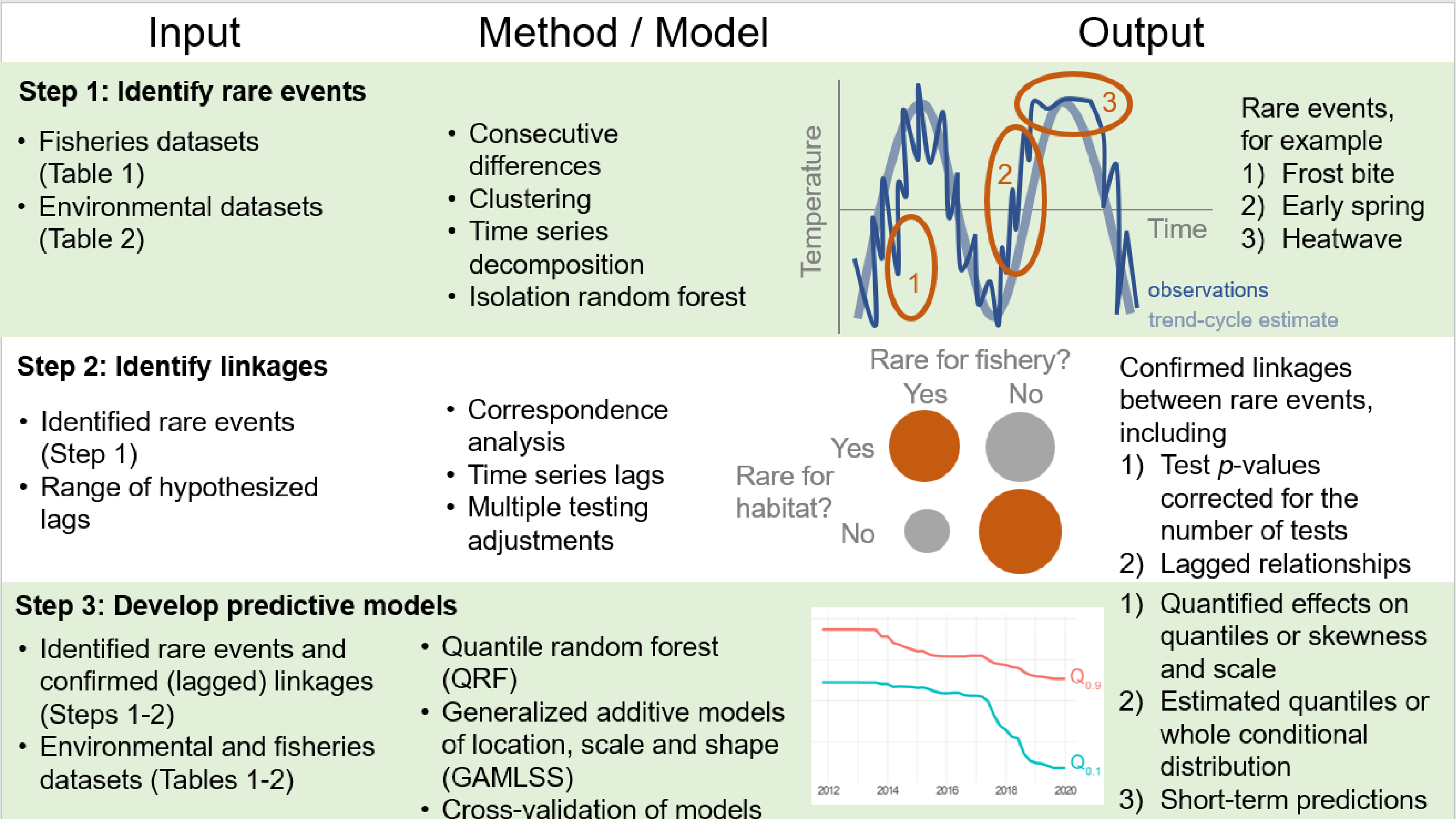
Data preparation



Chesapeake Bay Program subregions

Overview

Data and analysis approach → Methods development → Example results → Next steps


Input
Method / Model
Output
Step 1: Identify rare events
• Fisheries datasets (Table 1)
• Environmental datasets (Table 2)
• Consecutive differences
• Clustering
• Time series decomposition
• Isolation random forest
Temperature
Time
1
2
3
observations
trend-cycle estimate
Rare events, for example
1) Frost bite
2) Early spring
3) Heatwave
Step 2: Identify linkages
• Identified rare events (Step 1)
• Range of hypothesized lags
• Correspondence analysis
• Time series lags
• Multiple testing adjustments
Rare for fishery?
Yes
No
Rare for habitat?
Yes
No
Confirmed linkages between rare events, including
1) Test p-values corrected for the number of tests
2) Lagged relationships
Step 3: Develop predictive models
• Identified rare events and confirmed (lagged) linkages (Steps 1-2)
• Environmental and fisheries datasets (Tables 1-2)
• Quantile random forest (QRF)
• Generalized additive models of location, scale and shape (GAMLSS)
• Cross-validation of models
$Q_{0.9}$
$Q_{0.1}$
2012
2014
2016
2018
2020
1) Quantified effects on quantiles or skewness and scale
2) Estimated quantiles or whole conditional distribution
3) Short-term predictions

Methods: Time series decomposition

Trend-seasonal model for environmental data (without seasonality S_t for fish time series)

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

- Non-parametric estimates of trend and seasonality using **loess**

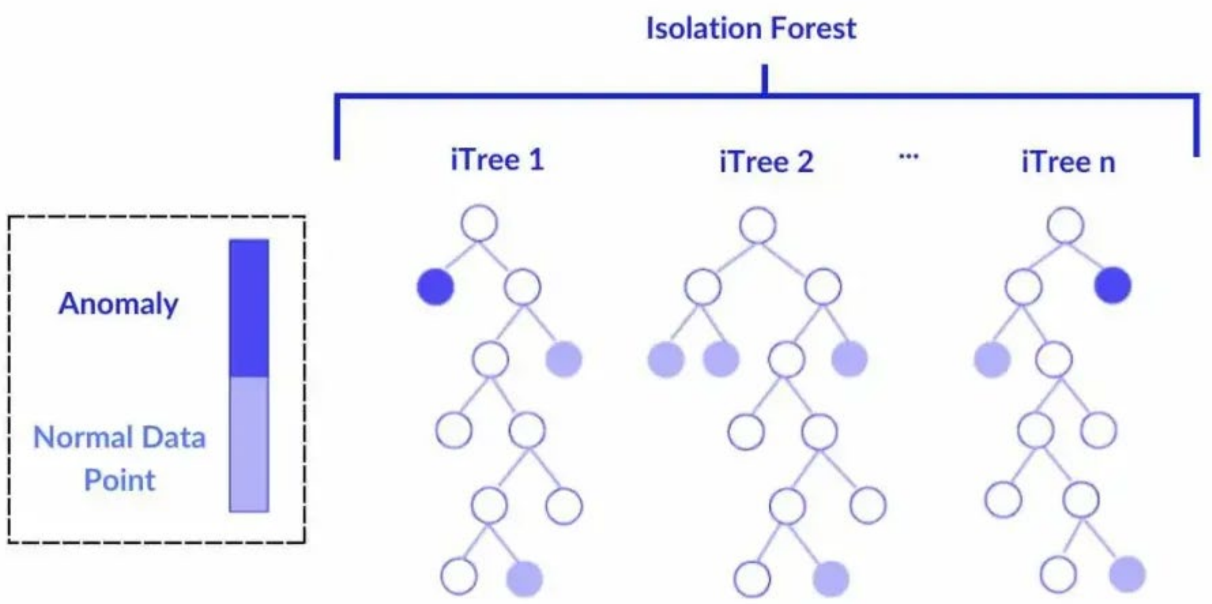
Compare seasonality estimates to determine if seasonality changing over time

