5-31-2019

Extracting Signal from the Noisy Environment of an Ecosystem

Emily Wefelmeyer
Harrisburg University of Science and Technology

Pranita Pramod Patil
Harrisburg University of Science and Technology

Sridhar Reddy Ravula
Harrisburg University of Science and Technology

Kevin M. Purcell
Harrisburg University of Science and Technology

Ziyuan Huang
Harrisburg University of Science and Technology

See next page for additional authors

Authors
Emily Wefelmeyer, Pranita Pramod Patil, Sridhar Reddy Ravula, Kevin M. Purcell, Ziyuan Huang, and Igor Pilja

Extracting Signal from the Noisy Environment of an Ecosystem

Emily Wefelmeyer, Pranita Patil, Sridhar Ravula, Kevin Purcell, Ziyuan Huang, & Igor Pilja

Harrisburg University of Science & Technology

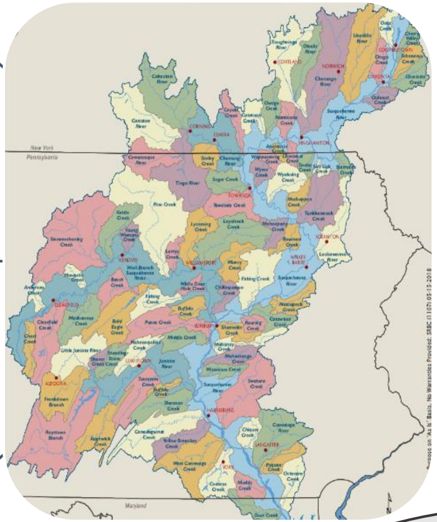
River Basin Management: Signal Vs Noise

River basins are complex systems.

Geological
- Rock- Soil
- Terrain
- Flow

Weather
- Precipitation
- Evaporation
- Snow

People
- Mining
- Withdrawal
- Effluents



SRBC
SUSQUEHANNA RIVER BASIN COMMISSION
NY ▪ PA ▪ MD ▪ USA

Stakeholder
- MD
- PA
- NY

Consumers
- Residents
- Farming Community
- Natural Gas Industry
- Other

Objectives
- Support Growth
- Sustain Environment

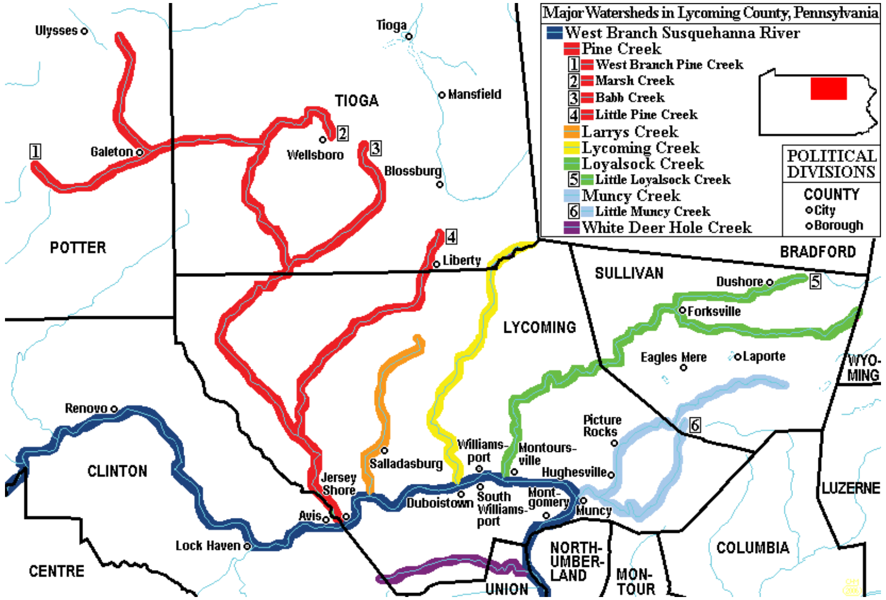
What “signal” should SRBC focus on in achieving the objectives?

Regulators need to identify “Signal” from noisy data

Signal Extraction and Definition

- Volume and variety of data
- Data scientist was sought
- How to systematically define and extract signal from noise
  - Lack of guiding framework
  - Limited tool set
- Explored available tools and literature
- Signal desired attributes
  - Clearly definable
  - Must be able to detect
  - Ability to explain movement in response variable with help of explanatory variables.
  - Has an effect on environment or is an indicator of change in environment
- Rational approach to variable selection
  - Maximize the information obtained from the analysis of variable
    - Linked to a large set of spatial, temporal environmental variables.

