Physical habitat is more than a sediment issue:

A multi-dimensional habitat assessment indicates new approaches for river management

STAC Workshop: Leveraging AI and ML

24 February 2025

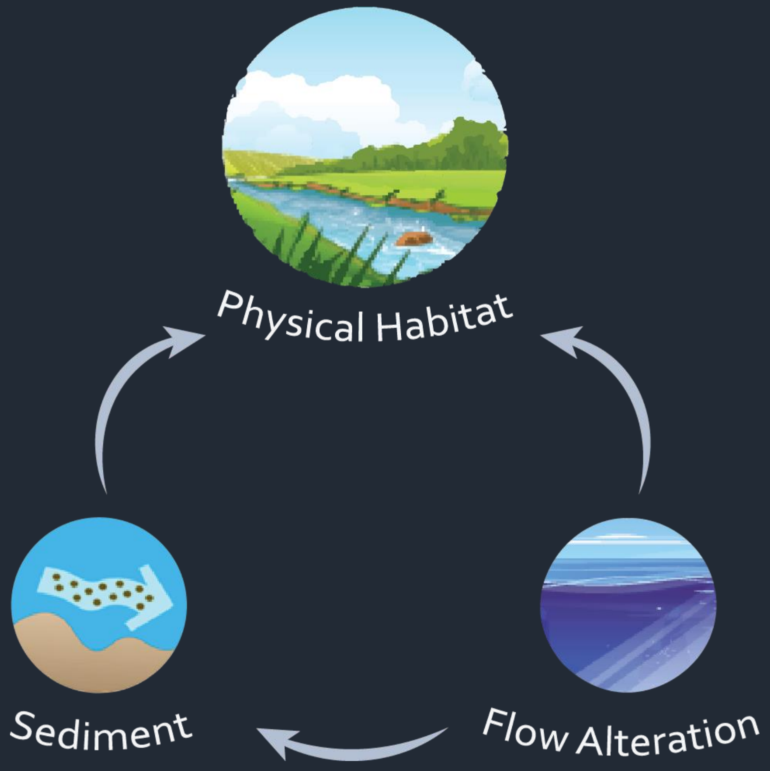
Matthew Cashman

US Geological Survey

Water Mission Area

Earth Surface Processes Division

mcashman@usgs.gov



Why ML? *Predicting habitat quality in unmonitored areas*

Physical habitat / ‘sediment’ is identified as a top stressor of stream health in the Chesapeake Bay watershed *(Fanelli et al., 2022)*

- ➢ #1 reason for ecological use impairments by mile – 303(d)

Journal of Environmental Management 371 (2024) 123139

Contents lists available at ScienceDirect

Journal of Environmental Management

journal homepage: www.elsevier.com/locate/jenvman

ELSEVIER

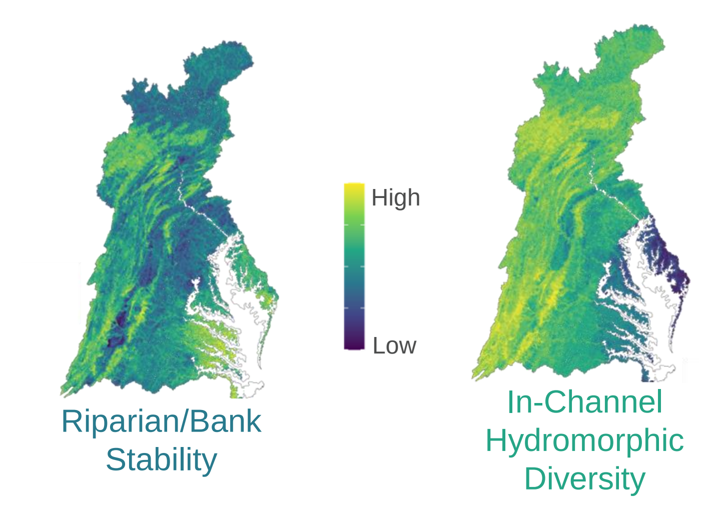
Research article

Physical habitat is more than a sediment issue: A multi-dimensional habitat assessment indicates new approaches for river management

Matthew J. Cashman [a,*], Gina Lee [b], Leah E. Staub [b], Michelle P. Katoski [c], Kelly O. Maloney [d]

Study goals:

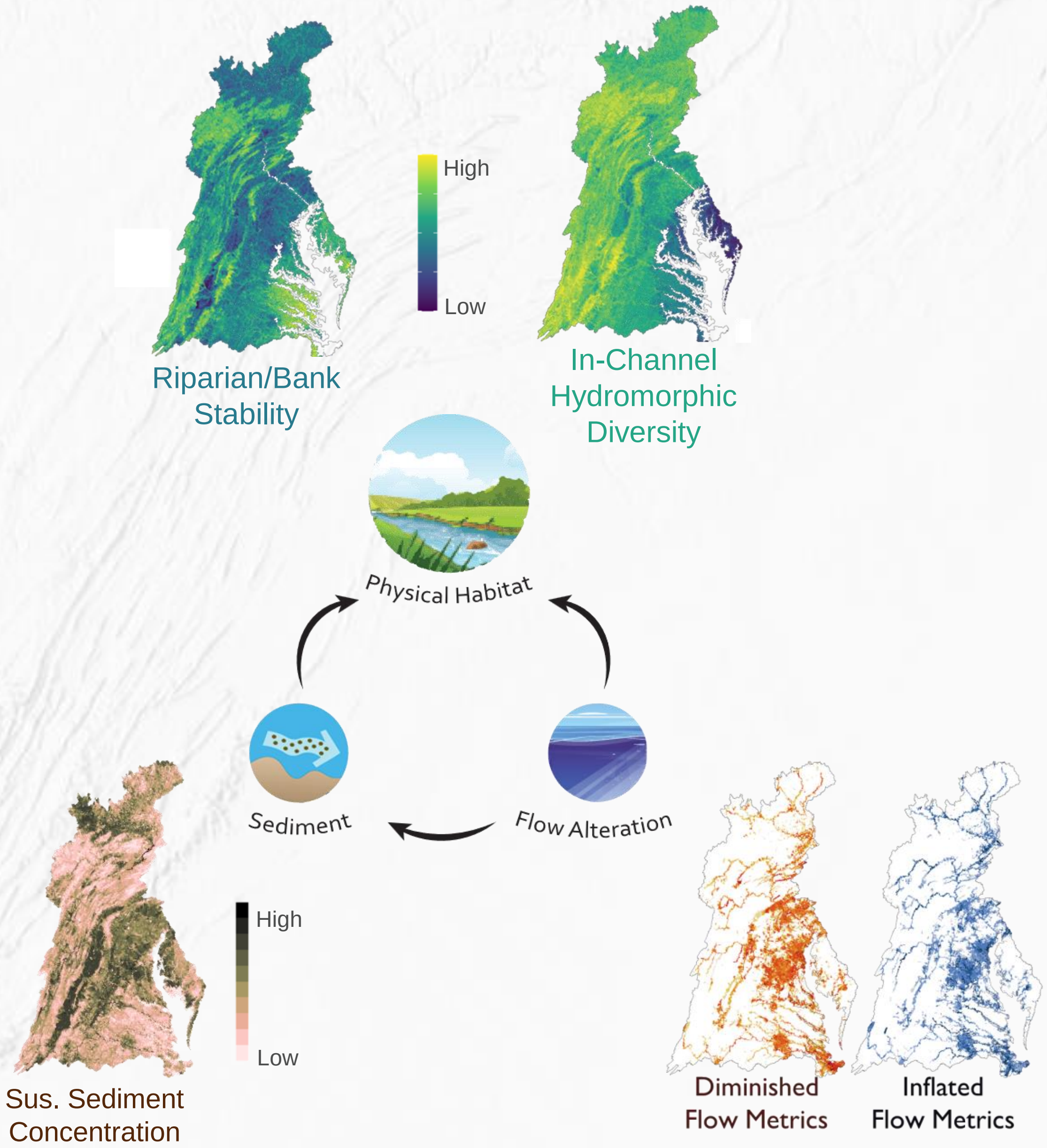
1. Where is physical habitat good/bad?
   - ❖ *Supervised continuous predictions*
2. Are there distinct dimensions of physical habitat?
   - ❖ *Unsupervised dimensionality reduction and clustering*
3. How is habitat affected by available management intervention pathways?



USGS
science for a changing world

Why ML? *Predictions allow model overlays*

Example regulatory process



Is degraded habitat caused by a sediment supply problem?

- Will restricting sediment supply improve habitat?
- What about confounding problems with flow alteration?

➢ Restorations that focus on restricting sediment, ***without addressing flows or in-channel hydromorphic diversity***, are unlikely to improve the habitat metrics that justified the TMDL.