- Parametric pairs of **Fourier** series (with periods 12 and 6 months)

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

- Identify rare events in residuals ε_t

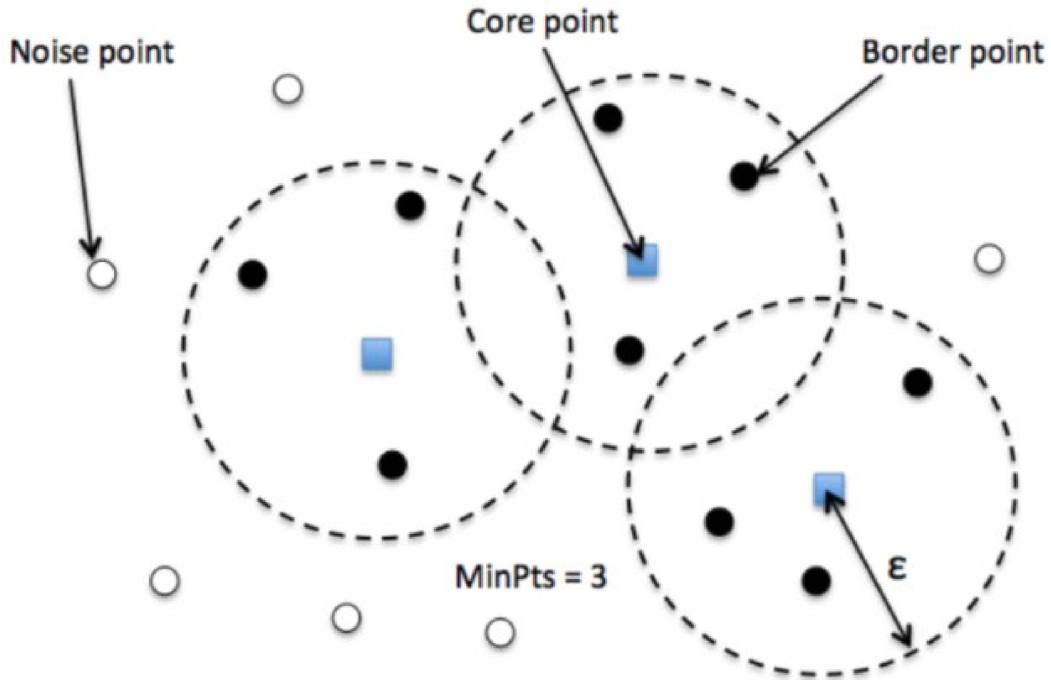
Methods: Rare event ID with iForest



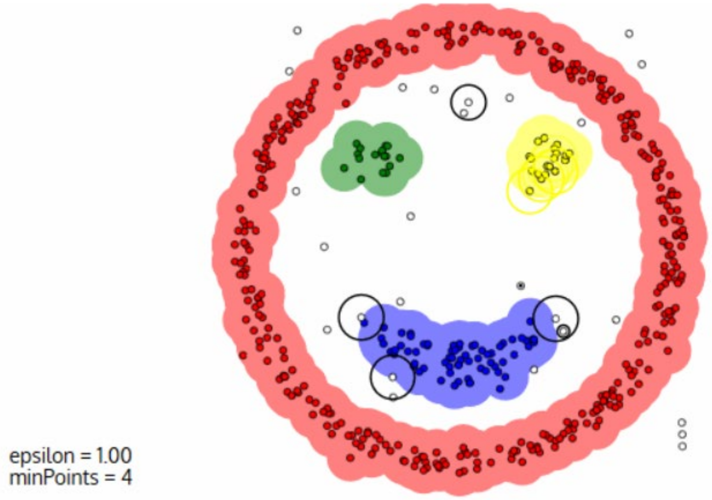
https://spotintelligence.com/2024/05/21/isolation-forest/

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

Methods: Rare event ID with 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/

Use the noise points

Methods: Example CBEFS

Example of CBEFS water condition estimates: surface & bottom (April – June)