Available Data



Initial scope limited to “Pine Creek”

- Largest tributary of the West Branch Susquehanna River
- 87.2 miles (140.3 km) long
- Largest watershed of all the West Branch’s tributaries

File	*n*	Parameters	Start	End	Missing Data	Frequency/Periodicity
Water Quality	195,142	6	6/23/2011 12:00	12/31/2017 23:45	<0.15%	4 hours until 10/9/2014; 15 minutes since then
Chemistry	754	51	8/10/1983 00:00	3/5/2018 08:45	~70%	No noticeable pattern; 1983, 1994, 2002, & 2008 – 2018
Fish	52	59	9/09/2008 12:00	8/03/2017 09:45	~0%	No noticeable pattern
Macroinvertebrates	134	223	8/10/1983 00:00	10/17/2017 14:00	<0.001%	No noticeable pattern; 1983 – 2002 & 2008 – 2017
Biotic integrity community data	59	95	No dates given	Seems to be averages	~22%	n/a

Community Metrics

- Community metrics (Diversity) from observed species count
  - For fish and macroinvertebrates
  - Available for numerous sites
- Community (Diversity) metrics advantages
  - Good overall metrics of environmental quality
  - Environmental quality is a core standard by which SRBC has to accomplish its mission
- Multiple response variables computed for diversity

Metric	Formula	Remark
Margalef's species Richness	$S_{Marg,y} = \frac{S_y - 1}{\log F_y}$	Where S_y is species count and F_y is the total count of all individual fish caught
Pielou evenness	$J_y = \frac{-\sum_{s=1}^{S_y} N_{y,s}/N_y \log(N_{y,s}/N_y)}{\log S_y}$	Where N_y is abundance and $N_{y,s}$ is average density of species 's' (individuals km^{-2})
Hill's N1 Diversity	$N1_y = \exp\left(-\sum_{s=1}^{S_y} \frac{N_{y,s}}{N_y} \log \frac{N_{y,s}}{N_y}\right)$	N_y and $N_{y,s}$ defined as above
Hill's N2 dominance	$N2_y = \frac{1}{\sum_{s=1}^{S_y} (N_{y,s}/N_y)}$	S_y,N_y and $N_{y,s}$ defined as above

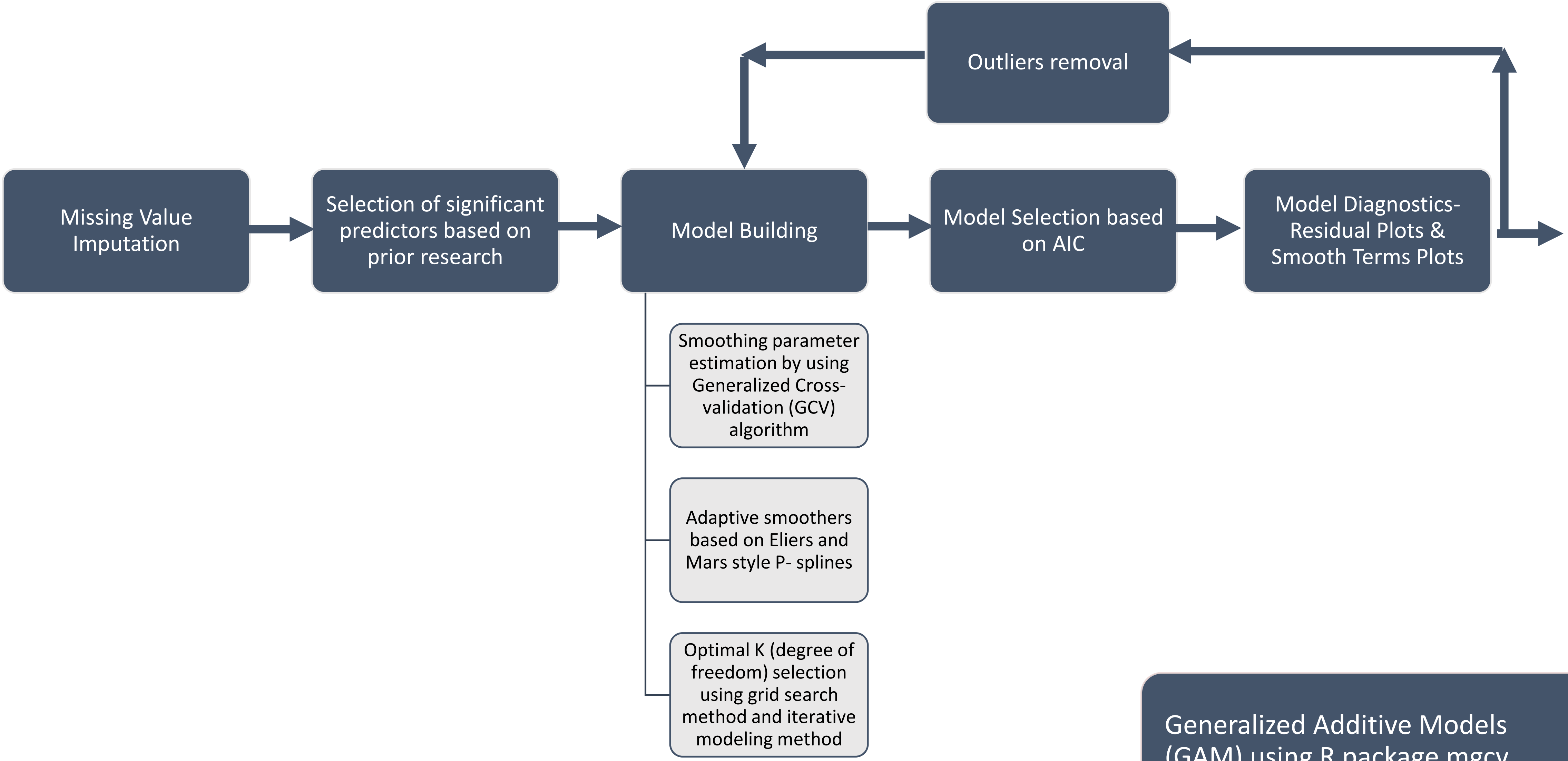
Metrics of composition, structure, and functioning of North Sea demersal fish community doi:10.1093/icesjms/fsr188 Greenstreet et al. 2012

