Template for Thesis Proposal

Feng Gu

August 12, 2025

Contents

1 Introduction

The Great Lakes, which occupy 84% of North America's surface fresh water and 21% of the world's supply of surface fresh water [14], are one of the world's largest surface freshwater ecosystems. Many nutrients and contaminants in the Great Lakes are stored in the sediments, playing a crucial role in supporting aquatic habitats and influencing water quality.

Contaminants in the sediments are, however, unavoidable. Among the regions, heavily developed shorelines are found around Lake Erie and the connecting channels such as the Detroit River, where human activities have left pronounced environmental footprints [1]. These impacts have raised public concerns about the potential ecological risks, and have prompted calls for stronger protective and restorative actions to safeguard these aquatic ecosystems.

In light of these challenges, scientific assessment of aquatic condition becomes fundamental to retaining the integrity of an ecosystem. One effective approach is **to assess the nature and extent of sediment contamination**, which directly reflects anthropogenic impacts and serves as a partial indicator of ecological integrity.

This rationale is grounded in a well-established body of research: sediment contamination is widely used as a proxy for assessing human-induced impacts in aquatic ecosystems. Numerous studies have demonstrated that chemical contaminants—especially trace metals and persistent organic pollutants—accumulate in sediments and can adversely affect benthic organisms and overall ecosystem health [5]. These impacts disrupt ecosystem structure and function, causing shifts in species composition, food web dynamics, and nutrient cycling. Such synchronous ecological changes provide the foundation for regression-based analysis between contamination levels and indicators of ecosystem condition. Among these indicators, the taxonomic composition of benthic macroinvertebrates is frequently used due to its sensitivity to sediment conditions and its practicality and cost-effectiveness compared to other biological measures.

Building on this relationship, one promising approach is to develop a model that links sediment contamination levels to shifts in benthic macroinvertebrate taxonomic composition. A simplified method involves constructing a composite index based on selected taxa and their relative abundances, which enhances communicability with stakeholders and simplifies interpretation. However, this index-based method inevitably sacrifices information compared to analyses using the full taxonomic dataset, and the selection of taxa and the index construction process may introduce subjectivity and limit generalizability[8]. On top of it, how to scale the index and measure the distance between indices is a crucial work to reveal the meaning of changes and compare shifts in the index values, which originally is determined by the raw benthic macroinvertebrate data. It is also important to note that benthic macroinvertebrates respond not only to anthropogenic chemical contaminants, but also to natural environmental variability such as sediment texture, organic content, and temperature. These complexities raise concerns about the reliability of such indices in attributing observed community shifts specifically to human-induced stress, thereby challenging the validity of stressor–index models that exclude natural variation.

To address this issue, it is necessary to consider environmental attributes in model development. Environmental variables such as sediment characteristics, water chemistry, and hydromorphology play key roles in structuring benthic communities. Therefore, the ability to distinguish natural variation from human-induced impacts becomes a critical question. Including environmental covariates in regression models can analytically partition their influence from that of anthropogenic stressors through multivariate analysis. (Do we need env-variables in integrity assessment? if no, remove this mention!) However, in the context of ecological integrity assessment or stress evaluation, this separation is often more complex due to the partial confounding of natural and human-induced factors. While the confounding effects of environmental variables exists in both modeling and assessment contexts, the methods required to account for them may differ substantially.

Additionally, the existence of unmeasured or unmeasurable factors—and their complex interactions with observed model inputs—may lead to non-linear relationships between variables, highlighting the importance of threshold detection. This detection introduces a crucial modeling component: identifying points at which small changes in environmental stressors lead to abrupt ecological responses. In the context of this study, such thresholds may reflect tipping points in sediment contamination levels, beyond which benthic community structures shift significantly. Detecting these thresholds enables more targeted and efficient bioassessment strategies, and may help guide the development of environmental quality criteria or inform restoration priorities.

To apply these concepts effectively, **this program will narrow its focus to the Huron-Erie corridor**, a critical aquatic link connecting Lake Huron and Lake Erie. This corridor forms a hydrologically

synchronous water system, characterized by faster flow caused by *channel constriction*—a well-known concept in fluid dynamics. These physical characteristics contribute to unique environmental conditions in the corridor and increase the complexity of sediment assessment, making traditional assessment approaches less suitable for this setting.

Our goal is to develop a more economical and efficient method to assess sediment contamination levels specifically tailored to the Huron-Erie corridor, with the potential for broader application to nearby aquatic ecosystems. The foundational idea is inspired by the work of Jian (Zhang 2008) [15], who investigated the composition of zoobenthic communities to infer contamination in this region.

Building on Jian's general framework, we aim to both enhance the original methodology and incorporate recent advancements—enabled by improved computational resources and emerging context-specific analytical techniques. Through this effort, we seek to create a more adaptive, data-driven approach to infer aquatic condition from zoobenthic measurements in complex freshwater systems.

2 Research Objectives

2.1 General objectives

The goal of this thesis is to enhance the analytical processes in Jian's work and to further design and implement the core idea of using zoobenthic measurements to infer the sediment contamination level of the Huron-Erie corridor. Specifically, the general objectives can be summarized as follows:

1. Assess anthropogenic adverse effects on aquatic sites using stressor measurements.
2. Design and implement a zoobenthic community indicator to assess taxonomic composition structure.
3. Explore the potential for using the zoobenthic community indicator to infer sediment contamination level with control of environmental variables. Design an inference model if appropriate.

2.2 Overview of the sub-objectives with the proposed framework

The above general goals will be supported and further detailed by the following specific objectives and corresponding analytical steps:

1. **Sediment contamination assessment**

   Apply an assessment method to chemical element concentrations to evaluate pollutant levels, interpret pollutant patterns among sites, and identify key contaminants.

2. **Zoobenthic community structure discriminant model of environmental variables**

   Filter reference sites based on assessed stress levels, assuming that human impact is minimal or absent, and that the community structure is primarily shaped by natural environmental variability.

   Explore the relationship between minimally impacted community structure and environmental variables, and build a predictive model to infer the expected community structure under natural conditions (without human disturbance).

   Given the limited environmental variables and other potentially unmeasured or unmeasurable environmental factors, it is nearly impossible to train a fully quantitative inference model from environmental variables to taxa.

   Therefore, constrained predicted values should be constructed and applied to avoid overfitting and to extract limited yet informative inferences about community structure using available environmental variables.

   It requires the following additional steps::

   - **Partition the reference sites into different groups** based on their taxa composition.
   - Build a discriminant model **between the taxa composition groups and environmental variables**.

   The resulting predicted group labels provide more reliable information, as the limited environmental variables are only used to classify sites into taxa composition groups that were predefined from the reference sites.

3. **Partition all sites into comparable taxa composition groups**

   Apply the discriminant model to all sites, including degraded and intermediate sites, to assign each site to one of the taxa composition groups. The group positions are fixed based on reference site partitioning.

4. **Measure the distance between each site and the reference and degraded endpoints within each taxa composition group**

   Scale the taxa composition structure and define a metric to measure the distance between each site and the reference and degraded endpoints within each group. Given similar environmental conditions across the group, differences in taxa composition are assumed to be caused by human disturbance.

5. **Quantitative regression to assess the relationship between the stressor level and the taxa composition structure**

   Variations in taxa composition from the reference endpoint should be explained by the relative stress level of each site. A quantitative regression model can assess this relationship. Considering the large number of potential ecological influences, non-linear quantile regression is employed to explore patterns beyond traditional mean regression.

2.3 Potential Impacts and Further Questions

Through achieving these objectives, there is a potential to significantly enhance the understanding of the sediment contamination within the Huron-Erie corridor and to develop a more robust zoobenthic community indicator that can be used to infer sediment contamination levels.

Several benthic integrity assessment relevant questions arise and might be answered from the implementation of these objectives, tentatively listed as follows:

- Is there a generally applicable sediment contamination assessment method that can be applied to different types of environmental conditions, or does the method need to be tailored to specific conditions?
- To what extent the clustering-discriminant model [1] can effectively capture the natural variability in taxa composition, and how the clustering parts, which are applied on different **partitions** in contamination assessment results, can shape the discriminant model?
- How to construct the measurement of zoobenthic community structure according to the contamination assessment results to support the inference model? To what extent should the two measurements be correlated but meanwhile keeping partial independence to ensure their respective interpretability?
- How to link the measurement of zoobenthic community structure to probability measurement of sediment contamination level - degraded, especially in areas with intermediate anthropogenic influence.

[1] The clustering and discriminant methods are bundled together to form a single step in the framework. Its purpose is to extract confined information(groups) from the selected sites, fit a discriminant model of environmental variables for the groups, and then apply the model to all sites to assign them to the groups.

3 Data Description

3.1 Data Collection

The data about Detroit River used in this program is provided by Professor Jan and was originally collected by Jian and her team during the Lake Huron–Lake Erie Corridor survey in July-August 2004. The survey collected 16 taxonomic variables, 8 environmental variables, and 30 stressors across the study zone.

The sampling site locations were determined prior to fieldwork by a stratified random sampling design to ensure representative coverage. The sampling zone encompassed the entire Detroit River, including both the upstream mixing zone with Lake Saint Clair and the downstream transition into the Lake Erie entrance.

To enhance the estimation and control of temporal variability, data from two previous studies—which collected environmental conditions in the same survey zone (Farara and Burt 1993; Wood 2004)—were compiled and incorporated by Jian and her team into the 2004 data set. Both of these earlier data sets followed the same field protocols as the 2004 Lake Huron-Lake Erie Corridor survey. For the taxonomic data, information from three separate benthic surveys was combined to increase sample size and provide more details on taxonomic. These combinations improved environmental clustering of aquatic sites and the construction of zoobenthic community indicators.

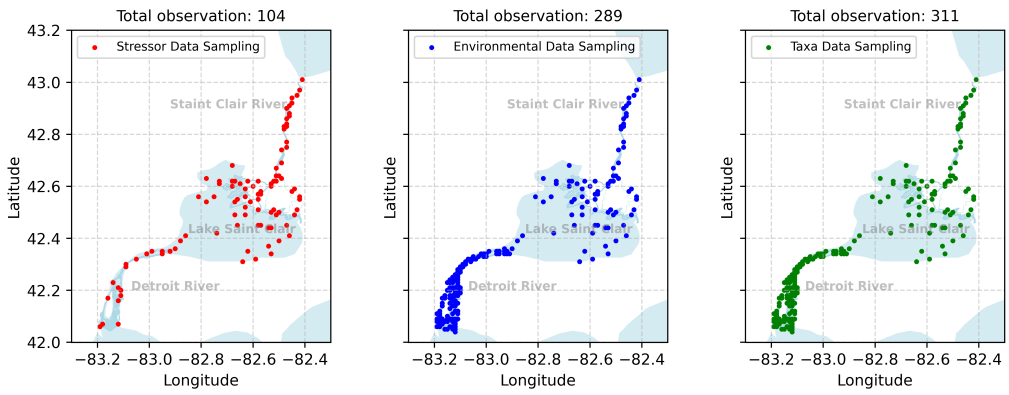

Figure 1: Different data sampling locations of the three types of data in the Lake Huron-Lake Erie Corridor survey area.

In data quality checking work, an evaluation of the sampling sites, as shown in Figure 1, reveals a spatial mismatch in the data sources: the taxonomic, environmental, and stressor datasets are not fully aligned across the survey area. [2] Additionally, Figure **??** visually illustrates how discrepancies in sampling locations and dataset sizes may affect the validity of modeling efforts. To address this issue, only those sites that contain all three types of data—taxonomic, environmental, and stressor—should be used to train the model.

A further comparison with the appendix of Jian's thesis (p. 164) shows that the full environmental dataset includes 311 sampling sites distributed across the Detroit River, Lake St. Clair, and St. Clair River zones. While the current dataset in use does not include all these sites, this limitation can be resolved by accessing the full environmental and taxonomic data from Jian's appendix and merging them by the site ID.

However, the stressor data presents a unique constraint: it is not included in the appendix and is only available in a separate file that contains measurements distributed across the three major zones. As such, this dataset reflects stressor impacts at a broader corridor scale, not limited to the Detroit River zone.

[2] By checking the thesis page 24, data from two previous studies was incorporated and compiled, but it did not mention what specific data type was included. Meanwhile, data from three separate benthic surveys was also combined into the 2004 data set.