-Salinity

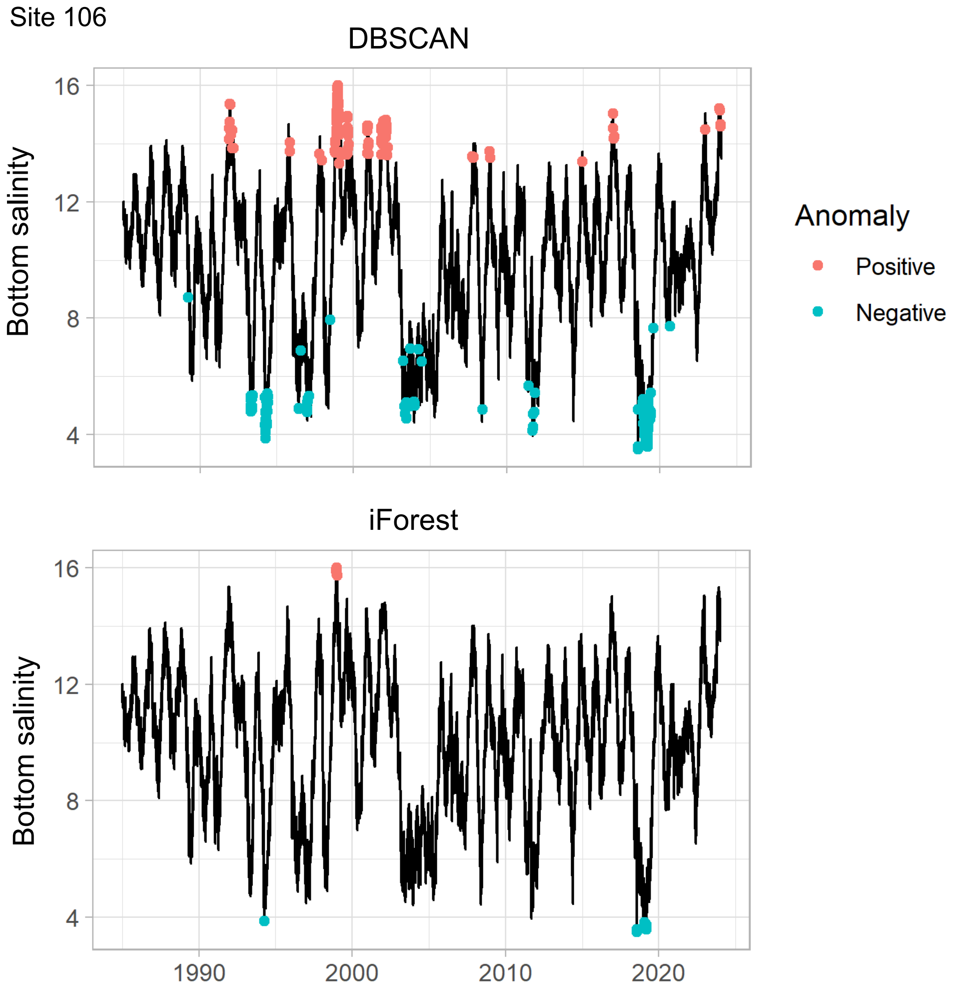
-Temperature

-pH

-DO

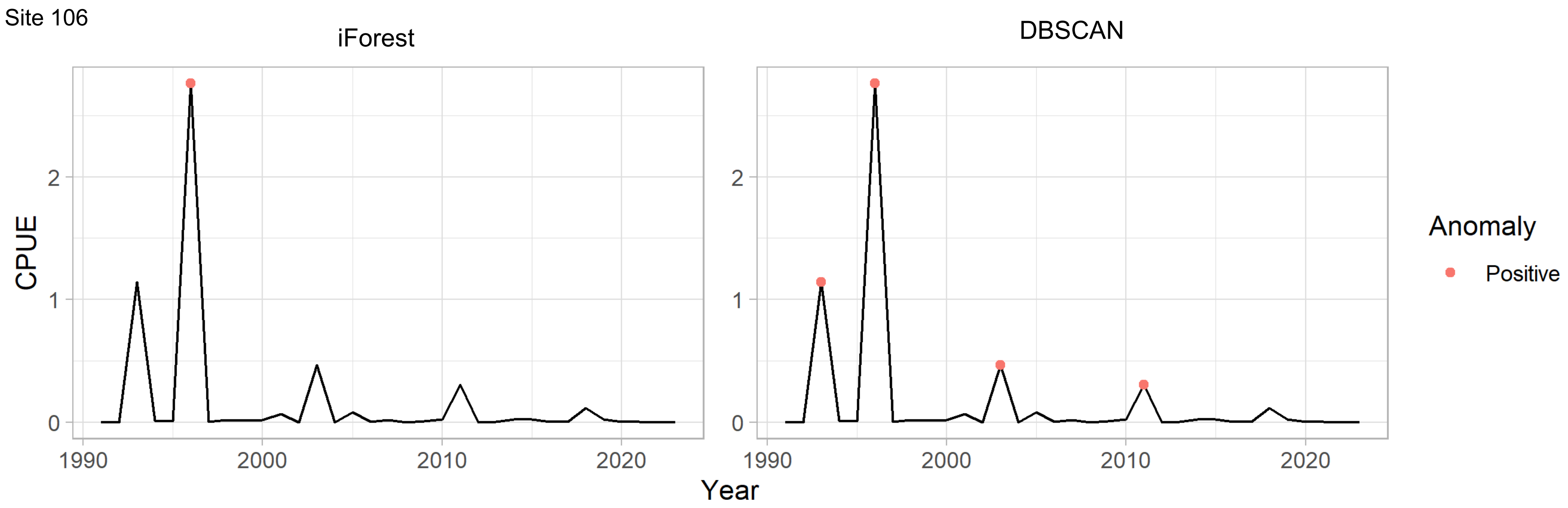
Surface:

-Diffuse attenuation



Methods: Example CPUE

- Same approach applied to fish survey CPUE



Methods: Identify linkages

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

- X(April) + X(May) + X(June) ~ Y

- Identify significant relationships between rare fish events and rare weather events detected by iForest using Pearson's chi-squared test

- Quantify proportion of significant linked rare fish ~ environmental events Baywide

Results – example Station 106

Environmental variable	Crossing of rare events: Fisheries x Environmental	
	Positive x Positive	Positive x Negative
Diffuse attenuation	-	-
Bottom DO	-	-
Bottom pH	**0.001**	-
Bottom salinity	-	0.858
Bottom temperature	-	-

Identified significant relationship between positive pH anomalies and striped bass CPUE rare events at this station

Results – Baywide summary

Proportion of statistically significant chi-squared test results across 49 seine survey stations in MD and VA

Environmental variable	Crossing of rare events: Fisheries x **Surface** Environmental		Crossing of rare events: Fisheries x **Bottom** Environmental	
	Positive x Positive	Positive x Negative	Positive x Positive	Positive x Negative
Diffuse attenuation	**0.12**	0.02	x	x
DO	**0.06**	**0.04**	0.02	0.00
pH	**0.04**	**0.06**	**0.06**	0.00
Salinity	0.00	0.02	0.00	0.02
Temperature	0.00	0.00	0.00	0.00

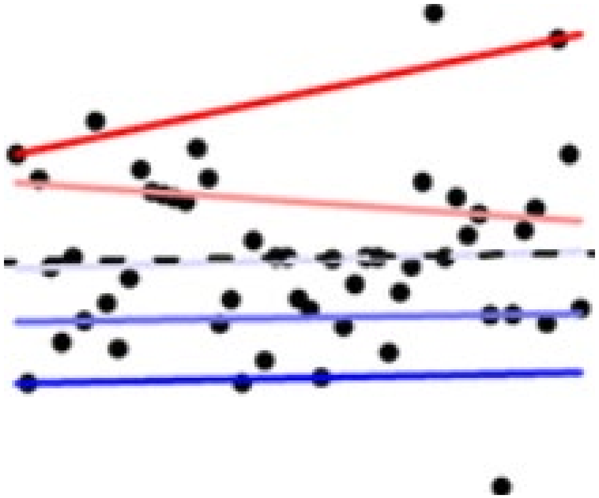
- Diffuse attenuation rare event most commonly associated with rare catch events for juvenile striped bass in seine surveys
- DO and pH also related to striped bass CPUE

Methods: Quantile modeling for rare events (baywide)

1) Generalized additive model for location, scale, and shape (GAMLSS) is a parametric model for the distribution of CPUE. Specifically, use a delta model, where presence of catches and mean (location parameter of the distribution) are related to Site ID and environmental parameters, and scale (variability) differs by Site ID. We use generalized gamma (GG) as the underlying distribution of CPUE.

2) Quantile regression (QR) is a linear modeling technique that estimates conditional quantiles (e.g., q5, q50 = median, q95) of the response variable, rather than the conditional mean as in the ordinary least squares regression. By targeting low or high quantiles, QR reveals relationships relevant to extreme or rare outcomes.

3) Quantile random forest (QRF) extends the traditional random forest by estimating conditional quantiles of the response variable, not just the mean prediction.



https://www.researchgate.net/figure/Quantile-regression-examples-for-winter-precipitation-at-station-a-168-880-in-Westwold_fig4_325068312

Methods: Variables

We explored relationships between annual mean CPUE (response variable) and 81 potential predictors from the CBEFS dataset:

min	DO_surface	April
	DO_bottom	
max	Salinity_surface	May
	Salinity_bottom	
avg	Temperature_surface	June
	Temperature_bottom	
	pH_surface	
	pH_bottom	
	Diffuse attenuation	

along with Year and SiteID.

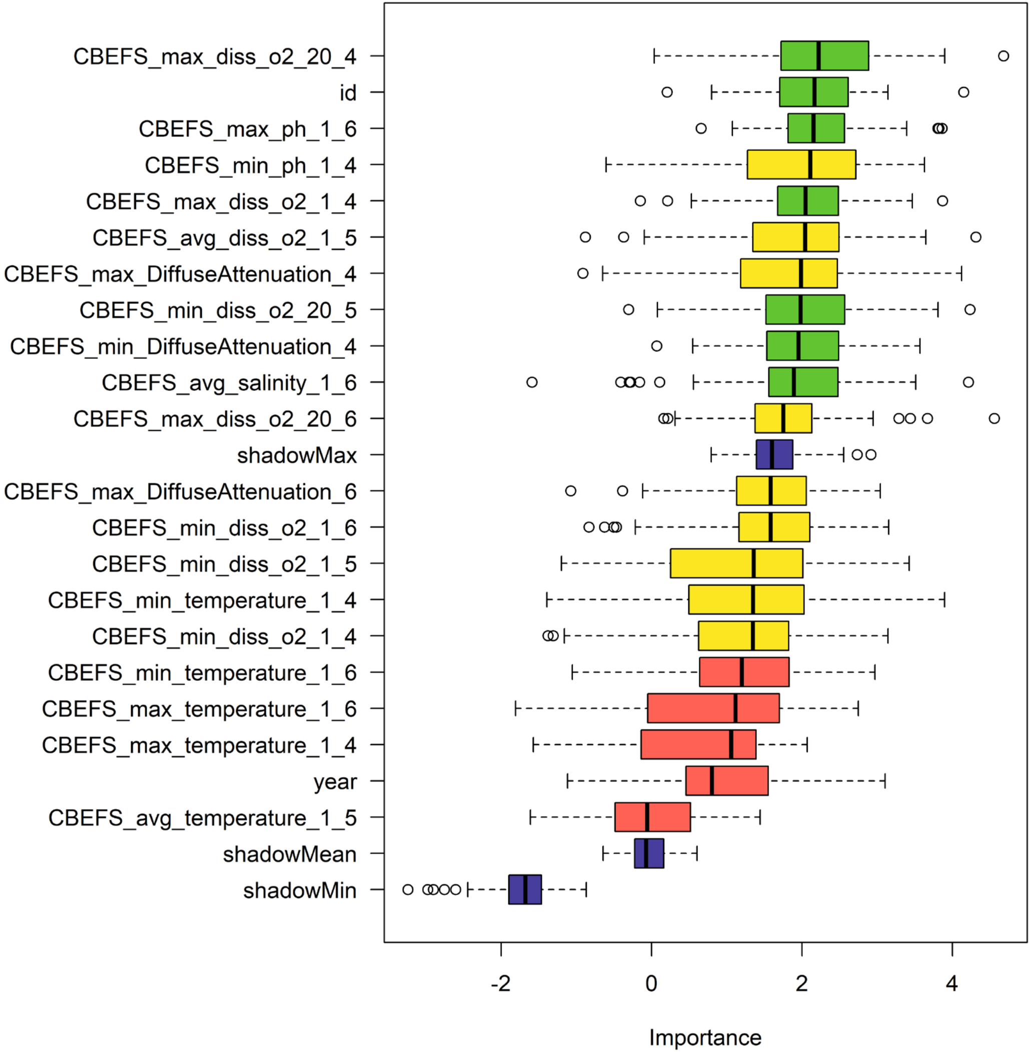
* In further results, surface is coded as 1, bottom = 20; April = 4, May = 5, June = 6