Why ML? *They are only the means, not the end*

Focus of paper is management implications, not ML methods

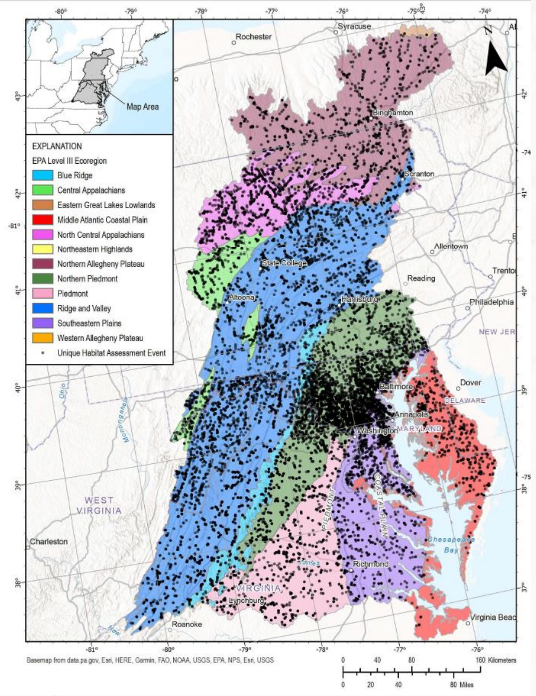
- ➢ If ML methods distracted from that focus, they were omitted

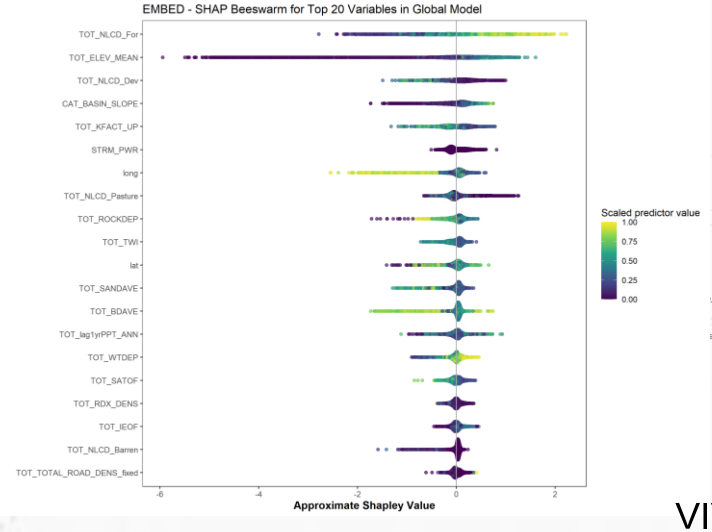
Tree-based: **random forest**, XGBoost, lightGBM, H2O AutoML, and ensemble stacking *~comparable performances*

- >20,000 observations, synoptic design
- Adjusted for mean-centered bias (Belitz and Stackelberg, 2021)

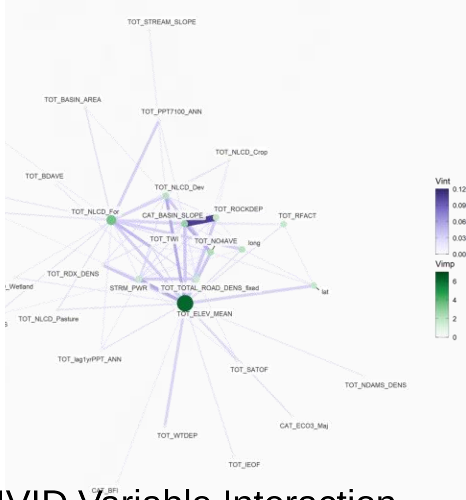
Tested many explainable AI techniques, most not in paper

- Expected relationships, nothing 'novel'
- Used for internal model consistency/validation