3.2 Environmental attributes and samples

In the 2004 Lake Huron-Lake Erie Corridor survey, eight environmental attributes were measured at each sampling site. The site **location** (longitude and latitude) was recorded based on the GPS reading.

Temperature (°C) and **dissolved oxygen concentration (mg/L)** were measured with a Hydrolab multimeter. **Water depth (m)** was recorded from the Ponar rope, and **water velocity (m/s)** (measured 0.5 meters below the surface) was obtained with an Ott C-3 portable current meter. **Loss on ignition (%)** and **median particle size (phi-units)** were determined during sediment processing but are treated as environmental attributes due to their fundamental roles in habitat characterization (details on their analysis are provided in a later subsection). [3]

The meanings and ecological relevance of these six attributes are as follows:

- **Temperature (°C):** Water temperature affects metabolic rates, distribution, and activity patterns of aquatic organisms.
- **Dissolved Oxygen Concentration (mg/L):** Indicates the amount of oxygen available for aquatic life; low levels can limit survival and exclude sensitive taxa.
- **Water Depth (m):** Influences light penetration, stratification, and habitat availability for benthic organisms.
- **Water Velocity (m/s):** Reflects the flow regime, which shapes sediment characteristics and determines which taxa can colonize a site.
- **Loss on Ignition (%):** Measures organic matter content in sediments, indicating productivity and food resource availability for benthic fauna.
- **Median Particle Size (phi-units):** Describes sediment texture, affecting substrate stability and the suitability of habitats for different taxa.

These environmental attributes collectively characterize the abiotic conditions at each site and strongly influence the taxonomic composition and abundance of benthic invertebrate communities.

Additionally, these attributes are commonly used to describe the baseline environmental conditions of aquatic habitats, as they are primarily governed by natural physical and chemical processes. By including these variables as covariates in the analysis, we can better control for natural environmental variation in the taxonomic composition of zoobenthic communities, thereby isolating the effects of stressors of interest. However, it is important to note that certain attributes (e.g., organic matter content or temperature) may also be indirectly influenced by anthropogenic activities in some settings. Care was taken to interpret their roles in the context of site history and land use, but in general, these variables serve as key descriptors of the underlying habitat conditions independent of direct contamination or disturbance.

To ensure comparability of the merged data, the 2004 Lake Huron-Lake Erie Corridor survey followed the same sampling protocols as those used in the two previous studies when recording environmental attributes.

3.3 Taxonomic attributes and samples

The zoobenthos were collected with a Petie Ponar grab sampler. After considering the fullness of each grap and the removing of fine materials, the team applied multiple grabs at each site until a total volume of $2L$ sediment was collected. The sediment samples for organic and metals analysis were preserved in corresponding professional containers, all these samples were stored frozen.

One zoobenthic sample replicate from each site was randomly selected and processed, while the other two were archived. Samples were sieved into size fractions (4 mm, 1 mm, 0.5 mm, 0.25 mm), then elutriated to separate lighter detritus and animals from inorganic sediments. Each fraction was sorted under a microscope and organisms were identified to the lowest possible taxonomic rank using standard keys. Zoobenthos were preserved in 70% ethanol in labeled vials and archived at the University of Windsor.

Specifically, there were 16 taxa recorded from the sediment samples, as shown in the table 7. According to their creature characteristics and preferred habitat, these taxa can be gently divided into three groups:

[3] Although loss on ignition and median particle size are derived from sediment samples, they are considered environmental variables because they reflect essential physical and chemical habitat features.

Table 1: Benthic Taxa with Explanation and Preferred Habitat Feature

Taxa	Explanation	Preferred Habitat
Nematoda	Roundworms	Broad
Chironomidae	Non-biting midges (larvae)	
Ceratopogonidae	Biting midges	
Amphipoda	Small crustaceans (scuds)	
Acari	Aquatic mites	
Hydrozoa	Small predatory animals	
Gastropoda	Snails and slugs	
Oligochaeta	Aquatic segmented worms	Depositional zone
Hexagenia	Mayfly genus (larvae)	
Dreissena	Zebra/quagga mussels	
Hirudinea	Leeches	
Turbellaria	Flatworms	
Sphaeriidae	Fingernail clams	
Caenis	Mayfly genus (larvae)	Erosional zone
Hydropsychidae	Net-spinning caddisflies	
Other Trichoptera	Other caddisfly families	

Immediately after the initial sorting of samples, ten samples were randomly selected to assess the sorting efficiency. One sample had a sorting efficiency of 91%, while the remaining samples had efficiencies of 96% or higher.

Each taxa responses to the changs in environmental and stress conditions differently, modeling this changes-response relationship will help us to understand how the zoobenthic community reacts to the external changes and use their response to assess the ecological integrity.

However, to consider the precise response of each taxa to the infinite environmental and stressor changes is impossible, but knowing how they respond to the stressors can help identify the level of impairment that was caused by anthropogenic activities. An alternative way to assess the taxonomic response to the stressors is to observe the changes in taxonomic composition, which is a generalized response of the zoobenthic community instead of the individual taxa. This generalized response is more representative than a single or a few taxa and partially eliminates the noise in taking limited taxa group, in which the noise may be caused by the subtle changes(too small to be detected) in the external conditions.

3.4 Stressors samples

Sediment samples from each site were thoroughly mixed to ensure homogeneity. The homogenized samples were then split into separate portions for different analyses, including median particle size, total organic carbon (TOC), organic contaminants, and metals.

- **Particle Size:** Median particle size analysis was performed by sieving dried sediment through a series of sieves of decreasing mesh size. Each size fraction was weighed and described using phi units ($\phi = -\log_2 d$), where d is particle size in mm.
- **Total Organic Carbon (TOC):** Sediment TOC(%OC) was determined using loss on ignition (LOI). Pre-weighed, dried sediment samples were combusted at 450°C for 24 hours, and organic carbon was determined gravimetrically by subtracting the remaining mass.
- **Organic Contaminants:** The concentrations of organic contaminants(including 1245-TCB, 1234-TCB, QCB, HCB, OCS, p,p'-DDe, p,p'-DDD, mirex, Heptachlor Epoxide, total PCB) were measured using a gas chromatograph equipped with a 63Ni electron capture detector, following standard operating procedures.
- **Metals:** Metal concentrations (including Al, As, Ca, Cd, Co, Cr, Cu, Fe, Mn, Ni, Pb, and Zn) were analyzed using an Inductively Coupled Plasma Optical Emission Spectrophotometer (ICP-OES). For total mercury (Hg), an atomic absorption spectrophotometer (AAS) was used with a vapor generation accessory for increased sensitivity. Liquid samples were introduced into the instrument for metal analysis.

Table 2: Sediment Stressors, Explanations, and Analytical Steps

Measurements	Explanation	Analytical Step
Metals		
Al	Aluminum (trace metal)	ICP-OES after acid extraction (Metals step)
As	Arsenic (toxic element)	
Ca	Calcium (major element, hardness)	
Cd	Cadmium (toxic metal)	
Co	Cobalt (trace element)	
Cr	Chromium (trace metal)	
Cu	Copper (trace metal, micronutrient)	
Fe	Iron (major element, micronutrient)	
Mn	Manganese (trace element)	
Ni	Nickel (trace metal)	
Pb	Lead (toxic metal)	
Sb	Antimony (trace element)	
V	Vanadium (trace element)	
Zn	Zinc (trace metal, micronutrient)	
Hg	Mercury (highly toxic metal)	Atomic Absorption Spectrophotometry (AAS, Metals step)
Total Organic Carbon		
%OC	Percent organic carbon	LOI combustion at 450°C for 24h (TOC step)
Organic Contaminants		
1245-TCB	1,2,4,5-Tetrachlorobenzene (organic pollutant)	Gas Chromatography with Electron Capture Detector (GC-ECD, Organic Contaminants step)
1234-TCB	1,2,3,4-Tetrachlorobenzene (organic pollutant)	
QCB	Quintachlorobenzene (organic pollutant)	
HCB	Hexachlorobenzene (organic pollutant)	
OCS	Octachlorostyrene (organic pollutant)	
p,p'-DDE	DDT breakdown product	
p,p'-DDD	DDT breakdown product	
mirex	Organochlorine insecticide	
Heptachlor Epoxide	Organochlorine pesticide breakdown product	
total PCB	Total polychlorinated biphenyls	

Quality assurance and chemical analyses were performed in collaboration with the Great Lakes Institute for Environmental Research (GLIER) at the University of Windsor[15].

Among the chemical variables analyzed, not all elements serve as stressors to benthic taxa in the same way. Major earth elements such as Al, Ca, Fe, K, Mg, and Na are generally considered non-toxic at typical environmental concentrations, as they are naturally abundant in sediments and primarily reflect the natural composition of the substrate. However, industrial pollution or other aquatic activities can elevate their concentrations, making them potential stressors, though their impacts are more challenging to assess due to naturally high background levels.

In contrast, trace metals—including As, Bi, Cd, Co, Cr, Cu, Hg, Mn, Ni, Pb, Sb, V, and Zn—are regarded as pollutants due to their anthropogenic origins and toxicological effects on aquatic organisms. Elevated concentrations of these metals usually indicate the overexploitation of natural resources and industrial activities, which can cause bioaccumulation and toxicity in benthic organisms. These metals are not biodegradable and ultimately accumulate in sediments, reflecting not only current but also historical pollution legacies.

Additionally, persistent organic pollutants such as PCBs, QCB, HCB, OCS, p,p'-DDE, p,p'-DDD, mirex, and Heptachlor Epoxide originate mainly from historical industrial activities and pesticide use. Due to their chemical stability, they persist in sediments and accumulate in aquatic food webs, where they can cause chronic and sub-lethal effects—including reproductive and developmental toxicity—in aquatic organisms, and thus pose long-term risks to ecosystem health and biodiversity.

The percent organic carbon (%OC) in sediments typically originates from the decomposition of plant and animal material, as well as inputs from terrestrial runoff and aquatic primary production. While not a pollutant itself, organic carbon plays a key role in sediment chemistry by influencing the binding and retention of contaminants, thereby affecting their mobility and bioavailability to aquatic organisms. High levels of organic carbon can thus modulate the ecological impact of other pollutants in the sediment.

By distinguishing between natural background elements and anthropogenic pollutants, these measurements enable a more accurate assessment of the sediment stress level and its ecological implications for the benthic community.

4 Literature Review(to majorly rewrite)

4.1 Sediment Contamination Evaluation Methods

Majorly to talk about the common sediment contamination evaluation methods, and distinguish which methods are suitable across various environmental conditions. The generality of the assessment method can support a good stress level-environmental distribution in the data, which is crucial to make sure enough data points with fixed stress level across the environmental values.

4.2 Taxa composition clustering with minimal (fixed) stress level

Majorly to talk how this clustering can help to identify the taxa composition patterns, which can be used as predicted values of the discriminant function analysis.

4.3 Discriminant Function Analysis of environmental variables for taxa composition clustering

4.3.1 Aquatic Integrity by Sediment Contamination Evaluation

Aquatic integrity refers to how well an ecosystem maintains its structure and function under both natural and human pressures. It encompasses the ability of aquatic systems to support and sustain a balanced, adaptive community of organisms having a species composition, diversity, and functional organization comparable to natural habitats within a region.

In aquatic ecosystems, sediments act as long-term pollutant archives. They accumulate and retain contaminants over time—especially heavy metals, pesticides, hydrocarbons, and other industrial or urban pollutants. Therefore, sediments exert a considerable influence on the biological health of benthic organisms. However, it is often difficult to determine whether metal accumulation is due to natural processes or anthropogenic sources [3].

Within such pollutant archives, anthropogenic contaminants accumulate through various pathways, including industrial discharges and agricultural runoff. Many of these pollutants—such as cadmium, lead, and mercury—are toxic to benthic organisms even at low concentrations. High contaminant levels are frequently associated with reduced biodiversity and shifts in community composition. These biological responses provide the foundation for using zoobenthic measurements to inversely infer sediment contamination levels and, in turn, evaluate the ecological integrity of aquatic systems.

However, anthropogenic pollution is not the only driver of changes in biological communities. Natural environmental variability—including sediment texture, organic matter content, and hydrological conditions—also plays a significant role in shaping benthic communities. As such, sediment contamination alone cannot fully explain or represent ecological integrity. To account for the full picture, environmental conditions must also be considered.