Methods: Variable selection

Variable selection 1 (V1): Remove collinear predictors (with pairwise correlation > 0.7). This removes duplicated information and improves model interpretability.

Variable selection 2 (V2): From V1, select important predictors of CPUE based on Boruta algorithm for selecting predictors in random forest

Dimensionality reduction: We additionally applied principal component analysis (PCA) to each set of predictors and selected the number of components in a data-driven way to represent the bulk of variability with fewer variables.



Methods: Quantile modeling cross-validation

We used a data-driven approach (10-fold cross-validation) to select among different methods (GAMLSS, QR, QRF), variable sets (V1 or V2), and variable processing (raw variables of PCA-transformed).

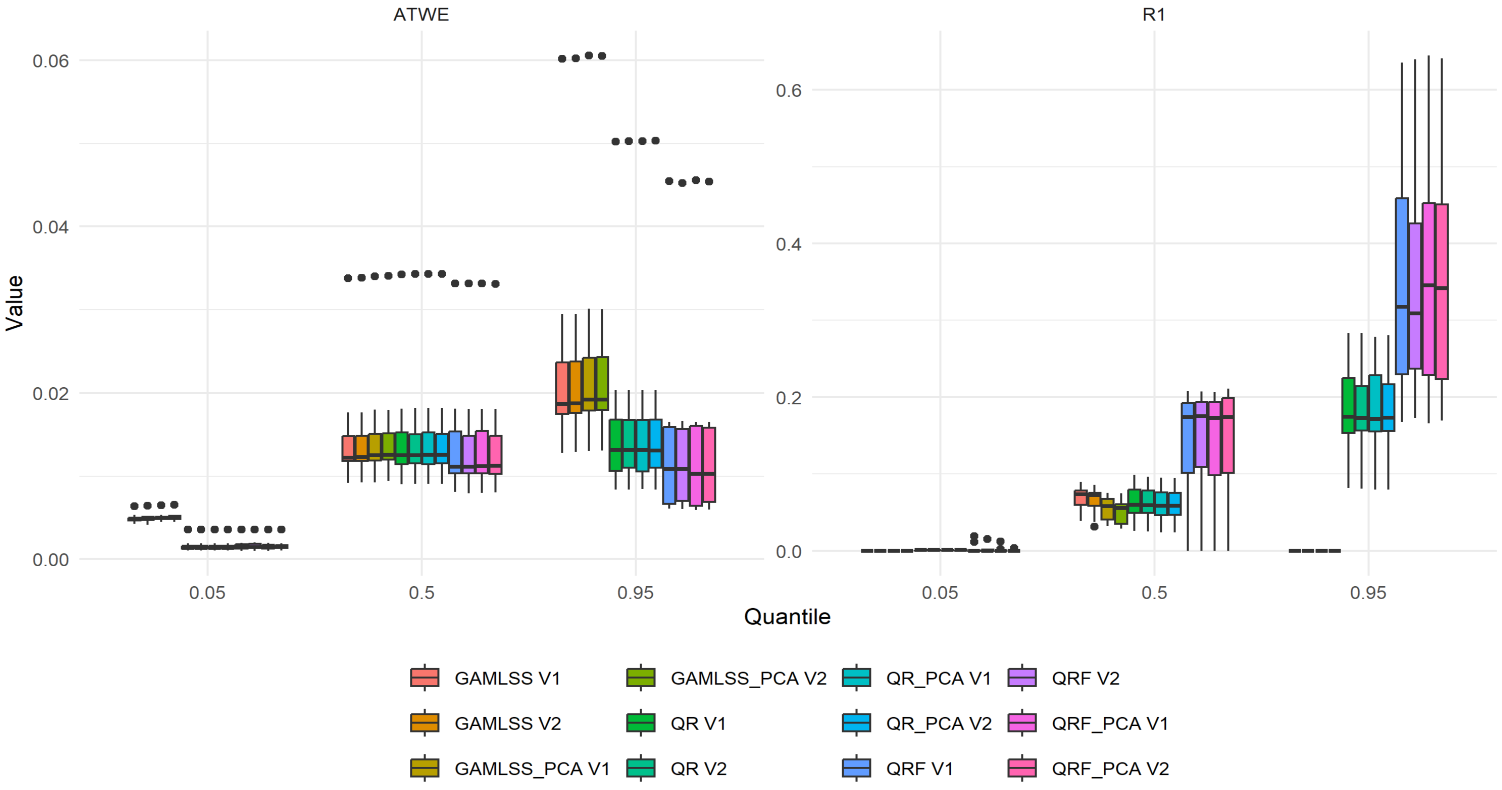
Performance metrics:

- ATWE (average quantile-weighted absolute error); smaller is better
- R^1 (analogue of coefficient of determination R^2) higher is better