Model Selection: Generalized Additive Model (GAM)

- More powerful than linear model due to inclusion of non-linear smoothers
- Parametric and non-parametric functions to explore linear and non-linear patterns
- Addition of smooth functions of covariates
- Smooth/basis function estimated from the data
- Mostly used when
  - Non-linear relationships
  - Distribution other than normal (response)
  - Need regularization(to avoid overfitting)
- Applications: air quality, ecology, medicine, genetics, molecular biology, etc.

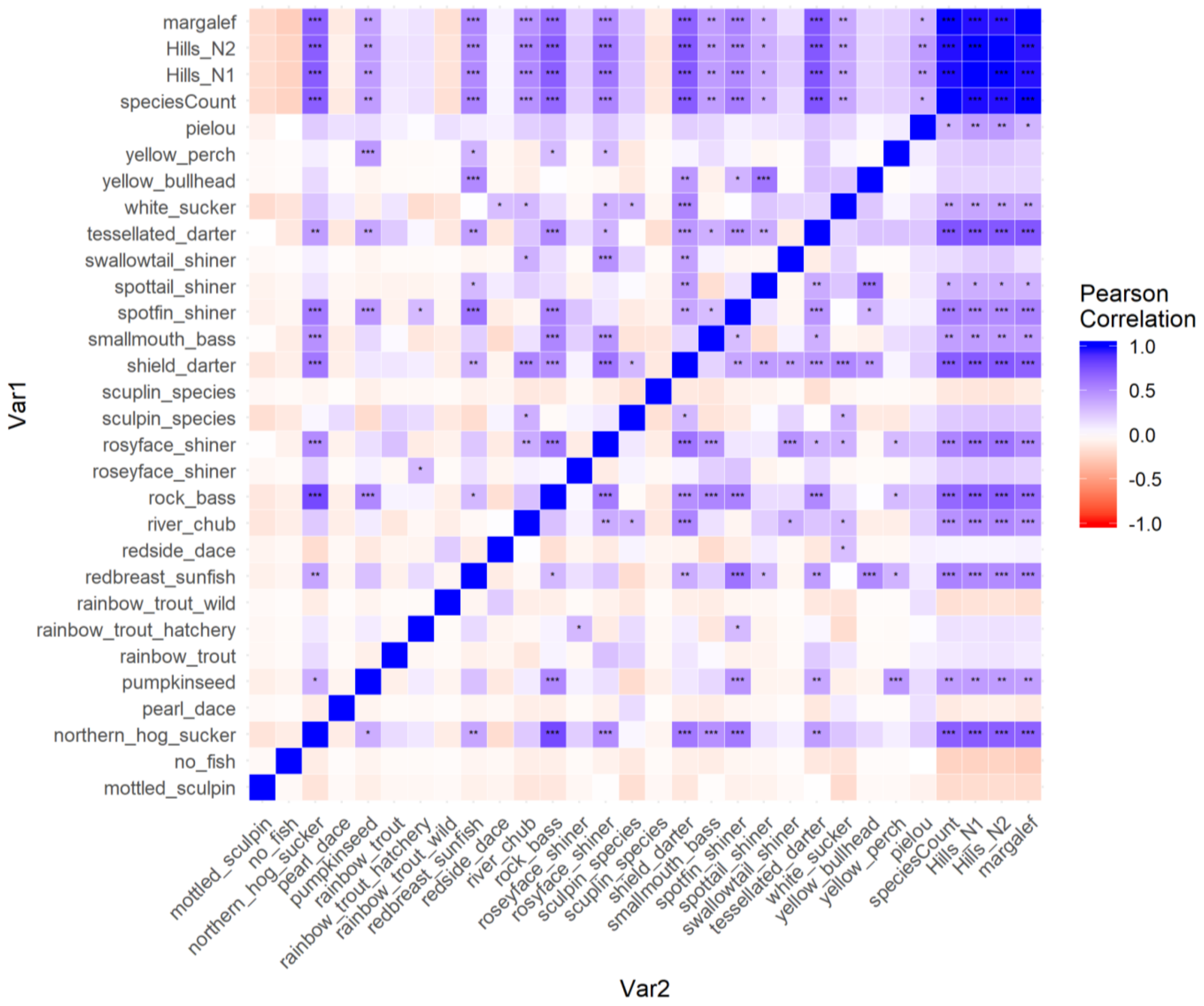
These factors made us chose GAM over other models