Nonetheless, in this program, we emphasize the controllable component of aquatic degradation—human-induced pollution—as our primary focus. While environmental variability is acknowledged, it is not the central concern of this study. Our goal is to develop a model that assesses ecological condition by leveraging sediment contaminant data and linking it to zoobenthic community structure, with the understanding that it may reflect only part of the full ecological integrity.

Frequently used evaluation methods include contaminant index-based approaches such as the Enrichment Factor (EF) and the Geoaccumulation Index (Igeo). These indices provide pollution scores relative to known background or reference values, often focusing on individual or selected chemical elements [2]. However, such methods typically ignore interactions between pollutants and rely on accurate background concentrations—which may be difficult to obtain or define. Furthermore, these approaches are less effective when analyzing large-scale datasets with many chemical elements and spatially distributed sampling sites.

Another type of evaluation method is Principal Component Analysis(PCA) based methods. Such methods are multivariate and data driven, considering all chemical variables and revealing their interaction pattern by dimensionality reduction. Additionally, they do not need for background/reference values, which saves the resources for identifying these benchmarks. By trade off, PCA methods may provide less intuitive scores(assessments) that lack direct interpretation, and the absence of benchmarks constrains the clear or comparable assessment results, no "moderately" or "severely" polluted sites can be identified as in the index methods.

In conclusion, both groups of methods serve different but complementary purposes: indices for direct contamination assessment, and PCA for exploring patterns or creating composite indicator.

However, the integrity of an aquatic site is determined not only by the anthropogenic impacts, but also by the natural pressures, ignoring natural pressures can make the assessment results biased, like over- or underestimating ecological degradation. The potential natural variabilities that shape the aquatic integrity include: geology, flow regime, temperature, elevation and so on. Consequently, a naturally metal-rich geology site might appear "degraded" but actually be natural, and a site with no pollution but naturally poor biological diversity [4] might be wrongly flagged as "degraded" in the inference process. Therefore, quantifying anthropogenic pollution through sediment contamination data mainly aims to assess the anthropogenic influence, and controlling for natural variation is necessary to assess the aquatic integrity through such quantification on pollution.

4.3.2 Aquatic Integrity by Biological Condition Gradient

Another type of assessment method is Biological Condition Gradient (BCG), which is different from the contaminant evaluation in its conceptual basis, data sources and focus.

BCG starts with a stress-response framework, categorizing sites into 6 biological condition levels, uses biological metrics(which the contaminant evaluation do not) to measure the biological responses and other ecological responses to the stressors and finally makes an integrative measure of ecosystem integrity. One of the key advantages of BCG is its widely applicable feature over a relatively large range [5] of aquatic ecosystems, which makes the biological condition to be interpreted independently of assessment methods[7]. However, it needs to build reference conditions from minimally disturbed sites, and deviation from these benchmarks(reference conditions) reflects biological degradation, which requires empirical data and expert judgment to assess the biological integrity.

Considering my research objectives and the available data sources, the contaminant evaluation are more suitable for my work, which reflects potential stress but not whether ecosystems are responding biologically. Such conceptual properties make the inference for degree of pollution through **biological responses** possible and avoid laking the biological information to the input data of the inference model.

4.4 Ecological Thresholds Detection and Inference

4.4.1 Ecological Thresholds Existence and Application Scope

Ecological thresholds are points at which a relatively small change in an external condition(like pollution, or nutrient level) causes a rapid and significant change in ecosystem structure or function.

The concept of ecological thresholds emerged in the 1970's from the idea that ecosystems often exhibit multiple 'stable' states, depending on environmental conditions[9](Holling 1973, as cited in Groffman et al. 2006). These stable states that are separated by thresholds can be long-lasting conditions that an ecosystem can exist in - such as clear-water lake with abundant vegetation versus a turbid, algae-dominated lake, with different ecological structures, functions and species compostion.

The shifts between these states occur when thresholds are crossed. Some shifted states may arise naturally and are not harmful to human societies or the surrounding environment, but others may be, leading to the loss of both economic and ecological value. For those potential shifts toward undesirable states that may or will cause such losses, detecting and inferring thresholds is economically and ecologically important, as it helps guide management policies and prevent potential degradation. Additionally, the identified thresholds can help set restoration targets, beyond which preservation efforts are more likely to support long-term ecological goals—especially when the recovered state is near a threshold and at risk of degrading again. Therefore, interest in ecological threshold has grown in both ecological management and restoration fields, with the aim of maintaining or restoring ecosystems in a desired state.

Turning to the scope of ecological threshold analysis, there are three main ways that threshold concepts have been applied in ecology: (1) analysis of dramatic and surprising "shifts" in ecosystem state; (2) the determination of "critical loads" (3) analysis of "extrinsic factor thresholds" [12]. Their explanations are summarized in Table 3.

The three scopes have respective preferred applications but they are not mutually exclusive. In some studies, they are used together to provide a more comprehensive understanding of ecological thresholds, extending the threshold application to multiple aspects of research and management.

[4] Biological diversity is not an environmental attribute, but a biological response variable.

[5] Large ranges: regional level(over several states or provinces), country level, continent level.

Table 3: Three main ways threshold concepts are applied in ecology (Peterson, etc. 2006 [12]).

Application Aspect	**Meaning**
Shifts in ecosystem state	Analyzing dramatic and surprising changes in ecosystem condition caused by small changes in a driver.
Critical loads	Determining the pollutant level an ecosystem can absorb without experiencing a shift in state or function.
Extrinsic factor thresholds	Analyzing how large-scale variable changes affect relationships between drivers and responses at smaller scales.

My project aligns with two ecological threshold applications. It fits the *shifts in ecosystem state* scope by using zoobenthic data to detect biological changes that signal ecosystem transitions. It also supports the *critical loads* approach by linking stressor levels to community responses, estimating pollutant thresholds beyond which ecological integrity declines, while controlling for environmental variation.

4.4.2 Ecological Thresholds Detection and Inference Methods

Both natural variables (e.g., temperature, altitude) and anthropogenic stressors (e.g., pollutant concentration) can show threshold behavior, where crossing a breakpoint in both these variables can lead to rapid shifts in ecosystem structure or function.

On top of that, the existence of thresholds often suggests the fact that the ecosystem is not linear in its response to external changes, not only in the rate and direction of change, but also in the heteroscedatic nature and other properties. It challenges the classical regression methods that constrain the strict linearity assumption and urges the implmentation of more flexible models that fit specific ecological contexts.

Typically, changes in conditional means are the exclusively focus of many traditional regression methods, which may fail to distinguish significant changes beyond the mean scope in heterogeneous distributions. Some extended traditional methods, such as weighted least squares (WLS) is designed by assigning weights inversely proportional to the variance at each level, to address *heteroscedasticity* where the variance is not constant. However, heteroscedasticity is only one type of complexity in ecological data[4]. Ecological relationships often involve *multiple sources of variation*, such as:

- **Unmeasured limiting factors:** Not all influential variables can be measured or included in the model, leading to variation in the response not explained by the predictors.
- **Heterogeneous responses:** Changes may occur primarily in certain portions of the response distribution (e.g., only in the upper quantiles), while mean trends remain flat.
- **Nonlinear and threshold effects:** Ecological thresholds may be apparent in certain quantiles or as abrupt changes not captured by mean regression models.

As a result, while methods like WLS can address heteroscedasticity, they may still fail to capture the full range of ecological relationships—especially when important ecological processes affect only a subset of the data or produce complex response patterns beyond simple mean-variance trends.

Quantile regression (QR) can be a desirable and practical solution to these challenges, benefiting from its less strict assumptions on parametrics and its ability to model the quantile-specific ($\tau \in [0, 1]$) relationship between variables[11]. It provides alternatives to reveal limiting factors by fitting different quantiles of the response distribution, and does not require the heteroscedasticity assumption in relationships, making the exploration beyond the mean scope possible. Additionally, the detection of nonlinear and threshold effects can be achieved by fitting piecewise quantile regression models, capturing abrupt changes in the response distribution at specific quantiles.

Horning (2013)[10] explores the concept of ecological thresholds in the context of behavioral physiology, and introduces the use of *constraint lines*—the boundaries delimiting point clouds in bivariate scatterplots—to reveal limiting factors and physiological constraints. In his study, quantile regression were employed to analyze these constraint lines, this method losses the assumptions on parametrics and therefore covers wider range of possible relationships between variables.

Cade and Noon (2003)[4] makes a gentle discussion on quantile regression and its application in ecology, highlighting the statistical properties and possible regulations in changes of estimated coefficients across different quantiles that a QR model may have under different complicated scenarios. They discussed several representative examples of hidden effects in ecological data, which can not be captured by mean-regression models. Additionally, a quantile regression model was fitted on a linearly increasing heteroscedasticity case, and bootstrapping method was employed to estimate the confidence intervals of the coefficients, showing a relatively stable and reliable performance of a QR model in such cases. They concluded that heteroscedasticity is the cause that leads to changes in the coefficients across quantiles, and that the QR model can be a good choice to address this issue. However, a "linearity" relationship between quantiles of the response and predictors should be the underlying factor that supports the use of QR models, because even a case of homoscedasticity (constancy in variance, not in distribution shape) can cause changes beyond the mean scope. Such a homoscedastic case—though rare in the real world—that results in changes beyond the mean further reinforces the need for quantile regression models in ecological data analysis.

Tests on accuracy and reliability of detected thresholds are necessary, several factors can affect these tests. Based on the work of Daily et al. (2012)[6], these factors influence the detection quality with different degrees in different contexts, including:

- Sample Size: Smaller sample sizes generally increase the rate of false threshold detection, where as larger sample sizes improve the reliability of threshold estimates.
- Sample-Environment Distribution(SED): The frequency and distribution of samples across the environmental gradient (SED) can greatly influence both the detection and the estimated location of thresholds. Non-uniform or uneven SEDs can lead to biased or misleading results.
- Rate of Change: The actual rate of linear or nonlinear change in ecological response can interact with statistical method properties, affecting detection outcomes.
- User-selected Model Parameters: The choice of model parameters-such as the quantile level(τ) in QR, or bandwidth in smoothing methods (e.g., sizer)-significantly impacts the detection rate and accuracy of threshold locations.

Correlatedly with the risks in false detection and/or inference, Spake et al. (2022)[13] synthesizes evidence on threshold detection and emphasizes that the concept of scale is fundamental to understanding, quantifying, and interpreting ecological thresholds. They organize the scale-dependence of threshold detection into four aspects, as shown in Table 4.

Table 4: The Scale Framework in Ecological Threshold Detection (adapted from Spake et al., 2022) (**Grain** refers to the smallest unit of measurement, affecting the resolution and detail of the data. **Extent** refers to the overall scope (area or time) of the study, affecting the range of environmental or temporal gradients observed.)

Scale Concept	**Description**
Grain (Resolution)	The size of the smallest unit of observation or measurement (e.g., plot size, pixel size, sampling interval); determines the level of detail captured.
Extent	The total area or duration covered by the study; determines the environmental or temporal gradient sampled and potential to observe thresholds.
Organizational Level	The biological or ecological hierarchy at which data are collected or analyzed (e.g., individual, population, community, ecosystem).
Analytical Method	The type of statistical or modeling approach used to detect thresholds, which can influence the sensitivity and interpretation of results.

In my project, piecewise quantile regression model (PQRM) will be the major analytical method to detect the potential thresholds between the stressors and the taxonomic composition, allowing for a

nuanced understanding of their relationships. Such scale framework can inform my project together with the factors mentioned above by guiding decisions on sampling design, data aggregation (if needed), and model selection, increasing the robustness and ecological relevance of threshold detection.

4.5 Synthetic Data in Machine Learning for Ecological Assessment

References already in hands include:

- *'Small Data' for big insights in ecology (citation to be added)*

There are more references to be added this month, after reviewing and confirming their relevance to the thesis objectives.

5 Methodology

There are various symbols used in the section, some of them are function names, some are vectors, and some are matrices. The following table summarizes the symbols used in this part:

Table 5: Summary of major mathematical symbols and their meanings, organized by subsection.