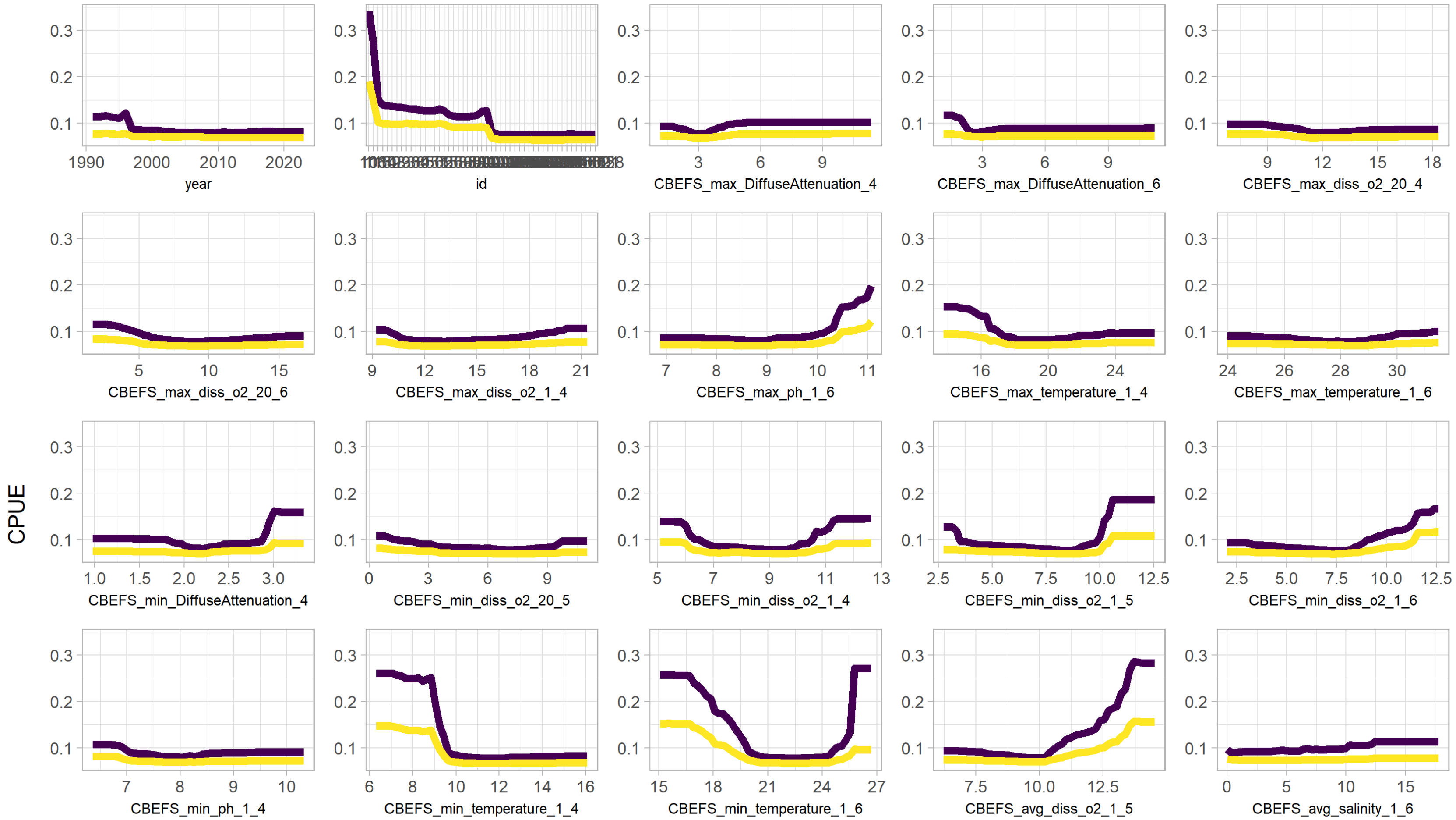
Image: Sontakke et al. (2019) DOI:10.1007/978-3-030-01057-7_35

Results: Quantile modeling cross-validation

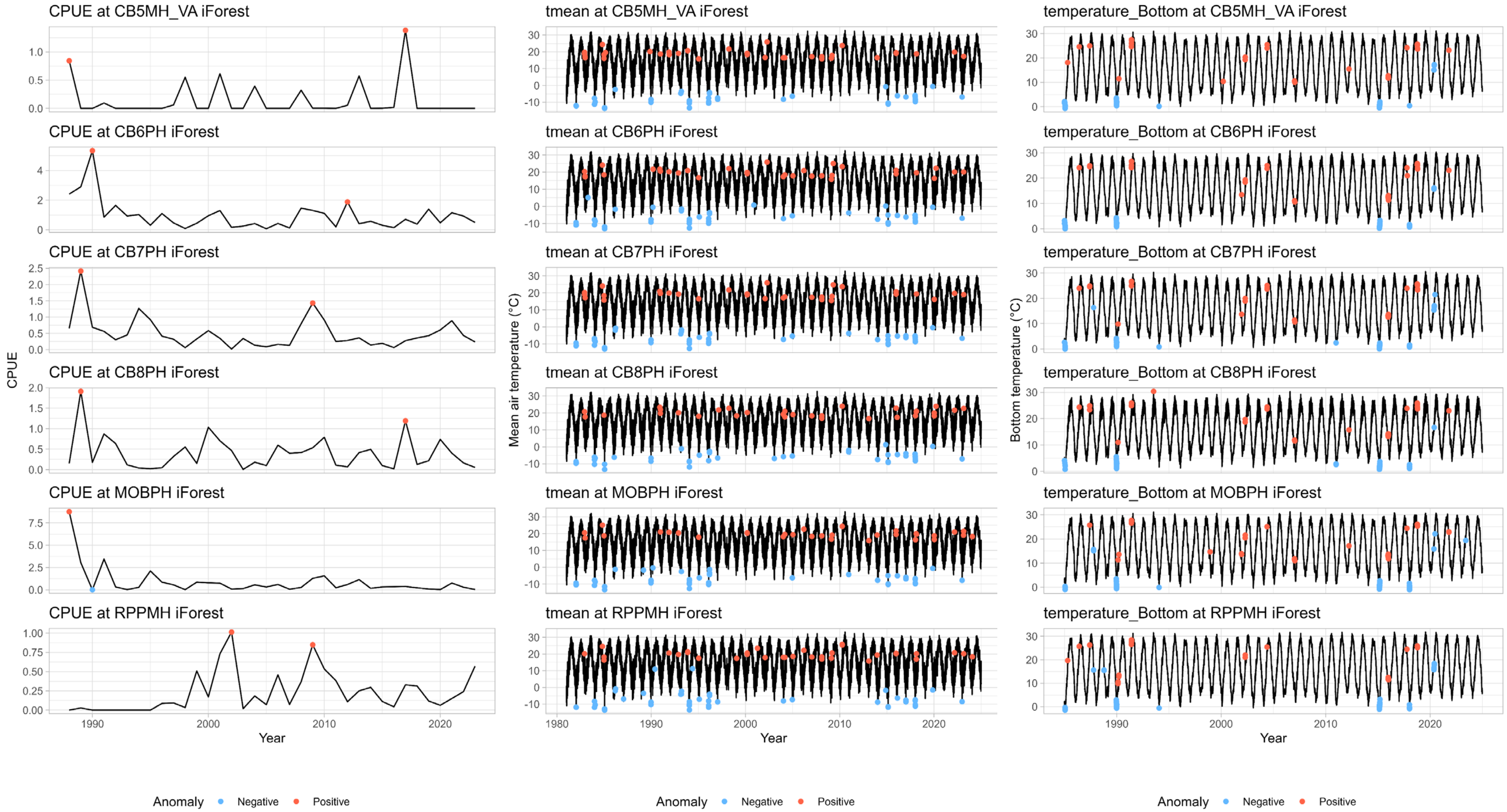
QRF consistently delivers lower errors (ATWE) and higher R^1,
PCA processing trivially improves the results, however,
V1 vs V2 variable sets do not lead to substantially different performance.