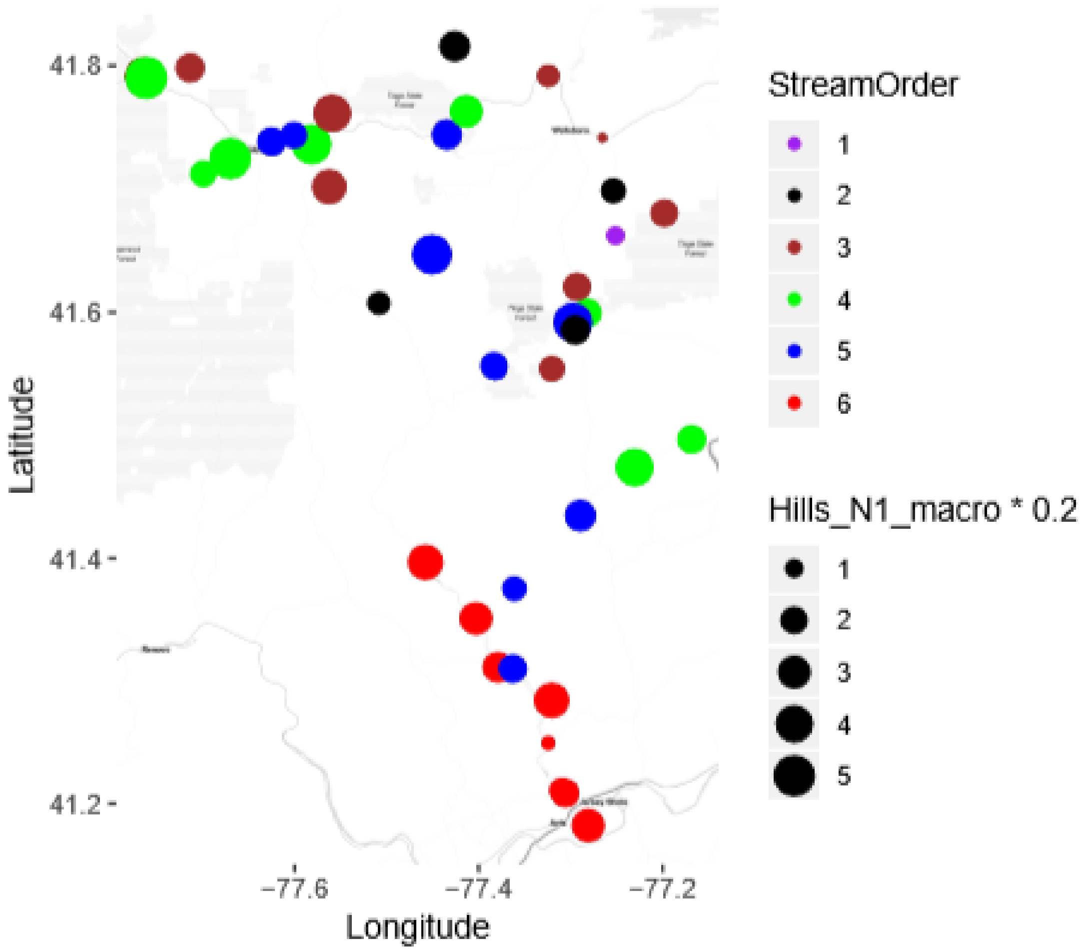
Modeling Steps



Generalized Additive Models (GAM) using R package mgcv

Model Inputs Selection

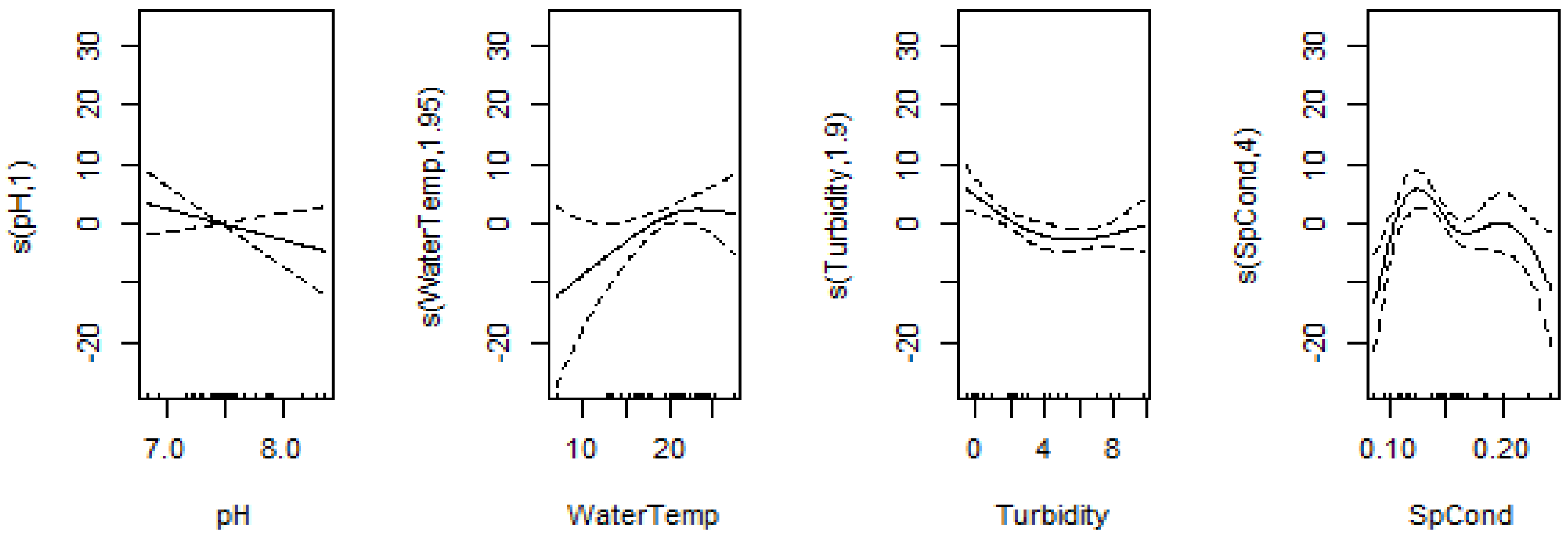




Hill’s N1 GAM Model: Fish

- Data
  - 50 observations
  - Repeated measurement at few stations
- Grid search for degrees of freedom (3 to 8)
- Best AIC: 288.94
- Deviance Explained:**74%**