Subsection	Symbol	Meaning
Data and grouping variables	m	Number of sampled sites
	$X \in \mathbb{R}^{m\times 30}$	Elemental concentration matrix (30 chemical elements)
	$E \in \mathbb{R}^{m\times 5}$	Environmental variable matrix (5 variables)
	$T \in \mathbb{R}^{m\times 16}$	Taxa abundance matrix (16 taxa)
	$s \in \mathbb{R}^m$	Composite stressor score (from PCA)
	$I_{\text{ref}} \in \mathbb{R}^m$	Indicator: 1 = reference site, 0 = disturbed site
Reference site clustering	$\mathcal{C}_K$	Cluster label (taxa composition group) from reference sites
	$\hat{\mathcal{C}}_K$	Predicted cluster label for disturbed sites
	$p\%$	Percentage of least stressed sites chosen as reference
	$\mathcal{R}_k$, $\mathcal{D}_k$	Sets of reference and disturbed sites in group k
Transformation and Gaussian modeling	ϕ_{Hel}	Hellinger transformation
	$\boldsymbol{\mu}_k$	Mean taxa composition vector for group k reference sites
	$\boldsymbol{\Sigma}_k$	Covariance matrix of taxa composition in group k
	λ	Ridge regularization term
	I_{16}	16×16 identity matrix
	$\tilde{T}_{k,j}$	Whitened deviation vector for site j in group k
Z-score Community Index (ZCI)	$\text{ZCI}_{k,j}$	Scalar Mahalanobis distance from reference centroid
	$\text{ZCI}_{k,j}^{(\text{diag})}$	Diagonal approximation ignoring correlations
	$\text{ZCI}_{k,j}^{(1)}, \text{ZCI}_{k,j}^{(2)}$	First two components of multi-dimensional ZCI
	$\text{ZCI}_{k,j}^{\star}$	0–100 scaled ZCI score
	V_k	PCA loading matrix from whitened reference deviations
Quantile regression modeling	$\mathcal{F}_{\text{dis}}$	Discriminant function mapping E_{ref} to $\mathcal{C}_K$
	$\mathcal{F}_{\text{reg},k}$	Regression function linking ZCI to stress level in group k
	$Q^{(k)}_{\delta X\mid Z}(\tau \mid z)$	Conditional τ-quantile of δX given ZCI z
	κ_m	Fixed breakpoint in piecewise regression
	$\beta^{(k)}_{0,\tau}, \beta^{(k)}_{1,\tau}$	Intercept and slope in quantile regression
	$\gamma^{(k)}_{m,\tau}$	Slope change after breakpoint κ_m
	$\delta X_{k,j}$	Stress level relative to group-k reference median
Hypothesis testing for degradation	$F^{(k)}_{\delta X\mid Z}(x \mid z)$	Conditional CDF of stress level given ZCI
	x^*_k	Group-k stress threshold for degradation classification
	p	One-sided p-value for degradation test

At the initial stage, the whole information about the sites can be shown in the matrix form:

$$\begin{bmatrix} X & E & T \end{bmatrix} \in \mathbb{R}^{m\times(30+5+16)}$$

where $X \in \mathbb{R}^{m\times 30}$ (elemental concentrations), $E \in \mathbb{R}^{m\times 5}$ (environmental variables), and $T \in \mathbb{R}^{m\times 16}$ (taxa abundances).

5.1 Find Reference Sites - Sediment Contamination Assessment

why to do this? one sentence

Let m be the number of sampled sites and $X \in \mathbb{R}^{m\times 30}$ denote the matrix of chemical element concentrations (each row represents a site and each column represents an element). Doing a principal component analysis (PCA) on X transforms it into a set of uncorrelated and high variation-loading components Z. On top of the Z, we can select $k(< 30)$ proper components with defined criteria to cover

the most variation in pollutant elements and define a composite stressor score $s \in \mathbb{R}^m$ by summing or weighting the selected raw principal components or their normalised variants:

1. **Principal component reduction** – Apply principal component analysis (PCA) to X. PCA transforms X into a set of uncorrelated components $Z = XW$, where $W \in \mathbb{R}^{30\times k}$ holds loadings of the first k principal components.

2. **Composite stress score** – Let $Z = [\, z_1, \ldots, z_k \,]$ with $z_i \in \mathbb{R}^m$ the vector of scores on the i-th principal component. Define a composite stressor score $s \in \mathbb{R}^m$ by summing or weighting the selected raw principal components:

$$s_j \;=\; \sum_{i=1}^{k} \omega_i \, z_{i,j}, \quad j \in \{1, \ldots, m\}$$

   where $z_{i,j}$ is the i-th PC score at site j and ω_i are predetermined weights (often set to 1 when components contribute equally).

After computing the composite stressor score, we can add this new information to the originally compound matrix:

$$\left[\; X \quad E \quad T \quad s \;\right] \in \mathbb{R}^{m\times(51+1)}$$

This s vector is used to rank the sites with respect to the stress level and filter the pristine reference sites where human impact is minimal or absent. Specifically, we rank sites by s and retain the least-stressed $p\%$ of the sites, create an indicator vector $I_{\text{ref}} \in \mathbb{R}^m$ where $I_{\text{ref},j} = 1$ if site j is a reference site and $I_{\text{ref},j} = 0$ otherwise.

$$\left[\; X \quad E \quad T \quad s \quad I_{\text{ref}} \;\right] \in \mathbb{R}^{m\times(52+1)}$$

To this sites with $I_{\text{ref},j} = 1$, we assume they represent the ideal taxa composition that is shaped by the given environmental conditions, supported by the minimal or absent human disturbance.

$$\left[\; X \quad E \quad T \quad s \quad I_{\text{ref}} \;\right]_{I_{\text{ref}}=1} \in \mathbb{R}^{(p\%\times m)\times(53)}$$

Therefore, in this submatrix, the X matrix only contains the minimal $p\%$ stress levels across all sites, controlling the human disturbance on the taxa composition.

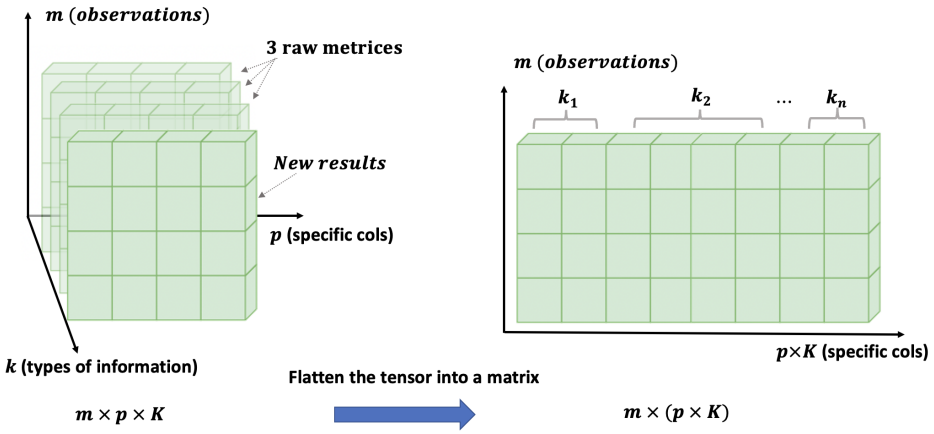

Figure 2: *Visualization of how the new information is generated and integrated into the existing matrix.*

5.2 Prepare metrics of pristine taxa composition - Cluster Analysis on References

In the matrix $\left[\; X \quad E \quad T \quad s \quad I_{\text{ref}} \;\right]_{I_{\text{ref}}=1}$, the set of taxa composition T_{ref} is assumed to be shaped by the environmental variables E_{ref}, a well-fitted regression model between the E_{ref} and T_{ref} matrices can numerically tell us how the taxa composition is multidimensionally shaped by the environmental variables.

However, considering that the $E_{\text{ref}} \in \mathbb{R}^{(p\%\times m)\times 5}$ only provides 5 environmental variables, and there are many other potentially unmeasured and unmeasurable environmental factors, it is nearly impossible to train a fully quantitative inference model that describes the below relationship well:

$$\mathcal{F} : E_{\text{ref}}^{(p\%\times m)\times 5} \rightarrow T_{\text{ref}}^{(p\%\times m)\times 16}, \quad \text{poorly fitted model}$$

To solve this issue, we can construct constrained predicted values $T_{\text{ref}}^{q}(q < 16)$ from the T_{ref} matrix, which provides limited yet information about the community structure, so that the model $\mathcal{F} : E_{\text{ref}}^{(p\%\times m)\times 5} \rightarrow T_{\text{ref}}^{(p\%\times m)\times q}$ can be trained to avoid overfitting and improve its prediction performance. One ideal way to do this information compression is to partition the reference sites into K different groups via clustering methods.

$$T_{\text{ref}}^{(p\%\times m)\times q} = \mathcal{C}_K^{(p\%\times m)\times 1} = clustering(T_{\text{ref}}^{(p\%\times m)\times 16}), \quad \text{where } q = 1$$

By the clustering analysis and merging the resulting information into the reference-base matrix, the reference-base matrix can be updated as:

$$\left[\begin{array}{cccccc} X & E & T & s & I_{\text{ref}} & \mathcal{C}_K \end{array}\right]_{I_{\text{ref}}=1} \in \mathbb{R}^{(p\%\times m)\times(53+1)}$$

Even though the C_K is computed from the clustering analysis on taxa composition matrix T_{ref}, the underlying environmental conditions(E_{ref}) are the actual drivers to lead to the clustering results, based on the foundamental assumption that "the reference taxa-composition is shaped by the environmental conditions".

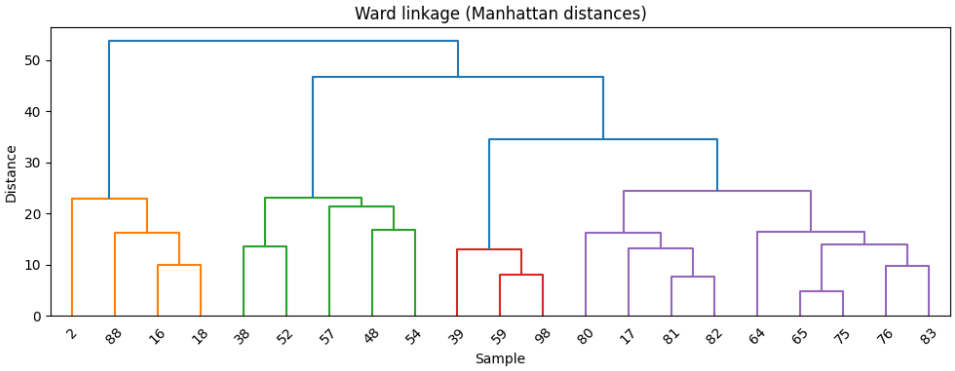

Figure 3: *An example of hierarchical clustering results on the taxa composition matrix of the references with selected clustering number K.*

5.3 Construct "ideal" taxa composition ruler of environmental factors - Fit a Discriminant Function

At this stage, there are constrained taxa composition information - cluster labels $\mathcal{C}_K$ that can be used as response variables in training the environmental-taxa composition regression. Specifically, a discriminant function can be fitted here:

$$\mathcal{F}_{\text{dis}} : E_{\text{ref}}^{(p\%\times m)\times 5} \rightarrow \mathcal{C}_K^{(p\%\times m)\times 1}$$

This discriminant function $\mathcal{F}_{\text{dis}}$ fitted on the reference sites tells us how the environmental variables - E roughly shape the taxa composition by assigning each site to one of the taxa composition groups $\mathcal{C}_K$.

During the training stage, the reference sites are partitioned into the K taxa composition groups, helping to fix the group positions in taxa composition space with the pristine taxa composition part in each group. **However, it does not mean there is only pristine taxa composition in each cluster. When human disturbance appears, the pristine taxa composition should be shifted to a new position in the taxa composition space, which is how the disturbed sites look like in the same taxa composition space.**

Therefore, these reference sites are partitioned (by clustering) into different clusters to play the role of 'ideal' metrics on a ruler of taxa composition (by discriminant function), this ruler measures the 'ideal' taxa composition structure that a site should have given its environmental conditions.

An imaginable scenario is that, when we use the fitted $\mathcal{F}_{\text{dis}}$ as a ruler to measure the taxa composition of sites that are effected by human disturbance, the measured 'ideal' taxa composition is not equal to the truly observed taxa composition. And this difference in taxa composition is caused by the human disturbance, which was measured by the sediment contamination assessment in the previous step.