Results: Quantile random forest estimated effects



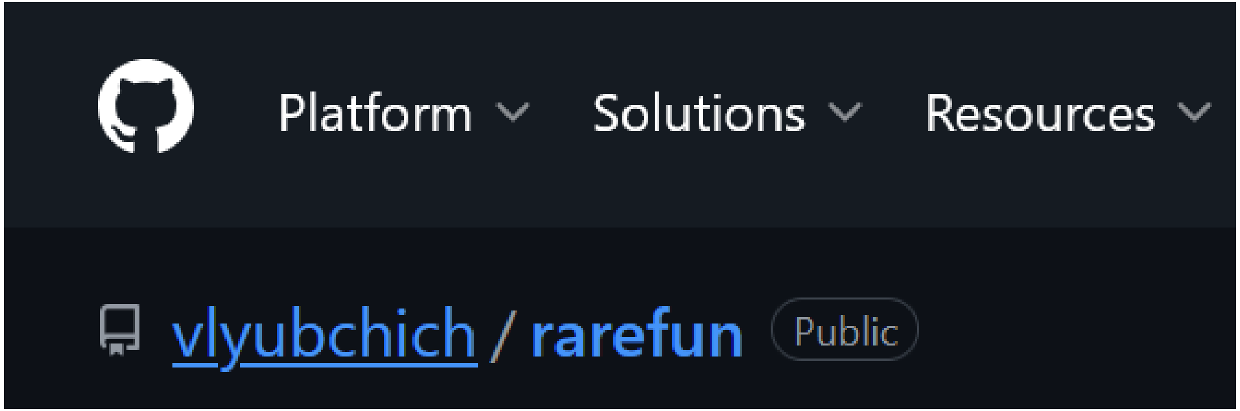
Next Steps: Regional Analysis



R package: rarefun

Source : github.com/vlyubchich/rarefun

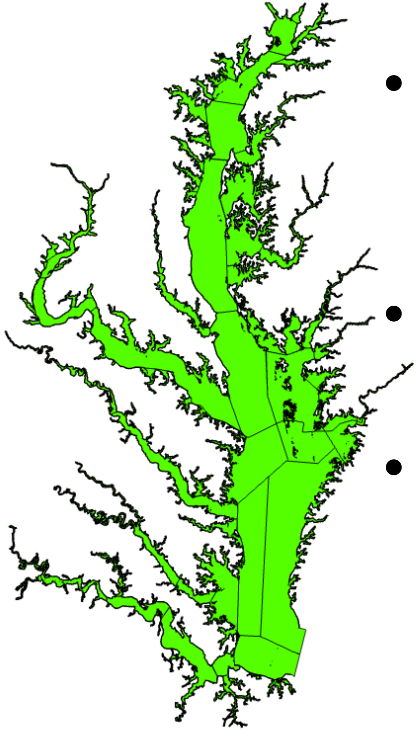
Vignette : Getting Started with rarefun



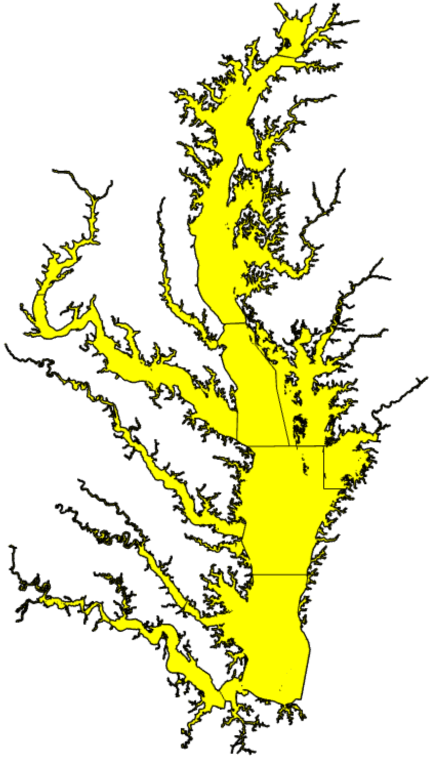
Overview

- Data and analysis approach
- Methods development
- Example results
- Next steps

Challenges and caveats

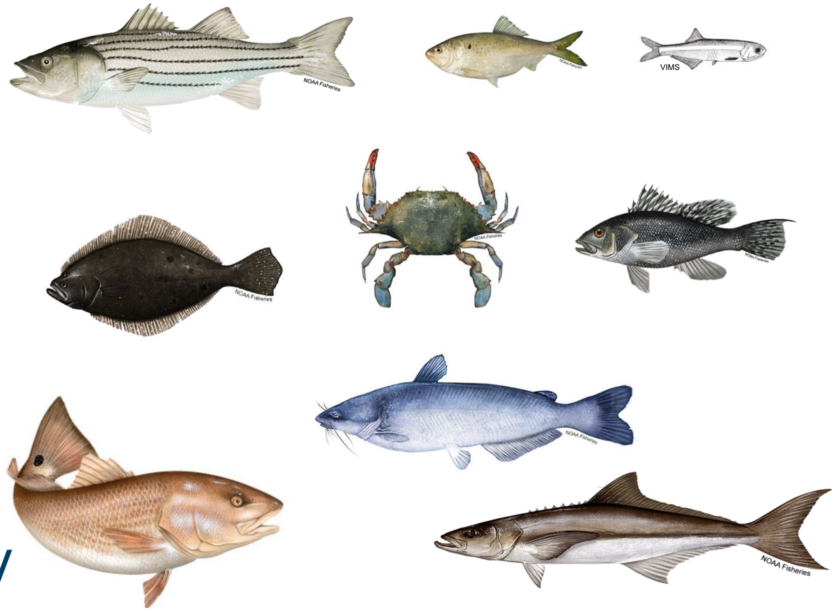


- Rare events in fish survey CPUE mostly positive anomalies
- Temporal resolution
  – lagging back prior 4 seasons
  – length of time series
- Spatial resolution of analyses
  – fine vs coarse region definitions
- Future data updates



Last steps

- Finalize analyses
- Explore sensitivity of approach to spatial resolution, test with known time series of rare events
- Compare with literature; generate hypotheses
- Identify suite of potential indicators for ecosystem monitoring
- Communicate results to fisheries/Bay managers
- Integrate results and **rarefun** tool into NCBO and other local EBFM efforts



Thank you!