- Outliers removed

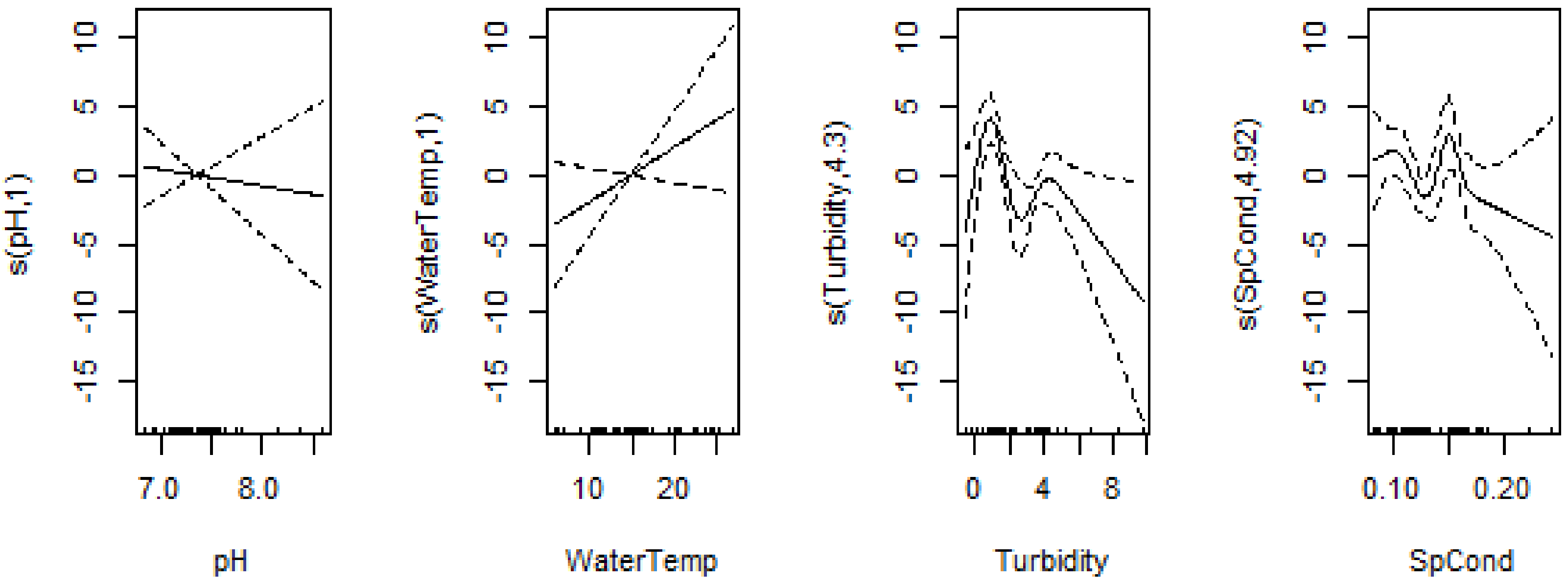
Model Smooth Functions Plots



Hill's N1 GAM Model: Macroinvertebrates

- Data
  - 115 observations
  - Repeated measurement at few stations
- Iterative modeling method for degrees of freedom
- Best AIC: 640.61
- Adaptive Smoothing Parameter
- Deviance Explained:**61.9%**
- Outliers removed