5.4 Mark the 'ideal taxa composition' for disturbed sites - Apply the Discriminant Function

Given the fitted discriminant function $\mathcal{F}_{\text{dis}}$, we can classify the rest $1-p\%$ of the sites into the taxa composition groups, where reference sites with similar environmental conditions are already assigned into.

Because the clustering analysis was done on the reference sites, the known information on the disturbed sites should look like:

$$\left[\begin{array}{ccccc} X & E & T & s & I_{\text{ref}} \end{array}\right]_{I_{\text{ref}}=0} \in \mathbb{R}^{((1-p\%)\times m)\times(53)}$$

After applying the discriminant function on these disturbed sites, we can know their environmental-deterministic taxa composition groups, $\mathcal{C}_K^{((1-p\%)\times m)\times 1}$. It expands the disturbed-base matrix to:

$$\left[\begin{array}{cccccc} X & E & T & s & I_{\text{ref}} & \hat{\mathcal{C}}_K \end{array}\right]_{I_{\text{ref}}=0} \in \mathbb{R}^{((1-p\%)\times m)\times(53+1)}$$

Compare it with the reference-base matrix, we can see that the sites having the same taxa composition cluster $\mathcal{C}_K$ are now comparable with the control of environmental variables E.

To the i th site in the matrix:

$$\left[\begin{array}{cccccc} X & E & T & s & I_{\text{ref}} & \mathcal{C}_K \end{array}\right]_{I_{\text{ref}}=1} \in \mathbb{R}^{(p\%\times m)\times(53+1)}$$

If the site has $\mathcal{C}_{K_i} = \hat{\mathcal{C}}_{K_j}$, then the i th reference site is comparable with the disturbed site j th site in the taxa composition space with the control of environmental conditions. The difference in their taxa composition, $\delta T_{i,j}$, is caused by the human disturbance, $\delta X_{i,j}$, between the two sites.

$$\mathcal{C}_{K_i} = \hat{\mathcal{C}}_{K_j} \Rightarrow E_{\text{ref},i}^{(1\times 5)} \approx E_{\text{dis},j}^{(1\times 5)} \Rightarrow \delta T_{i,j} = \mathcal{F}_{reg}(\delta X_{i,j})$$

Therefore, the sites within the same taxa composition group will be used to fit the regression model - $\delta T_{i,j} = \mathcal{F}_{reg,k}(\delta X_{i,j})$, and these completed groups can be found through the merging-dismantle process of the two base matrices.

Merging the reference-base matrix and the disturbed-base matrix:

$$\text{stack}\left(\left[\begin{array}{cccccc} X & E & T & s & I_{\text{ref}} & \mathcal{C}_K \end{array}\right]_{I_{\text{ref}}=1}, \left[\begin{array}{cccccc} X & E & T & s & I_{\text{ref}} & \mathcal{C}_K \end{array}\right]_{I_{\text{ref}}=0}\right)$$

$$\Rightarrow \left[\begin{array}{cccccc} X & E & T & s & I_{\text{ref}} & \hat{\mathcal{C}}_K \end{array}\right]$$

Split the merged matrix into K submatrices, where each submatrix contains the same cluster label $\mathcal{C}_k$:

$$\left[\begin{array}{cccccc} X & E & T & s & I_{\text{ref}} & \hat{\mathcal{C}}_K \end{array}\right] = \begin{cases} \left[\begin{array}{cccccc} X & E & T & s & I_{\text{ref}} & \mathcal{C}_1 \end{array}\right] & \text{if } \mathcal{C}_K = 1 \\ \left[\begin{array}{cccccc} X & E & T & s & I_{\text{ref}} & \mathcal{C}_2 \end{array}\right] & \text{if } \mathcal{C}_K = 2 \\ \vdots & \vdots \\ \left[\begin{array}{cccccc} X & E & T & s & I_{\text{ref}} & \mathcal{C}_K \end{array}\right] & \text{if } \mathcal{C}_K = K \end{cases}$$

Within each submatrix, $\left[\begin{array}{ccccc} X & E & T & s & I_{\text{ref}} & \mathcal{C}_k \end{array}\right]$, we will numerically measure the difference in the taxa composition between the degraded sites and the reference sites, δT_k, this distance in taxa composition will be explained by the relative stress level of each site, δX_k.

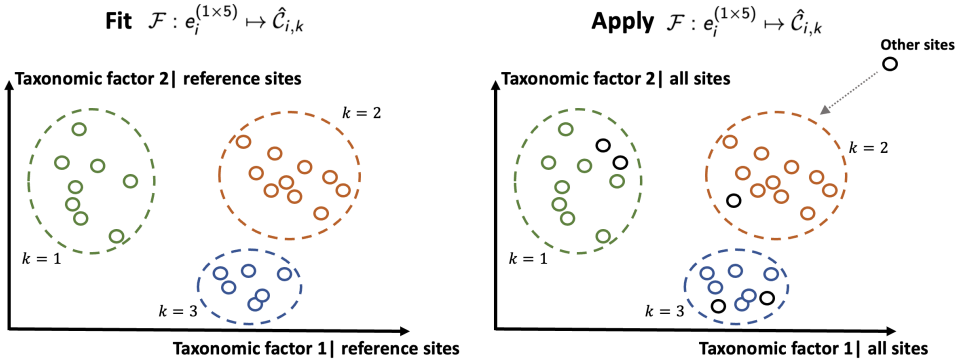

Figure 4: *Visualization of the fitting and application of the discriminant function that assign disturbed sites to the environmentally-determined taxa composition groups. to correct the symbols in the figure*

5.5 Measure the difference from 'pristine' to 'true' taxa composition - Multivariate Gaussian Deviation Index

Within each taxa-composition group $\mathcal{C}_k$, let $\mathcal{R}_k$ denote the set of reference sites ($I_{\text{ref}} = 1$) and $\mathcal{D}_k$ the set of disturbed sites ($I_{\text{ref}} = 0$). We construct a site-level deviation metric that quantifies how far a site's observed community is from the pristine expectation of its group while controlling for environmental setting via $\mathcal{C}_k$.

Because taxa compositions are multivariate and often compositional/zero-inflated, we first work on a transformed scale using the Hellinger transformation:

$$\phi_{\text{Hel}} : \mathbb{R}^{16}_{\geq 0} \to \mathbb{R}^{16}, \quad \phi_{\text{Hel}}(\mathbf{t}) = \left(\sqrt{\frac{t^{(1)}}{\sum_{\ell=1}^{16} t^{(\ell)}}}, \ \ldots, \ \sqrt{\frac{t^{(16)}}{\sum_{\ell=1}^{16} t^{(\ell)}}} \right).$$

This transformation converts each taxon abundance to the square root of its relative abundance, reducing the influence of highly dominant taxa while preserving ecological distance relationships. All subsequent quantities are computed on $\phi_{\text{Hel}}(T)$; to simplify notation we overwrite $T \leftarrow \phi_{\text{Hel}}(T)$.

There are other transformations that may be preferred depending on the context (e.g., log-ratio transforms, or raw counts), the Hellinger transformation is tentative and can be replaced as needed.

After the transformation, for group k, compute the reference centroid and covariance

$$\boldsymbol{\mu}_k \ = \ \frac{1}{|\mathcal{R}_k|} \sum_{i \in \mathcal{R}_k} T_i, \qquad \boldsymbol{\Sigma}_k \ = \ \text{Cov}\{T_i : i \in \mathcal{R}_k\} + \lambda I_{16},$$

where $\lambda > 0$ is a small ridge term to ensure invertibility and numerical stability, and I_{16} is the 16×16 identity matrix. These parameters $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ define a multivariate Gaussian-like distribution in the 16-dimensional taxa space, representing the *pristine community cloud* for group k. Under this view, each reference site is a draw from $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, and the geometric shape of this cloud is an ellipsoid whose orientation and size are determined by $\boldsymbol{\Sigma}_k$.

5.5.1 Value-based Measurement: Z-score Community Index (ZCI)

For any site j in group k (reference or disturbed), define the multivariate standardized deviation from the pristine centroid as the Mahalanobis distance:

$$\text{ZCI}_{k,j} \ = \ \sqrt{(T_j - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} (T_j - \boldsymbol{\mu}_k)}.$$

This measures how far T_j lies from the center of the pristine Gaussian cloud, in units that account for both taxon-specific variability and cross-taxon correlations. It effectively reduces the 16-dimensional deviation vector to a *single scalar score* while preserving the anisotropic geometry of the reference distribution.

When a diagonal approximation is preferred, use the "sum of squared z-scores" variant:

$$\text{ZCI}^{(\text{diag})}_{k,j} \ = \ \sqrt{\sum_{\ell=1}^{16} \left(\frac{T_j^{(\ell)} - \mu_k^{(\ell)}}{\sigma_k^{(\ell)}} \right)^2},$$

where $\sigma_k^{(\ell)}$ is the reference standard deviation of taxon ℓ in group k (robust alternatives such as median absolute deviation may also be used). This ignores inter-taxon correlations, treating the pristine cloud as an axis-aligned hypersphere, which can be more stable when the number of reference sites is small relative to the number of taxa.

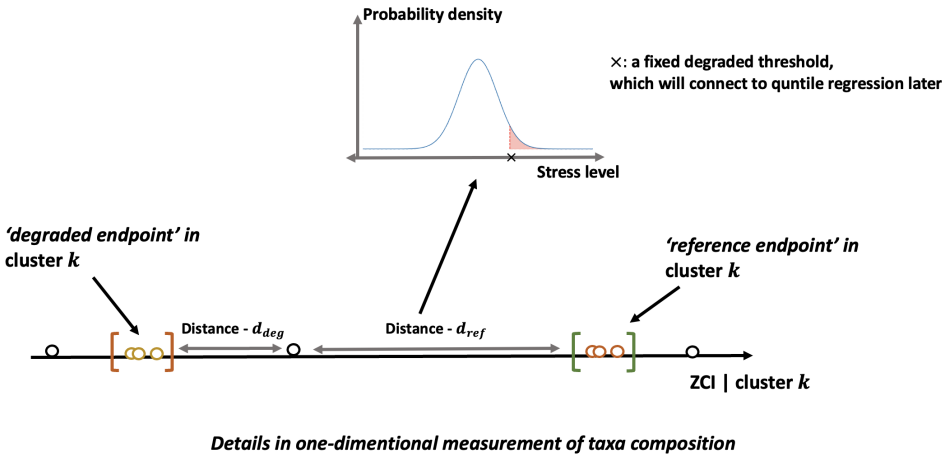

Figure 5: *Visualization of the details of taxa community structure differences measured in one-dimension ZCI.*

5.5.2 Vector-based Measurement: multi-dimensional ZCI

The scalar $\mathrm{ZCI}_{k,j}$ summarizes deviation magnitude but discards the *direction* of change in community composition. To retain more structure, the same Gaussian framework can be used to construct a multi-dimensional ZCI:

1. **Whitening of deviations** [6] **:** For each site j in group k, compute the whitened deviation vector

$$\tilde{T}_{k,j} = \boldsymbol{\Sigma}_k^{-1/2}(T_j - \boldsymbol{\mu}_k),$$

   where $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are estimated from the reference sites. Denote the matrix of whitened deviations for *reference* sites as $\tilde{T}_{k,\mathrm{ref}} \in \mathbb{R}^{n_{\mathrm{ref},k} \times 16}$. In this whitened space, the reference cloud is isotropic and centered at the origin.

2. **PCA fitted on whitened reference sites:** Perform principal component analysis (PCA) on $\tilde{T}_{k,\mathrm{ref}}$ to obtain the loading matrix V_k. Retain the first d principal axes $V_{k,(1:d)}$, where $d = 2$ gives a two-dimensional ZCI.

3. **PCA applied to disturbed sites:** For each disturbed site j, compute its whitened deviation $\tilde{T}_{k,j}$ using the *same* $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k^{-1/2}$ from the reference sites, and project it onto the retained principal axes:

$$\left(\mathrm{ZCI}_{k,j}^{(1)}, \mathrm{ZCI}_{k,j}^{(2)}\right) = \tilde{T}_{k,j}\, V_{k,(1:2)}.$$

These coordinates preserve both d magnitude and orientation of deviation in the most informative subspace of the pristine community cloud, enabling more nuanced comparisons between sites that have similar scalar ZCI values but differ in the *type* of community shift. The scalar $\mathrm{ZCI}_{k,j}$ can be recovered as the Euclidean norm of these coordinates.