NOAA FISHERIES

MARYLAND DEPARTMENT OF NATURAL RESOURCES

VIMS | WILLIAM & MARY

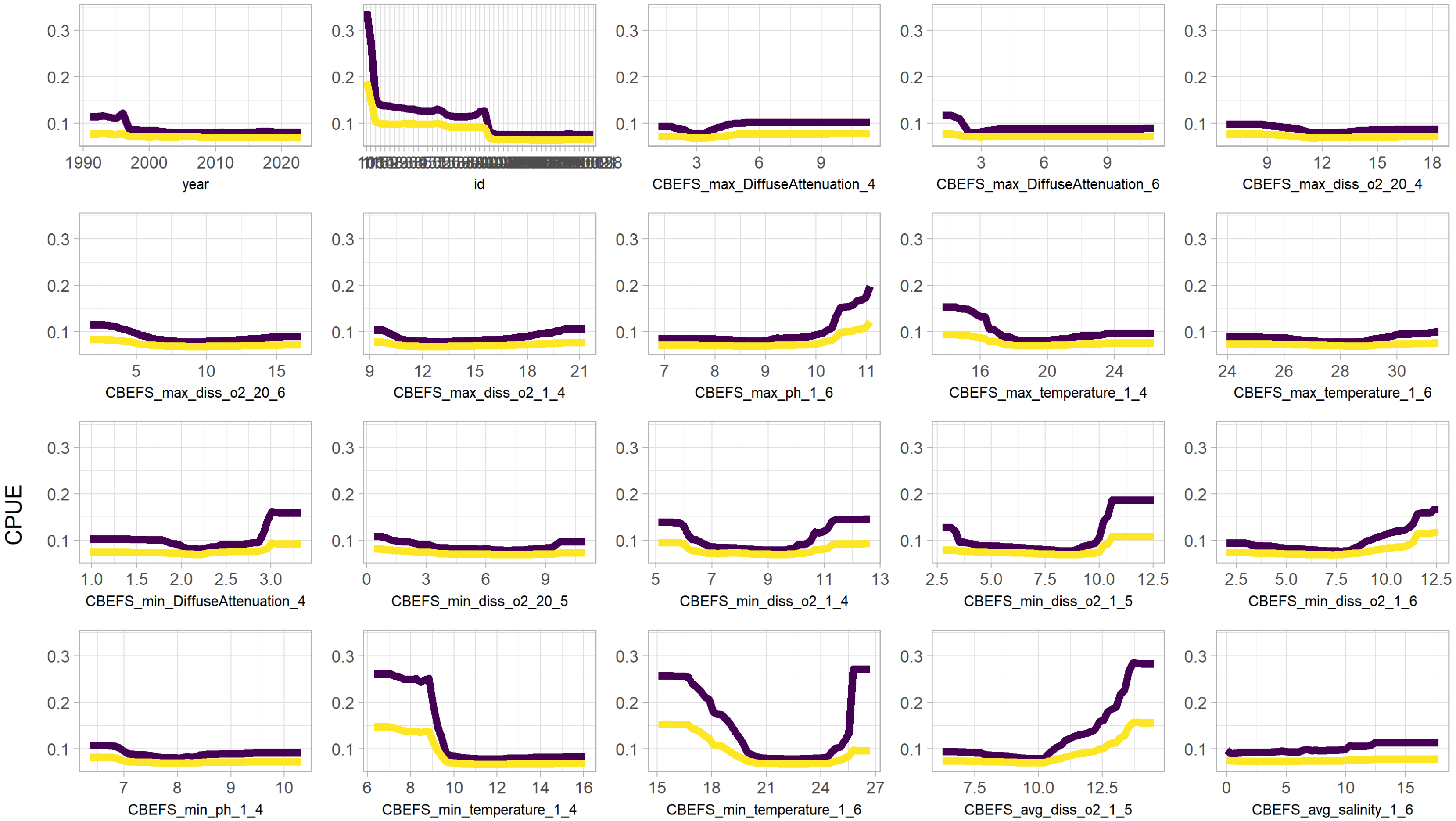
VIRGINIA INSTITUTE OF MARINE SCIENCE

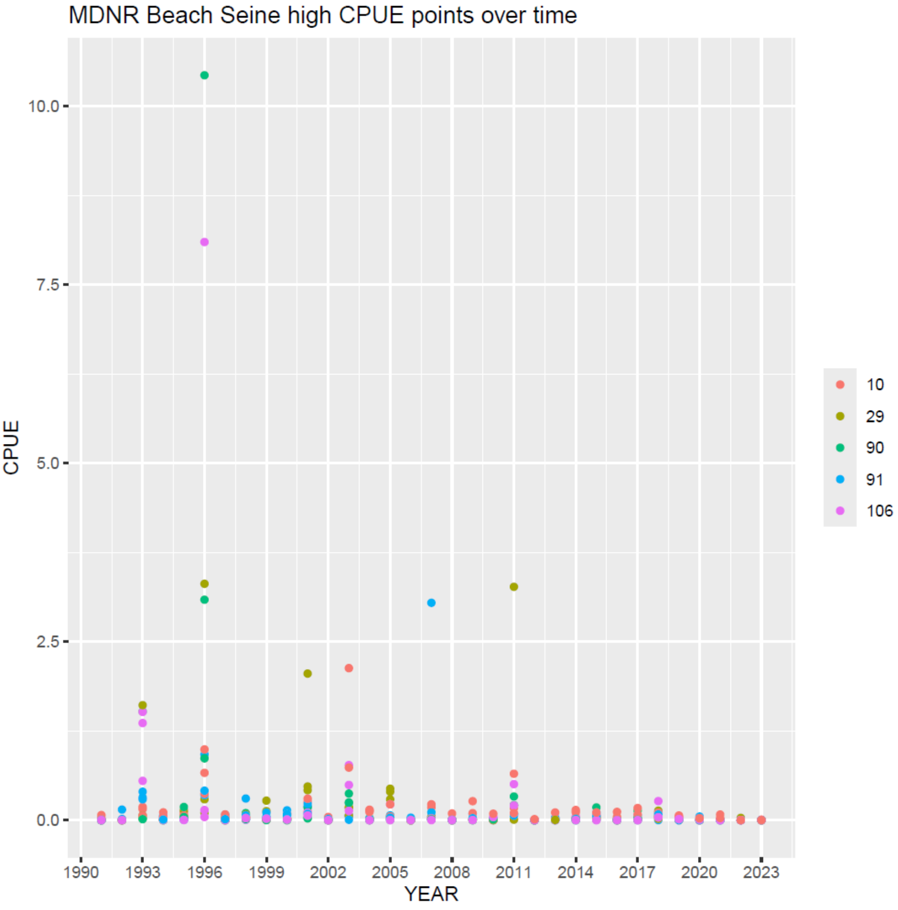
Questions?

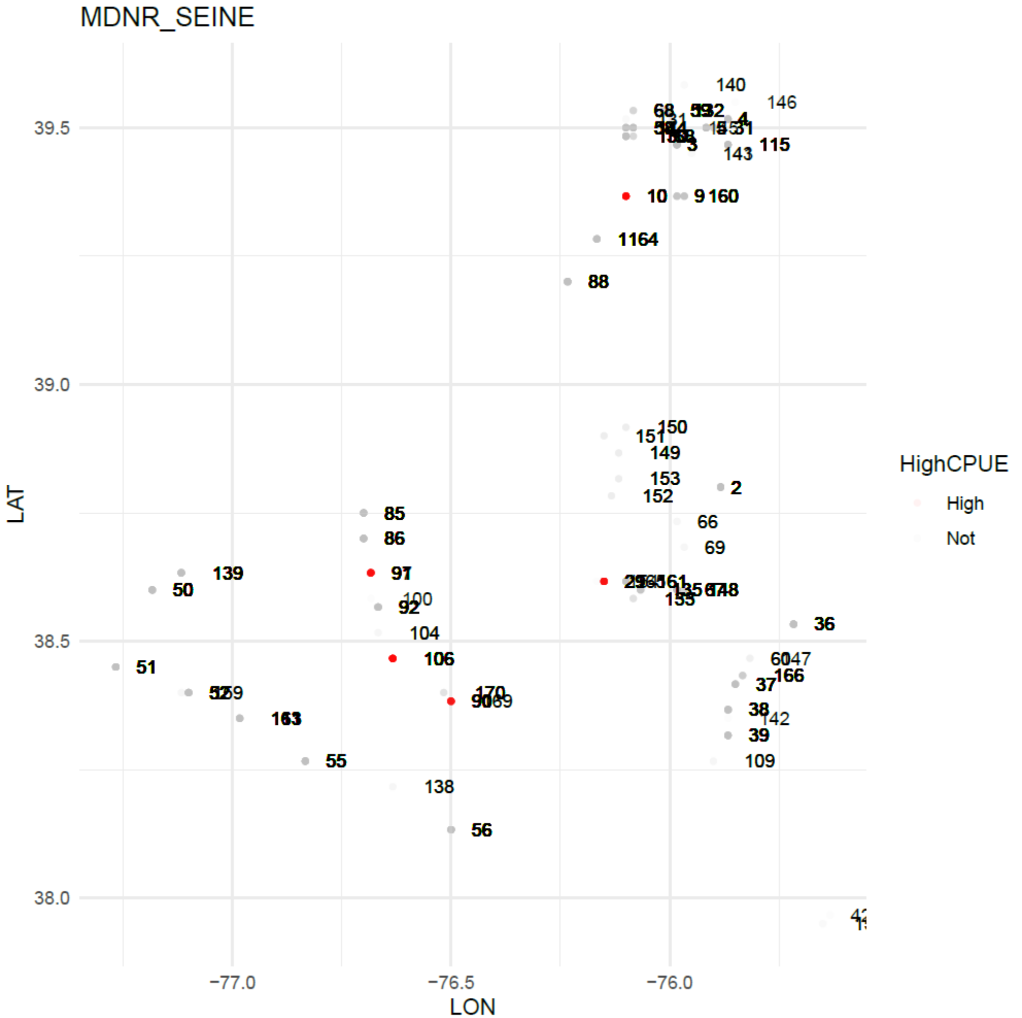
Genny Nesslage – nesslage@umces.edu
Slava Lyubchich – lyubchich@umces.edu