Model Smooth Function Plots



GAM Models Summary

- Extract significant information from noisy dataset
- Explained significant % of variance in Hill’s N1, Hill’s N2, and Margalef diversity metrics

Fish GAM Model

- 72% to 75% variance explained

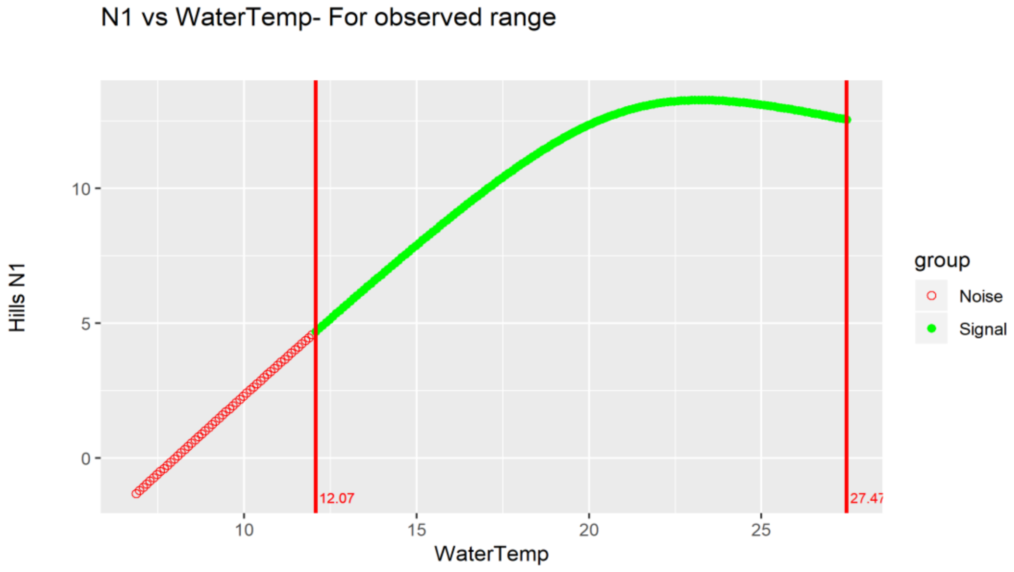
Diversity Metric	% Variance Explained	AIC
Hill’s N1	73.8	289
Hill’s N2	72.4	283
Pielou	54	-156
Margalef	74.9	111