[6] Whitening means: Centering (subsracting μ_k), rescaling and rotating so that the reference covariance becomes the identity matrix. Knowing that $\Sigma_k = \frac{1}{|T|-1}(T-\mu)^T(T-\mu)$, replacing the T with $\tilde{T} = \Sigma_k^{-1/2}(T-\mu)$, the new covariance matrix $\tilde{\Sigma}_k$ becomes $\frac{1}{|\tilde{T}|-1}(\tilde{T}-\tilde{\mu})^T(\tilde{T}-\tilde{\mu}) = I$. T and μ are both matrices.

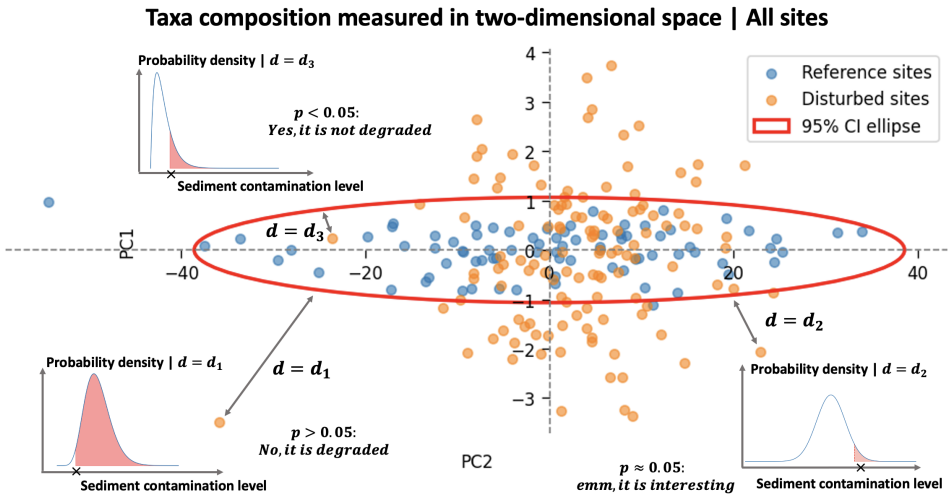

Figure 6: *Visualization of the details of taxa community structure differences measured in two-dimension ZCI.*

5.5.3 Direction, interpretation, and optional 0–100 scaling.

By construction, smaller values indicate communities closer to the pristine expectation for their environment; larger values indicate stronger deviation (putative impact). For reporting, we optionally map ZCI to a condition scale where larger is better:

$$\mathrm{ZCI}^{\star}_{k,j} \;=\; 100\,\big(1 - \widehat{F}_k(\mathrm{ZCI}_{k,j})\big),$$

with $\widehat{F}_k$ the empirical CDF of ZCI computed from *reference* sites in group k. Under this calibration, reference sites cluster near higher scores (closer to 100), while progressively disturbed sites trend toward 0.

5.6 Build the ZCI-based inference model of stress level – Piecewise Quantile Regression Model

The ZCI score reflects the degree of deviation in taxa composition from the pristine expectation within each group k, given the same environmental context. Because both the stress level and the community condition are influenced by a range of measured and unmeasured factors, it is reasonable to model the conditional distribution of the stress level *given* the community departure. This regression is used only as a statistical association to infer likely stress levels from observed ZCI values and does **not** imply a causal relationship between stress and community departure.

Within each group k, we relate the relative stress level to the measured ZCI distance:

$$\delta X_{k,j} \;=\; \mathcal{F}_{\mathrm{reg,k}}\big(\,\mathrm{ZCI}_{k,j}\,\big) + \varepsilon_{k,j},$$

where $\delta X_{k,j}$ may be the composite stressor score s_j (or selected PCs of X) centered at the group-k reference level (e.g., $s_j - \mathrm{median}\{s_i : i \in \mathcal{R}_k\}$).

Considering the potential nonlinearity of this association and the effects brought by hidden factors, non-linear quantile regression models can better reveal the underlying patterns and are chosen to fit the regression function $\mathcal{F}_{\mathrm{reg,k}}$.

With this setting, $\mathcal{F}_{\mathrm{reg},k}$ is the conditional τ-quantile of $\delta X_{k,j}$ given $z_{k,j} := \mathrm{ZCI}_{k,j}$, modeled by a continuous piecewise-linear (hinge) form:

$$Q^{(k)}_{\delta X|Z}(\tau \mid z) \;=\; f_{k,\tau}(z) \;=\; \beta^{(k)}_{0,\tau} \;+\; \beta^{(k)}_{1,\tau}\, z \;+\; \sum_{m=1}^{M} \gamma^{(k)}_{m,\tau}\,(z-\kappa_m)_+, \qquad (z-\kappa_m)_+ := \max\{0,\, z-\kappa_m\},$$

where $\kappa_1 < \cdots < \kappa_M$ are fixed breakpoints on the ZCI axis (e.g., selected by domain knowledge, grid search, or sample quantiles of z). The parameters are estimated by minimizing the check loss:

$$\widehat{\boldsymbol{\theta}}^{(k)}_{\tau} \;\in\; \arg\min_{\boldsymbol{\theta}} \sum_{j \in \mathcal{C}_k} \rho_\tau\Big(\delta X_{k,j} - Q^{(k)}_{\delta X|Z}(\tau \mid z)\Big), \quad \rho_\tau(u) = u\{\tau - \mathbf{1}(u<0)\}.$$

The fitted conditional quantile is

$$\widehat{Q}^{(k)}_{\delta X|Z}(\tau \mid z) = \widehat{\beta}^{(k)}_{0,\tau} + \widehat{\beta}^{(k)}_{1,\tau}\, z + \sum_{m=1}^{M} \widehat{\gamma}^{(k)}_{m,\tau}\, (z - \kappa_m)_+.$$

5.6.1 Hypothesis testing for site degradation – Quantile-based threshold inference

In many applications, a binary classification of a site as "degraded" or "non-degraded" is more actionable than estimating its exact stress level. This decision problem can be formulated as a one-sided hypothesis test, conditioning on the observed community departure (ZCI):

- **Step 1 – Define degradation threshold.** For each group k, choose a stress threshold x^*_k (e.g., a regulatory limit or an ecologically relevant benchmark) that separates degraded from non-degraded sites.
- **Step 2 – Predict conditional stress distribution.** For a site j with observed $\mathrm{ZCI}_{k,j} = z$, use the fitted quantile regression model $Q^{(k)}_{\delta X|Z}(\tau \mid z)$ over a grid of quantile levels $\tau \in (0, 1)$ to approximate the conditional distribution $F^{(k)}_{\delta X|Z}(x \mid z)$. This is done by inverting the quantile function across τ.
- **Step 3 – Compute p-value for degradation.** The hypothesis test is:

  $$H_0 : \delta X_{k,j} \le x^*_k \quad \text{vs.} \quad H_a : \delta X_{k,j} > x^*_k.$$

  The conditional p-value is then

  $$p = 1 - F^{(k)}_{\delta X|Z}(x^*_k \mid z)\,,$$

  which represents the probability, given the site's ZCI, that the stress level exceeds the degradation threshold.

If p is below a chosen significance level α (e.g., 0.05), the site is classified as degraded; otherwise, it is classified as non-degraded. This approach converts the regression output into a probabilistic decision rule while controlling for environmental setting via the group-specific model.

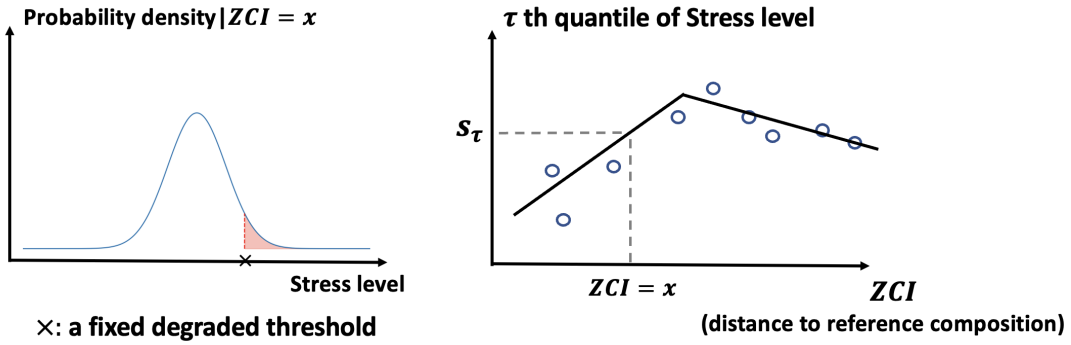

Figure 7: *A pre-fixed degradation threshold on the conditional stress level distribution and the correspondingly predicted quantile value* $\hat{s}_\tau$

6 Preliminary exploration

In this section, I implemented a simplified framework of Jian's analysis of the ZCI-stress score relationship using the quantile regression.

1. **Collect comparable data.**

   I currently have three datasets: zoobenthic community data (311×16), chemical data (104×30), and environmental data (289×7). These datasets were merged by station ID, resulting in a combined dataset with 104 rows and 53 columns, containing all three types of data. Column indices were structured by data type to improve readability and consistency for future processing.

2. **Pre-process taxa data.**

   The zoobenthic dataset has issues such as small sample size and potential outliers. I used the IQR method to detect outliers and then applied octave transformation, as suggested in Jian's analysis, to reduce their impact. The transformed data showed fewer outliers and a more even distribution.

3. **Assess sediment contamination.**

   Instead of standardization, I applied log-transformation to the chemical data to reduce dominance by high-value variables. Then I conducted PCA and selected principal components based on variance explained, pollutant specificity, and balanced loadings. These selected components were normalized and summed (with attention to loading directions) to produce a composite "SumReal" score reflecting sediment contamination. This score was added to the dataset as an indicator of stress level.

4. **Identify reference and degraded sites.**

   Sites were classified by their stress scores. The bottom 20% were considered minimally disturbed (reference sites), and the top 20

5. **Cluster reference sites by community composition.**

   Since taxa composition varies even among reference sites, clustering was applied to identify dominant community patterns under undisturbed conditions. These clusters represent "normal" taxa structures, each likely shaped by distinct environmental conditions.

6. **Build a discriminant model for habitat classification.**

   A discriminant model was trained to predict cluster membership using environmental variables from reference sites. This model was applied to all sites, assigning each to one of the community clusters. This setup allows comparisons between reference and degraded sites within the same habitat type to define "best" and "worst" endpoints.

7. **Construct endpoints and compute ZCI via ordination.**

   Within each cluster, endpoints were constructed using mean taxa abundances from the most reference-like and most degraded sites. These endpoints were used in Bray-Curtis ordination to assign scores to each site, scaled between 0 (worst) and 1 (best) to compute the Zoobenthic Condition Index (ZCI). I do not yet fully understand the ordination method used, and I plan to explore whether a vector-based ZCI might provide a more accurate assessment.

8. **Evaluate the ZCI vs SumRel relationship by quantile regression.**

   I plotted ZCI against SumRel and applied quantile regression to examine their relationship.

References

[1] U.S. Environmental Protection Agency and Environment Canada. State of the great lakes 2007: Status and trends of great lakes shoreline hardening. Technical report, U.S. Environmental Protection Agency, Great Lakes National Program Office, 2007. Accessed: 2025-07-15.

[2] Gavin F. Birch. A review and critical assessment of sedimentary metal indices used in determining the magnitude of anthropogenic change in coastal environments. *Science of The Total Environment*, 823:153623, 2022.

[3] Stephen Birch and Amiram Gafni. Being naughty about nice? questioning the methods used to maximize health gains from nhs resources. *Health Economics, Policy and Law*, 2(2):217–221, 2007.

[4] Brian S. Cade and Barry R. Noon. A gentle introduction to quantile regression for ecologists. *Frontiers in Ecology and the Environment*, 1(8):412–420, 2003.

[5] Aurea C. Chiaia-Hernández, Carmen Casado-Martinez, Pablo Lara-Martin, and Thomas D. Bucheli. Sediments: sink, archive, and source of contaminants. *Environmental Science and Pollution Research*, 29:85761–85765, 2022.

[6] Jonathan P. Daily, Nathaniel P. Hitt, David R. Smith, and Craig D. Snyder. Experimental and environmental factors affect spurious detection of ecological thresholds. *Ecology*, 93(1):17–23, 2012.