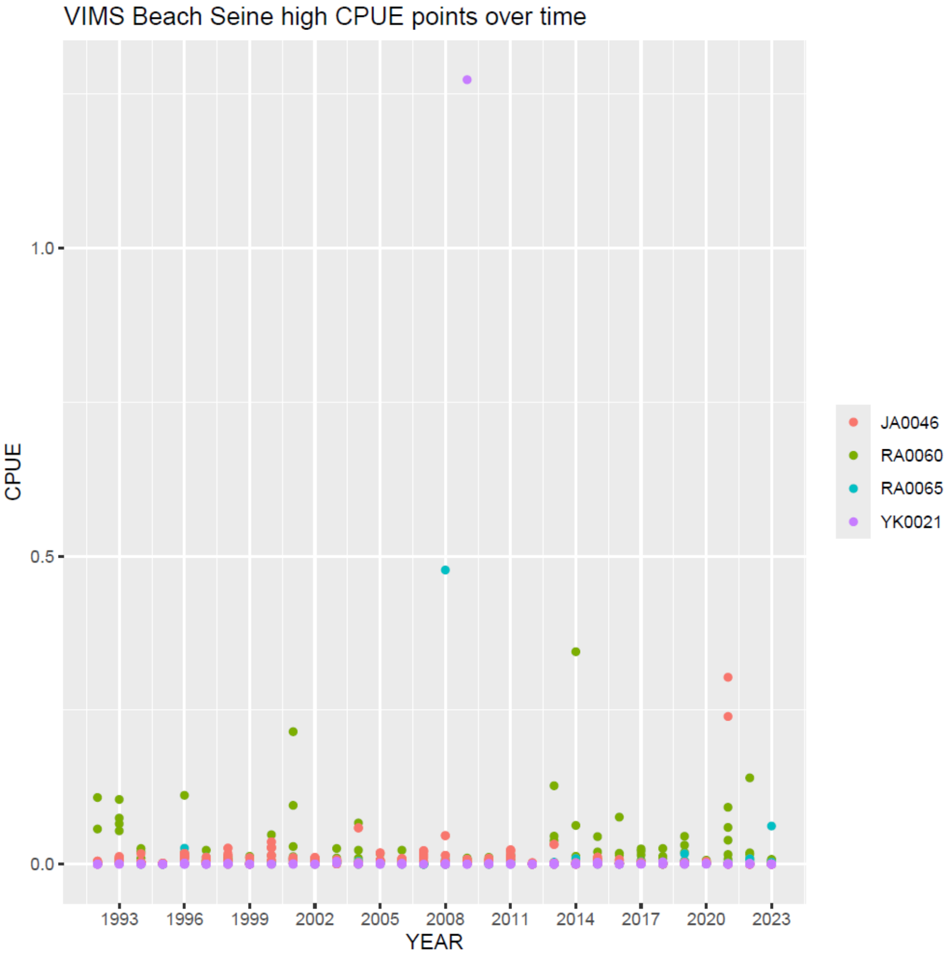
University of Maryland
CENTER FOR ENVIRONMENTAL SCIENCE
CHESAPEAKE BIOLOGICAL LABORATORY

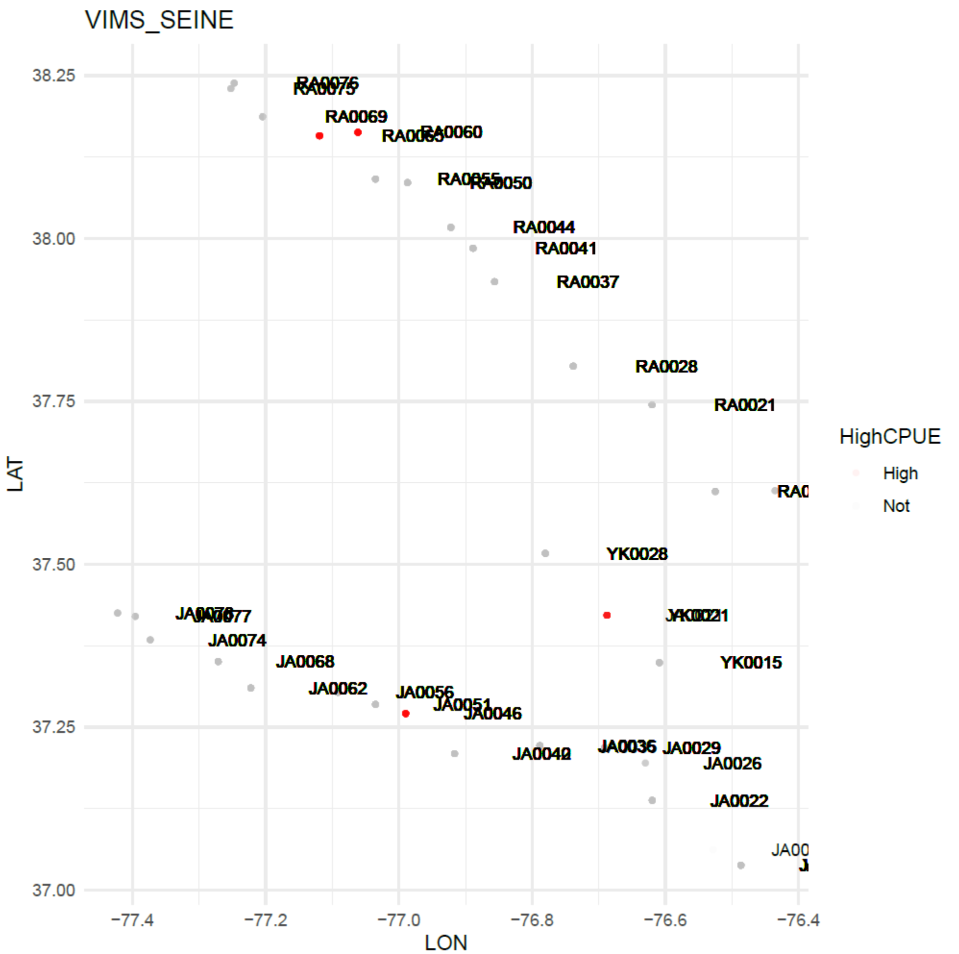
Results: Quantile random forest estimated effects




MDNR Beach Seine high CPUE points over time
CPUE
YEAR
10.0
7.5
5.0
2.5
0.0
1990
1993
1996
1999
2002
2005
2008
2011
2014
2017
2020
2023
10
29
90
91
106


MDNR_SEINE
LAT
LON
39.5
39.0
38.5
38.0
-77.0
-76.5
-76.0
HighCPUE
High
Not
140
146
115
143
10
9
160
1164
88
150
151
149
153
152
2
66
69
85
86
139
50
97
100
92
104
106
36
51
170
163
55
138
56
166
37
38
142
39
109


VIMS Beach Seine high CPUE points over time
CPUE
YEAR
0.0
0.5
1.0
1993
1996
1999
2002
2005
2008
2011
2014
2017
2020
2023
JA0046
RA0060
RA0065
YK0021


VIMS_SEINE
LAT
LON
38.25
38.00
37.75
37.50
37.25
37.00
−77.4
−77.2
−77.0
−76.8
−76.6
−76.4
HighCPUE
High
Not
RA0069
RA0060
RA0044
RA0041
RA0037
RA0028
RA0021
YK0028
JA0074
JA0068
JA0062
JA0056
JA0046
YK0015
JA0029
JA0026
JA0022

Rare: Infrequently occurring; uncommon, extraordinary

Extreme: Being far beyond the norm; most remote in any direction; outermost or farthest; Being in or attaining the greatest or highest degree; very intense