SHAP plots / breakdowns

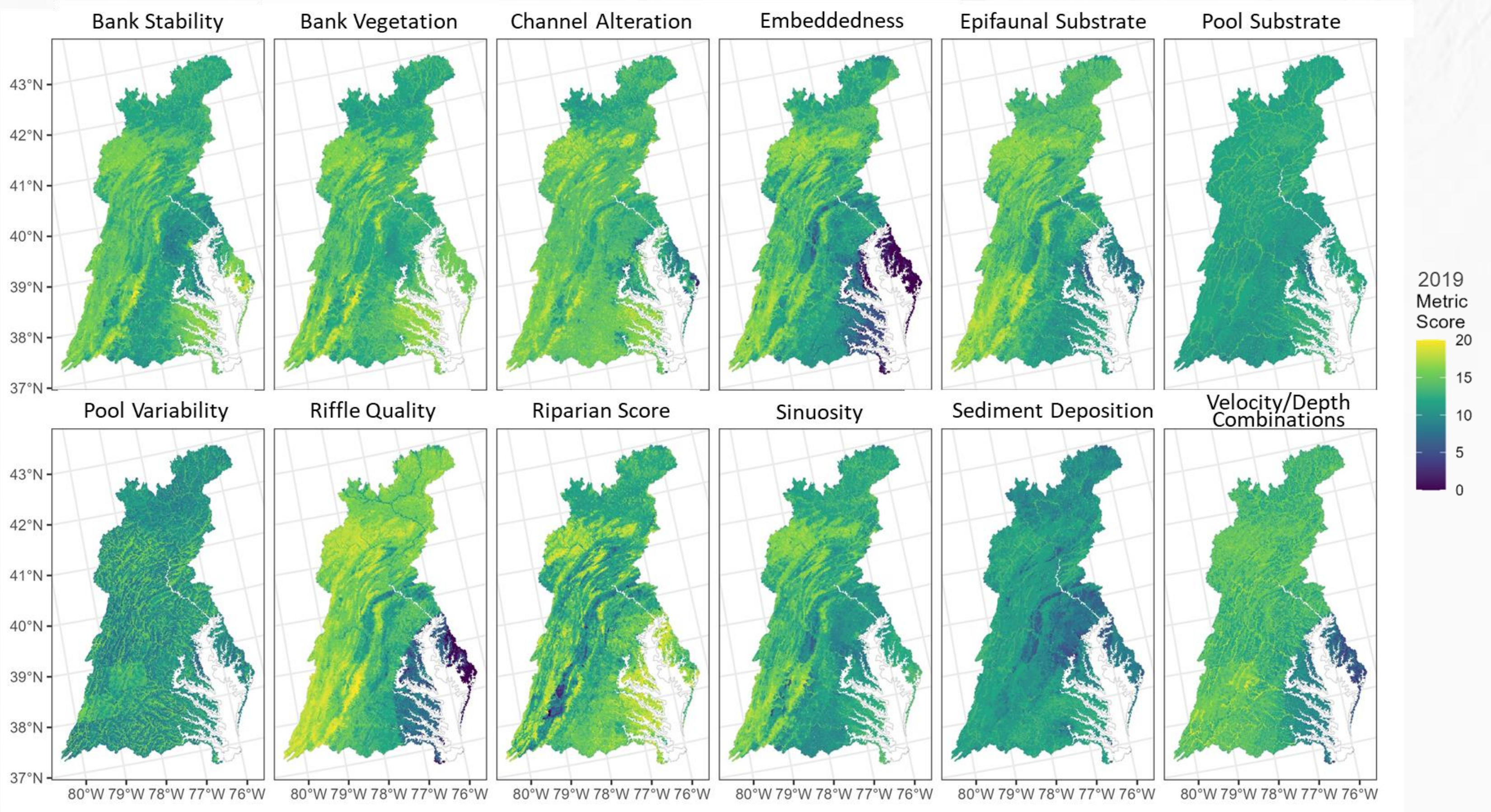


VIVID Variable Interaction network

Aligning with CBP goals? *Restoration targeting and design*

Modeled metrics are routinely used in field monitoring and well-known by stakeholders.

Models can help prioritize and identify areas for restoration or conservation (by itself or with co-occurring stressors).

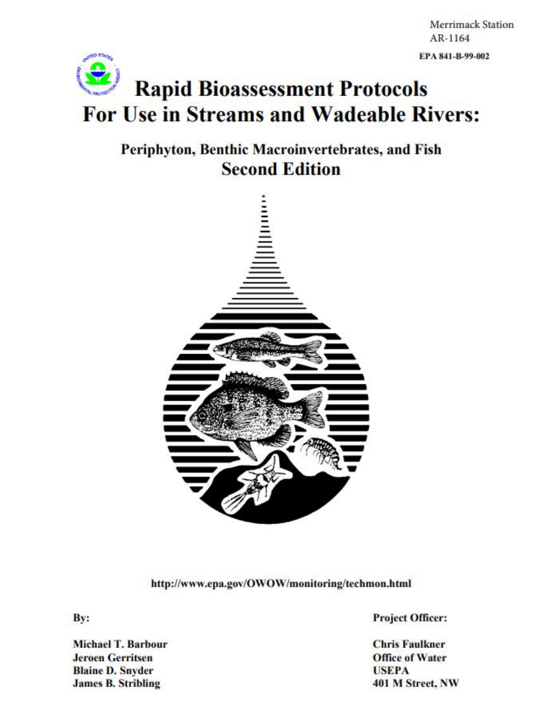


USGS
science for a changing world

Challenges: *Input data quality*

Limited by quality of input data

- Metrics are visually scored, semi-quantitative, subjective
- Field-measurement uncertainty accounted for ~80% of RMSE in our ML models