[7] Susan P. Davies and Susan K. Jackson. The biological condition gradient: A descriptive model for interpreting change in aquatic ecosystems. *Ecological Applications*, 16(4):1251–1266, August 2006.

[8] Mélanie Desrosiers, Bernadette Pinel-Alloul, and Charlotte Spilmont. Selection of macroinvertebrate indices and metrics for assessing sediment quality in the st. lawrence river (qc, canada). *Water*, 12(12):3335, 2020.

[9] C. S. Holling. Resilience and stability of ecological systems. *Annual Review of Ecology and Systematics*, 4:1–23, 1973.

[10] Markus Horning. Constraint lines and performance envelopes in behavioral physiology: the case of the aerobic dive limit. *Frontiers in Physiology*, 3:381, 2012.

[11] Qi Huang, Hanze Zhang, Jiaqing Chen, and Mengying He. Quantile regression models and their applications: A review. *Journal of Biometrics & Biostatistics*, 8(3):354, 2017.

[12] Garry D. Peterson, Stephen R. Carpenter, William A. Brock, Jonathan M. Hanson, Joel Carson, Lynne Haskins, Milos Holmgren, Tim Eason, Christine Engels, et al. Ecological thresholds: The key to successful environmental management or an important concept with no practical application? *Ecosystems*, 9(1):1–13, 2006.

[13] Rebecca Spake, Martha Paola Barajas-Barbosa, Shane A. Blowes, Diana E. Bowler, Corey T. Callaghan, Magda Garbowski, Stephanie D. Jurburg, Roel van Klink, Lotte Korell, Emma Ladouceur, Roberto Rozzi, Duarte S. Viana, Wu-Bing Xu, and Jonathan M. Chase. Detecting thresholds of ecological change in the anthropocene. *Annual Review of Environment and Resources*, 47(Volume 47, 2022):797–821, 2022.

[14] U.S. Environmental Protection Agency. Great lakes facts and figures, 2024. Accessed: 2025-07-15.

[15] Jian Zhang. Zoobenthic community composition and chironomidae (diptera) mouthpart deformities as indicators of sediment contamination in the lake huron-lake erie corridor of the laurentian great lakes. Master's thesis, University of Windsor, Windsor, Ontario, Canada, 2008. A thesis for the degree of Master of Science.

7 Appendix

7.1 Tables

Table 6: Environmental Variables and Their Explanations

Variable Name	Explanation
Site_ID	Unique identifier for each sampling site
Lake_or_River	Indicates whether the site is in a lake or river
Latitude	Geographic latitude coordinate
Longitude	Geographic longitude coordinate
Total_Organic_Carbon_LOI_percent	Total organic carbon content (loss on ignition, as %)
Water_Depth_m	Water depth at the sampling location (meters)
Water_Temperature_C	Water temperature in degrees Celsius
Dissolved_Oxygen_Concentration_mgL	Dissolved oxygen concentration in milligrams per liter
Median_Particle_Size_Phi	Median particle size of sediment (Phi scale)

Table 7: Taxonomic Variables and Their Explanations

Taxonomic Group	Explanation
Oligochaeta	Aquatic segmented worms
Nematoda	Roundworms
Chironomidae	Non-biting midges (larvae)
Ceratopogonidae	Biting midges
Hexagenia	Mayfly genus (larvae)
Caenis	Mayfly genus (larvae)
Hydropsychidae	Net-spinning caddisflies
Other Trichoptera	Other caddisfly families
Amphipoda	Small crustaceans (e.g., scuds)
Dreissena	Zebra/quagga mussels
Acari	Aquatic mites
Hydrozoa	Small predatory animals (hydroids)
Hirudinea	Leeches
Turbellaria	Flatworms
Gastropoda	Snails and slugs
Sphaeriidae	Fingernail clams

Table 8: Stressors and Their Explanations

Chemical/Contaminant	**Explanation**
Al	Aluminum (trace metal)
As	Arsenic (toxic element)
Bi	Bismuth (trace element)
Ca	Calcium (major element, hardness)
Cd	Cadmium (toxic metal)
Co	Cobalt (trace element)
Cr	Chromium (trace metal)
Cu	Copper (trace metal, micronutrient)
Fe	Iron (major element, micronutrient)
Hg	Mercury (highly toxic metal)
K	Potassium (major element)
Mg	Magnesium (major element, hardness)
Mn	Manganese (trace element)
Na	Sodium (major element)
Ni	Nickel (trace metal)
Pb	Lead (toxic metal)
Sb	Antimony (trace element)
V	Vanadium (trace element)
Zn	Zinc (trace metal, micronutrient)
%OC	Percent organic carbon
1245-TCB	1,2,4,5-Tetrachlorobenzene (organic pollutant)
1234-TCB	1,2,3,4-Tetrachlorobenzene (organic pollutant)
QCB	Quintachlorobenzene (organic pollutant)
HCB	Hexachlorobenzene (organic pollutant)
OCS	Octachlorostyrene (organic pollutant)
p,p'-DDE	DDT breakdown product
p,p'-DDD	DDT breakdown product
mirex	Organochlorine insecticide
Heptachlor Epoxide	Organochlorine pesticide breakdown product
total PCB	Total polychlorinated biphenyls

Table 9: Explanation, Habitat, and Survival Rate in Fast/Slow Water for Each Taxon

Taxa	Explanation	Habitat Type	Survival Rate (Fast/Slow Water)
Oligochaeta	Aquatic segmented worms	Both (lentic/lotic)	Moderate/High
Nematoda	Roundworms	Both (lentic/lotic)	Moderate/High
Chironomidae	Non-biting midge larvae	Both (lentic/lotic)	Moderate/High
Ceratopogonidae	Biting midge larvae	Both (lentic/lotic)	Moderate/Moderate
Hexagenia	Burrowing mayflies	Lentic (lakes/ponds)	Low/High
Caenis	Small mayflies	Both (lentic/lotic)	Low/Moderate
Hydropsychidae	Net-spinning caddisflies	Lotic (streams/rivers)	High/Low
Other Trichoptera	Other caddisflies	Both (lentic/lotic)	Moderate/Moderate
Amphipoda	Small crustaceans	Both (lentic/lotic)	Moderate/High
Dreissena	Zebra/quagga mussels	Lentic, large rivers	Low/High
Acari	Aquatic mites	Both (lentic/lotic)	Moderate/High
Hydrozoa	Predatory invertebrates	Lentic (lakes/ponds)	Low/High
Hirudinea	Leeches	Both (lentic/lotic)	Low/High
Turbellaria	Flatworms	Both (lentic/lotic)	Low/High
Gastropoda	Aquatic snails	Both (lentic/lotic)	Low/High
Sphaeriidae	Fingernail clams	Both (lentic/lotic)	Low/High

7.2 Index-based methods for quantitative stress metrics

Introduce the index-based methods for quantitative stress metrics

PCA is mainly for exploratory work, like finding patterns, associations in the data. It may be used to score the sediment contamination level, but only comparing relative abundance of stressor elements among the sites, lacking a little standards to accurately assess the contamination level. Index-based methods can be used to quantitatively assess the sediment contamination level. Such method needs more empirical and theoretical support than PCA, but it offers a more standardized way to assess the contamination level, which is theoretically more accurate.

Combining the PCA and index-based methods is a good idea to explore the sediment contamination level. I am thinking about what do they mean in contamination assessment, and how to combine them with a reasonable logical flow.

7.3 Principal Component Analysis based methods to explore contaminant association and patterns

Discuss the reason why and how PCA can be used to assess the sediment contamination level and reflect the aquatic ecological integrity.

In SVD, a data matrix X is decomposed as $X = U\Sigma V^T$, where U and V are orthogonal matrices and Σ is a diagonal matrix of singular values. For PCA, the principal components correspond to the directions of maximum variance, which are given by the right singular vectors in V. By incorporating weights into the data matrix, Weighted PCA modifies the SVD process to emphasize certain observations(row or column), allowing for more flexible dimensionality reduction tailored to the importance of each data point.

Given a data matrix $X \in \mathbb{R}^{m\times n}$, the SVD decomposes X as $X = U\Sigma V^T$, where:

- $U \in \mathbb{R}^{m\times m}$ contains the left singular vectors (columns),
- $\Sigma \in \mathbb{R}^{m\times n}$ is a diagonal matrix of singular values,
- $V \in \mathbb{R}^{n\times n}$ contains the right singular vectors (columns).

To solve for U and V:

1. Compute X^TX and find its eigenvectors and eigenvalues. The eigenvectors form the columns of V.
2. It can be proven that $\mu_i^T = \frac{Xv_i}{\sigma_i}$ is the i-th column of U, where σ_i is the i-th singular value(square root of eigenvalue λ_i).

3. The singular values in Σ are the square roots of the nonzero eigenvalues from either XX^T or X^TX.

Mathematically, given a matrix X of shape (m, n), the SVD can be expressed as:

$$(X^TX)V = V\Lambda, \quad \Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$$

According to spectral theorem, the eigenvalues λ_i are non-negative and can be ordered as $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \geq 0$. V is an orthogonal matrix, meaning $V^TV = I$, where I is the identity matrix. It gives:

$$(X^TX) = V\Lambda V^T$$

To a centered sample matrix of size m with n features, its covariance matrix is $\frac{1}{m-1}(X^TX)$. Via the eigenvalue decomposition, the variation in it can be expressed by the paris of eigenvalues and eigenvectors, which are a series of rank-one matrices that carry different and independent information:

$$\frac{1}{m-1}(X^TX) = \frac{1}{m-1}\sum_{i=1}^{n}\lambda_i v_i v_i^T$$

Therefore, when columns in X represent different features, the columns v_i of V are the principal components that have its values as the linear combinations of the original features, and the scaled eigenvalues $\frac{\lambda_i}{m-1}$ as the amount of variation explained by each principal component.

Weighted PCA leverages the Singular Value Decomposition (SVD) to extract principal components from weighted data.

Preparing to explain why this ordinal decomposition is not good for our case, and how we can improve it by considering the weights for different chemical elements and filtering out the from the PCs

7.4 Hierarchical Clusering analysis for Zoobenthic Communicty Indicator Construction

Application of Hierarchical Clustering in Stressor-Community Analysis Overview Hierarchical clustering offers a data-driven approach to group sampling sites based on environmental or community-level similarities. Unlike partitioning methods, hierarchical clustering builds a nested tree (dendrogram) that captures the progressive grouping of observations based on their dissimilarities. In the context of my program, hierarchical clustering can be leveraged to define natural environmental classes prior to modeling the biological response to stressors.

Let each observation $x_i \in \mathbb{R}^p$ denote a site with p-dimensional attributes (e.g., environmental variables). The dissimilarity between two observations x_i and $x_{i'}$ is measured by a distance function, often chosen as the squared Euclidean distance:

$$d(x_i, x_{i'}) = \sum_{j=1}^{p}(x_{ij} - x_{i'j})^2$$

Based on the distance metric, we can define the similarity measure on variable and observation levels as the following:

$$\text{Variable level:} \quad d_j(x_{ij}, x_{i'j}) = (x_{ij} - x_{i'j})^2, \quad j = 1, \ldots, p$$

$$\text{Observation level:} \quad d(x_i, x_{i'}) = \sum_{j=1}^{p}(x_{ij} - x_{i'j})^2$$

To cluster level D_C, we can define other alternatives to the dissimilarity measures, commonly including:

$$D_{SL}(G, H) = \min_{i \in G,\, i' \in H} d(x_i, x_{i'}) \quad \text{(Single Linkage)}$$

$$D_{CL}(G, H) = \max_{i \in G,\, i' \in H} d(x_i, x_{i'}) \quad \text{(Complete Linkage)}$$

$$D_{GA}(G, H) = \frac{1}{|G||H|}\sum_{i \in G}\sum_{i' \in H} d(x_i, x_{i'}) \quad \text{(Average Linkage)}$$

- **Use in the Program:**

  - Hierarchical clustering is used to categorize sampling sites into clusters with similar environmental conditions before building predictive models linking taxonomic composition to stressor levels.
  - This enables a two-stage analysis:
    * Use environmental variables to form environmental clusters (reference site classification).
    * Model stressor-community relationships within or across these clusters to control for natural variation.