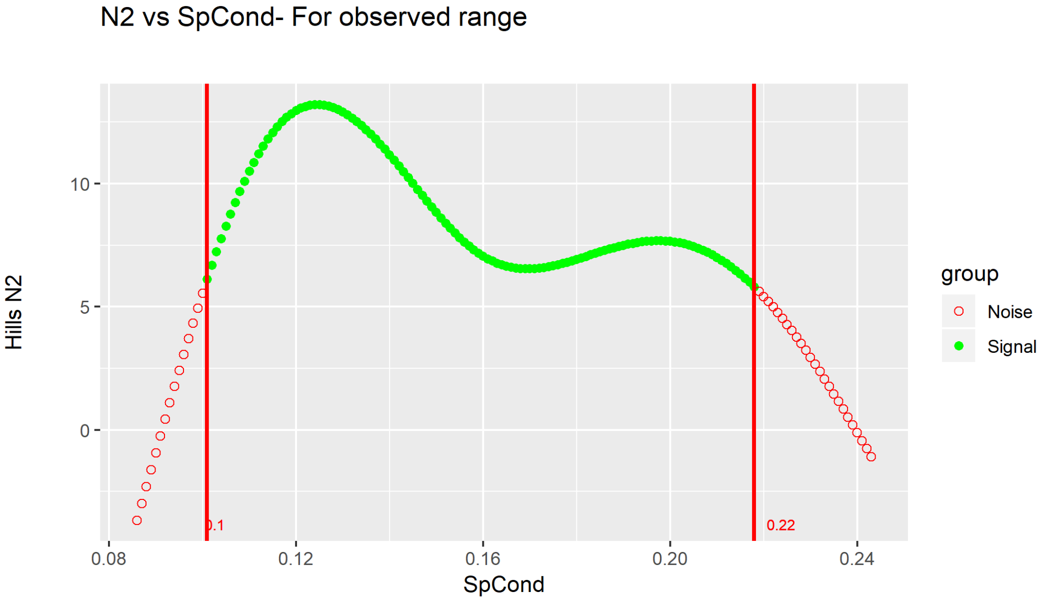
Macroinvertebrate GAM Models

- 54% to 66% variance explained

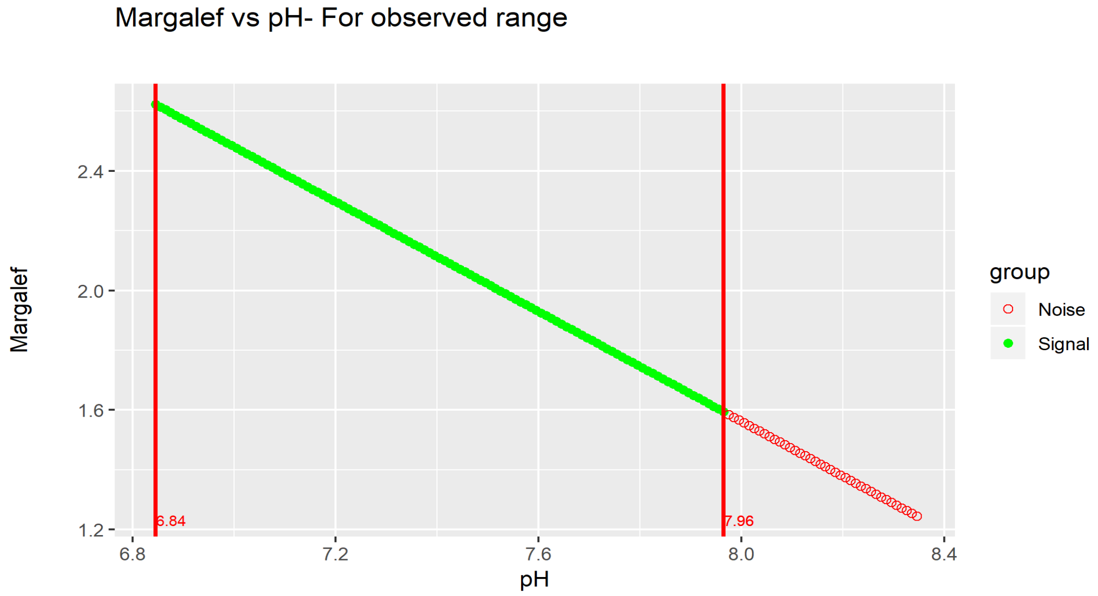
Diversity Metric	% Variance Explained	AIC
Hill's N1	61.9	641
Hill's N2	54.2	598
Pielou	45.5	-251
Margalef	65.9	296