Merrimack Station
AR-1164
EPA 841-B-99-002

Rapid Bioassessment Protocols
For Use in Streams and Wadeable Rivers:

Periphyton, Benthic Macroinvertebrates, and Fish
Second Edition

http://www.epa.gov/OWOW/monitoring/techmon.html

By:
Michael T. Barbour
Jeroen Gerritsen
Blaine D. Snyder
James B. Stribling

Project Officer:
Chris Faulkner
Office of Water
USEPA
401 M Street, NW

2a. Embeddedness—High Gradient


Optimal Range *(William Taft, MI DNR)*


Poor Range *(William Taft, MI DNR)*

Habitat Parameter	Condition Category			
	Optimal	**Suboptimal**	**Marginal**	**Poor**
2.a Embeddedness (high gradient)	Gravel, cobble, and boulder particles are 0-25% surrounded by fine sediment. Layering of cobble provides diversity of niche space.	Gravel, cobble, and boulder particles are 25-50% surrounded by fine sediment.	Gravel, cobble, and boulder particles are 50-75% surrounded by fine sediment.	Gravel, cobble, and boulder particles are more than 75% surrounded by fine sediment.
SCORE	20 19 18 17 16	15 14 13 12 11	10 9 8 7 6	5 4 3 2 1 0

Challenges: *Causal questions need causal methods*

Judea Pearl's Causal Hierarchy

	Layer (Symbolic)	Typical Activity	Typical Question	Example Research Question	Statistical Methods
L_1	Associational $P(y\|x)$	Seeing	**What is?**	Where are areas quality physical habitat across the Chesapeake Bay watershed?	**Supervised / Unsupervised ML**
L_2	Interventional $P(y\|do(x), c)$	Doing	**What if I do?**	How does dam removal affect physical habitat?	Reinforcement Learning Randomized Controlled Trials A/B Testing "Observational Experiments"
L_3	Counterfactual $P(y_x\|x', y')$	Imagining	**What if instead? Why?**	What would physical habitat be if there was no anthropogenic effects at all?	Causal Mediation/Path Synthetic Control ***Causal ML***

Modified from Bareinbom et al 2022[1]

Traditional non-causal ML

Causal methods

- Questions at higher layers *cannot be accurately answered* with information and methods from lower levels
- Traditional ML can be accurate for "*What is?*" Qs (L_1) and inaccurate for *"What if?"* or counterfactual scenario Qs (L_2, L_3)
  - ML methods can infer info for a variable without it being in the dataset (Kratzert et al., 2019[2])
  - Correlation, confounding, and hierarchical dependency causes problems for estimating cause-effect L_2, L_3 scenarios, less so for L_1 predictions

➢ But what about progress-guided DL models pre-trained on cause-effect process models?

[1] Bareinboim, E., Correa, J. D., Ibeling, D., & Icard, T. (2022). On Pearl's hierarchy and the foundations of causal inference. In *Probabilistic and causal inference: the works of Judea Pearl, Association for Computing Machinery, 2021.* (pp. 507-556).
[2] Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., & Nearing, G. S. (2019). Toward improved predictions in ungauged basins: Exploiting the power of machine learning. *Water Resources Research*, 55, 11344–11354.

USGS science for a changing world

Challenges: *Causal machine learning is rapidly developing*

- Causal Inference, Causal Machine Learning, and Causal Discovery
  - Used largely in public health, econometrics, genomics, nascent in ecology (some private co., Google, Uber, Microsoft)

- Causal inference/ML techniques are specifically designed to accurately estimate cause and effect
  - Propensity score matching/weighting
  - Causal forests
  - Double-machine learning
  - Targeted Maximum Likelihood Estimation (TMLE)
  - Highly Adaptive LASSO (HAL)
  - Causal Impact (using Bayesian structural time-series models)
  - Deep End-to-end Causal Inference (DECI)
  - *Among many others…*

- This is a rapidly developing field, and not all methods are suitable (yet)
  - Depends on your dataset, specific questions, and assumptions
  - Lots of nuance, little “off-the-shelf” accessibility

Useful intro texts to causal inference