- **Advantages:**
  - Does not require pre-specification of the number of clusters.
  - Produces a dendrogram showing how clusters are formed step-by-step.
  - Enables ecological interpretation of clusters through tree visualization.
  - Allows separation of natural variation from anthropogenic stress impacts.

- **Model-Based Interpretation:**
  - Assume that each cluster k corresponds to a latent distribution $p_k(x)$, with the overall mixture model:
$$p(x) = \sum_{k=1}^{K} \pi_k p_k(x)$$
  - Each observation is generated as $x \sim p_k(x)$ conditional on cluster membership k, providing a statistical grounding for hierarchical clustering in unsupervised structure discovery.

waiting to add more specific details after applying the clustering on the data and getting some preliminary results.

7.5 Piecewise Quantile Regression for Threshold Determination

Let $m_\tau(x; \boldsymbol{\beta}_\tau, \boldsymbol{\alpha}_\tau)$ be the τth quantile of the conditional distribution of the ecological response given environmental condition. Then we define:

$$y_\tau = m_\tau(x; \boldsymbol{\beta}_\tau, \boldsymbol{\alpha}_\tau) + \varepsilon_\tau.$$

The form is similar to the linear regression model, but there are no restrictions on the distribution of error term ε_τ, which means the error terms can be heteroscedastic(non-constant variance).

Then the **PQRM** with two breakpoints is defined as:

$$m_\tau(x_i; \boldsymbol{\beta}_\tau, \boldsymbol{\alpha}_\tau) = \begin{cases} \beta_{0\tau} + \beta_{1\tau} x_i & \text{for } x_i \le \alpha_{1\tau} \\ \beta_{0\tau} + \beta_{1\tau} x_i + \beta_{2\tau}(x_i - \alpha_{1\tau}) & \text{for } \alpha_{1\tau} < x_i \le \alpha_{2\tau} \\ \beta_{0\tau} + \beta_{1\tau} x_i + \beta_{2\tau}(x_i - \alpha_{1\tau}) + \beta_{3\tau}(x_i - \alpha_{2\tau}) & \text{for } x_i > \alpha_{2\tau} \end{cases} \tag{3}$$

The τ subscript indicates the quantile level. Breakpoints are spotted in the predictor space, which seems uncorrelated with the response variable and its quantile, but they are not fixed and should be estimated under different quantile levels so that the piecewise model provides a better performance on the data with such quantile level. That is, the breakpoints are also functions of the quantile level τ.

$$Y_\tau = X(\alpha_{(\tau;1)}, \alpha_{(\tau;2)}) \cdot \boldsymbol{\beta_\tau} + \boldsymbol{\varepsilon_\tau}$$

Because the model are not predicting the means but the quantiles of y's at given x's, the measurement for the model performance should no longer be the mean squared error(MSE), [7] the check function is the loss function, which is defined as:

$$\rho_\tau(u) = \begin{cases} \tau u & \text{if } u \ge 0 \\ (\tau - 1)u & \text{if } u < 0 \end{cases} \quad \text{where } u = y - \hat{y}$$

[7] If a model regresses the mean of Y, it should outperform other models in the measurement of MSE, assuming the assumptions are satisfied.

The loss function is a random variable that derivatives from Y, and it also can be viewed as a function of the prediction $\hat{y_\tau}$, which is not a RV. A specific quantile can be found by minimizing the expected loss function - $E(\rho_\tau(Y - \hat{y}_\tau))$ with respect to $\hat{y}_\tau$ across the observations:

$$q_Y(\tau) = \arg\min_{\hat{y}_\tau} E(\rho_\tau(Y - \hat{y}_\tau)) = \arg\min_{\hat{y}_\tau} \left\{ (\tau - 1) \int_{-\infty}^{\hat{y}_\tau} (y - \hat{y}_\tau) dF_Y(y) + \tau \int_{\hat{y}_\tau}^{\infty} (y - \hat{y}_\tau) dF_Y(y) \right\}$$

The expectation is no more a random variable, but a function of $(\hat{y}_\tau, y, F_Y(y))$. By computing the derivative of it with respect to $\hat{y}_\tau$, the equation is determined by $(\hat{y}_\tau, F_Y(y))$. Settting it to zero and letting q_τ the solution of $\hat{y}_\tau$ to the equation, we have:

$$0 = (1 - \tau) \int_{-\infty}^{q_\tau} dF_Y(y) - \tau \int_{q_\tau}^{\infty} dF_Y(y).$$

It gives:

$$F_Y(q_\tau) = \tau$$

Therefore, $\rho_\tau(u)$ is a valid loss function for inferring the quantile of Y. However, when building quantile regression on Y with X, we need to have enough observations on each condition of X. Becasue there is no assumption that residuals are homoscedastic, the globally minimized loss function does not necessarily bring good predictions on all conditions, which is different to the linear regression model.

To a given condition x, the quantile of $\tilde{y}$ is defined as:

$$q_{\tilde{y}|x}(\tau) = \arg\min_{\hat{y}_\tau} E(\rho_\tau(\tilde{y} - \hat{y}_\tau)|X) = \arg\min_{\hat{y}_\tau} \left\{ (\tau - 1) \int_{-\infty}^{\hat{y}_\tau} (y - \hat{y}_\tau) dF_{\tilde{y}|x}(y) + \tau \int_{\hat{y}_\tau}^{\infty} (y - \hat{y}_\tau) dF_{\tilde{y}|x}(y) \right\}$$

The expected loss function that provides quantiles over all observations is:

$$E(\rho_\tau(\tilde{y} - \hat{y}_\tau)) = \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(\tilde{y}_i - x_i^\top \beta_\tau)$$

Optimizing it to minimum with respect to β_τ gives the optimal β_τ [8]:

$$\hat{\beta}\tau = \arg\min_{\beta_\tau} E(\rho_\tau(\tilde{y} - \hat{y}_\tau))$$

When given two breakpoints, the $(\alpha_{(\tau;1)}, \alpha_{(\tau;2)})$ can be iteratively searched by Newton-Raphson method. Within each iteration, the loss function is minimized to find the optimal $\beta_{(\tau)}$ vector.

waiting to add more specific details after applying the clustering on the data and getting some preliminary results.

7.6 Synthetic Data Generation

Synthetic data generation refers to the process of artificially creating data that mimics the statistical properties and structure of real-world datasets. The core motivation is to generate data when real data are limited, imbalanced, private, or costly to obtain. Synthetic data are widely used in machine learning, data privacy preservation, and simulation-based research.

This approach can simulate various data types, including tabular (structured), image, text, and time series data. The generated data should reflect similar distributions, correlations, and interactions as the original data while maintaining flexibility and scalability.

Synthetic data generation can play a supportive role in this study by addressing several data-related challenges commonly encountered in ecological assessments:

- **Enhancing model robustness**: By generating additional synthetic observations that mimic the real data structure, we can augment the existing data pool, which helps reduce model overfitting and improve generalization when predicting stressor impacts across unsampled sites.
- **Balancing site distribution across gradients**: Many ecological datasets are imbalanced with respect to environmental gradients or stressor intensity levels. Synthetic data can help balance the representation of different ecological zones or stressor conditions, ensuring the model learns from a more evenly distributed set of scenarios.

[8] There should be an assumption of linear relation on the amount changing between x and τ quantile of y, and β is the coefficient of that relation.

- **Supporting rare condition modeling**: Stressor levels or taxa responses in extreme or undersampled regions (e.g., highly degraded or pristine sites) can be underrepresented. Synthetic data can simulate these rare cases to aid in threshold detection or to inform prediction in sparsely observed domains.
- **Facilitating model testing and validation**: Controlled synthetic datasets can be used to evaluate the behavior and sensitivity of the modeling framework under different stressor-environment-taxa scenarios, helping identify model biases or limits.
- **Exploring uncertainty and variability**: Synthetic data allow the introduction of controlled noise and variability, which is useful for evaluating how robust the model is to observational or measurement uncertainty.

A simple yet effective approach to synthetic data generation is random sampling (bootstrapping) from the observed dataset. Given a dataset $\mathcal{D} = \{x_1, \ldots, x_N\}$, where $x_i \in \mathbb{R}^p$, synthetic samples are generated by sampling with replacement:

$$x_i^* \sim \hat{F}, \quad i = 1, \ldots, N^*$$

where $\hat{F}$ is the empirical distribution of $\mathcal{D}$, and N^* is the desired number of synthetic samples.

This method preserves the multivariate structure by sampling entire rows, maintaining correlation between features. It is non-parametric and computationally efficient.

Use in this project: Random sampling can augment small or imbalanced training sets, support bootstrapped threshold estimation, and improve model robustness across environmental gradients.

More common methods for synthetic data generation are summarized in Table 11.

waiting to add more specific details after applying the clustering on the data and getting some preliminary results.

Table 10: Chemical Descriptive Statistics by Site Label

	SumReal	degraded	intermediate	reference
Al	mean	4276.423	6380.140	4319.381
	std	2888.769	5523.949	1767.861
As	mean	2.186	1.777	2.232
	std	1.602	1.290	1.041
Bi	mean	17.085	17.505	17.622
	std	10.352	10.273	9.722
Ca	mean	28180.500	33518.930	28480.714
	std	14031.433	11400.266	11870.107
Cd	mean	0.535	0.351	0.271
	std	0.649	0.202	0.233
Co	mean	4.049	4.497	3.984
	std	1.733	2.209	1.118
Cr	mean	13.254	12.830	9.007
	std	16.373	11.835	2.937
Cu	mean	16.958	18.082	12.946
	std	22.388	29.120	9.003
Fe	mean	9495.000	11246.789	9650.905
	std	5392.824	6804.654	3856.739
Hg	mean	0.474	0.324	0.196
	std	1.230	0.420	0.365
K	mean	818.927	1285.558	845.657
	std	638.053	1092.550	411.332
Mg	mean	12849.500	15204.175	12269.143
	std	6104.202	5764.037	5281.794
Mn	mean	161.228	188.900	161.905
	std	76.973	86.663	57.883
Na	mean	118.998	134.042	123.611
	std	49.081	43.693	41.021
Ni	mean	11.225	12.399	9.136
	std	8.851	8.424	3.542
Pb	mean	12.515	8.774	8.573
	std	32.312	22.204	18.750
Sb	mean	17.262	16.765	18.001
	std	11.879	13.115	13.743
V	mean	15.274	18.353	15.183
	std	7.012	9.560	4.408
Zn	mean	52.732	46.181	35.677
	std	48.896	44.586	17.938
%OC	mean	2.110	2.405	1.779
	std	1.599	1.458	0.682
1245-TCB	mean	0.906	1.201	0.555
	std	2.321	2.143	1.035
1234-TCB	mean	0.252	0.234	0.253
	std	0.257	0.240	0.332
QCB	mean	0.729	1.255	0.636
	std	1.015	3.055	0.871
HCB	mean	2.759	17.713	2.904
	std	4.291	83.487	6.011
OCS	mean	1.213	1.502	0.721
	std	3.395	3.606	1.874
p,p'-DDE	mean	0.679	0.485	0.324
	std	1.255	0.930	0.328
p,p'-DDD	mean	3.879	0.772	0.862
	std	14.634	1.039	0.923
mirex	mean	0.253	0.212	0.134
	std	0.682	0.332	0.242
Heptachlor Epoxide	mean	0.098	0.051	0.071
	std	0.235	0.250	0.211
total PCB	mean	15.137	10.705	7.715
	std	32.189	36.285	16.795

Table 11: Summary of common synthetic data generation methods and their applications.

Method	Description	Example Use Case
Random Sampling	Generate synthetic data by sampling from known statistical distributions (e.g., normal, uniform, Poisson)	Simulating sensor readings or generating test input data
SMOTE (Synthetic Minority Oversampling Technique)	Generates new instances of the minority class by interpolating between existing instances	Addressing class imbalance in classification problems
GANs (Generative Adversarial Networks)	Uses a generator and a discriminator to iteratively create highly realistic synthetic data that mimics the real distribution	Synthetic images, medical data, or tabular data with complex dependencies
VAEs (Variational Autoencoders)	Learns a probabilistic latent space from real data and generates synthetic samples by decoding random samples from that space	Generating synthetic patient records or anomaly detection
Agent-based Simulations	Models behavior of individuals or agents and their interactions to generate data over time	Simulating traffic systems or disease spread
Rule-based Generation	Applies predefined rules and logical templates to construct synthetic data	Automated form testing, fake transactions for system validation