Signal Prediction Analysis: Fish Diversity



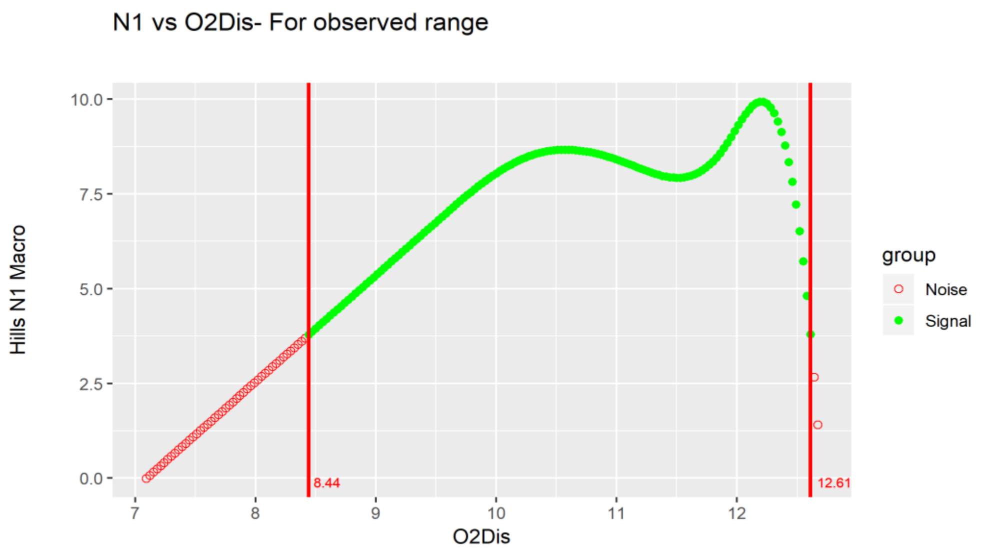


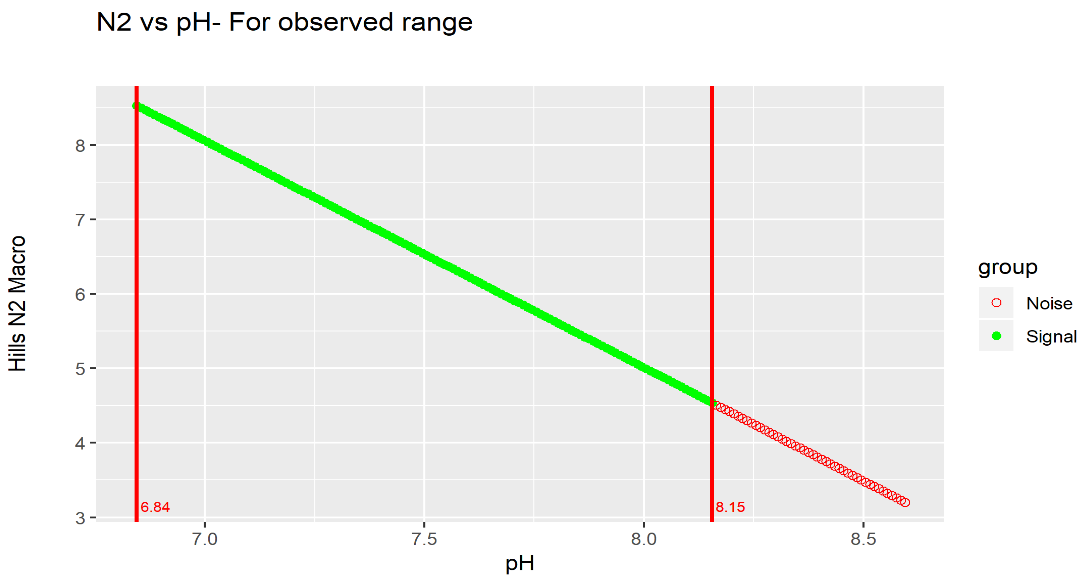
- Posterior simulations
- Change one environmental variable while keeping other variables constant
- Threshold calculations based on quantile method

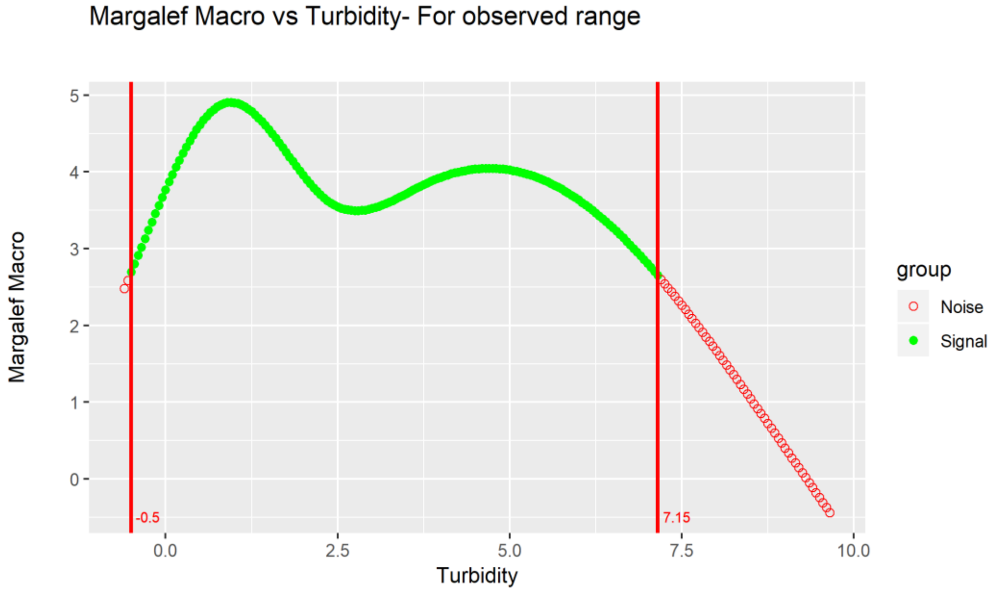


- Fish diversity deteriorates when
  - Dissolved oxygen < 8
  - Specific conductivity > 0.2

Signal Prediction Analysis: Macroinvertebrate Diversity







- Macroinvertebrate diversity deteriorates when
  - Dissolved oxygen < 8.5
  - Turbidity > 7

Our Findings

River basin monitoring

- A framework to synthesize multiple datasets
  - Enables regulators to prioritize and communicate
  - Can help stakeholder in understanding the impact of their actions
- Flexible: diversity metric can be replaced with another metric
- Dynamic dashboard application to monitor biotic response

Next Steps

Refine synthesis in collaboration with SRBC

Action Threshold Identification

Incorporate Variable Interactions

Adding More Information

Questions?

Special thanks to the following people/organizations:

John Quigley with the Center for Environment, Energy and Economy at Harrisburg University for the opportunity.

SRBC for the data and resources.

Dr. Kevin Purcell for his help and support through the project.

Contact Information

Emily M. Wefelmeyer (emwefelmeyer@my.harrisburgu.edu)
Pranita P. Patil (pppatil@my.harrisburgu.edu)
Sridhar Ravula (sravula@my.harrisburgu.edu)
Kevin M. Purcell (kpurcell@harrisburgu.edu)
Ziyuan Huang (zhuang@my.harrisburgu.edu)
Igor Pilja (ipilja@my.harrisburgu.edu)