Detecting and Modeling Changes in Stream Nutrient Dynamics

JSM 2024, Portland, Oregon

Thomas J. Fisher

Joint work with Hannah Waler

Miami University, Oxford, OH

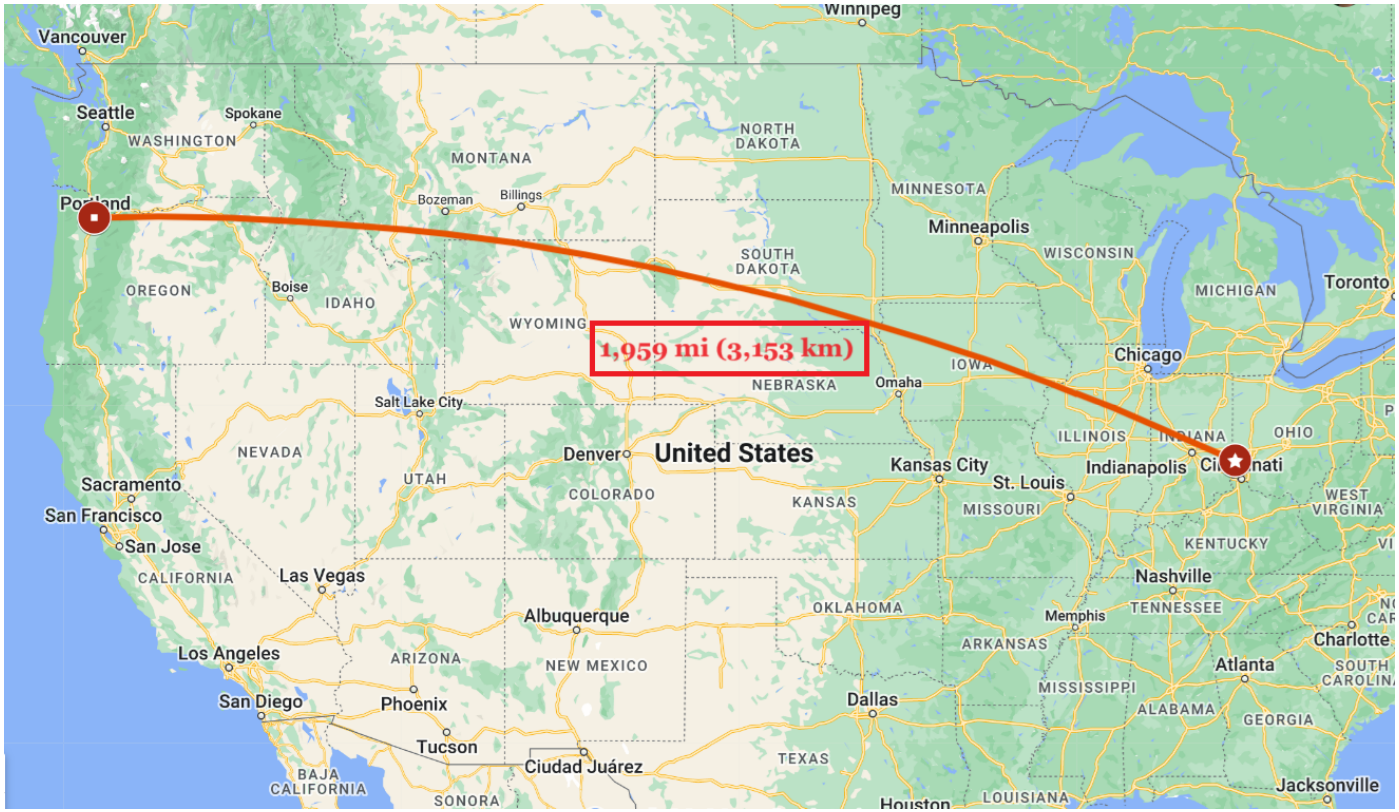
05 August 2024

MIAMI UNIVERSITY

OXFORD, OH • EST. 1809

Acton Lake

Where is this?



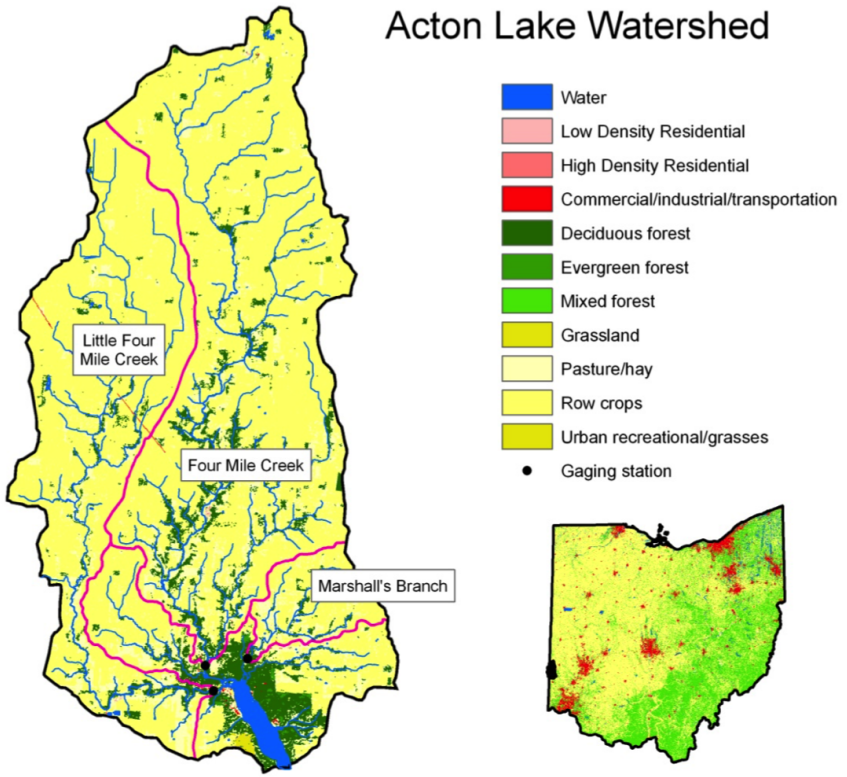
Acton Lake Watershed & Nutrients

Three different streams

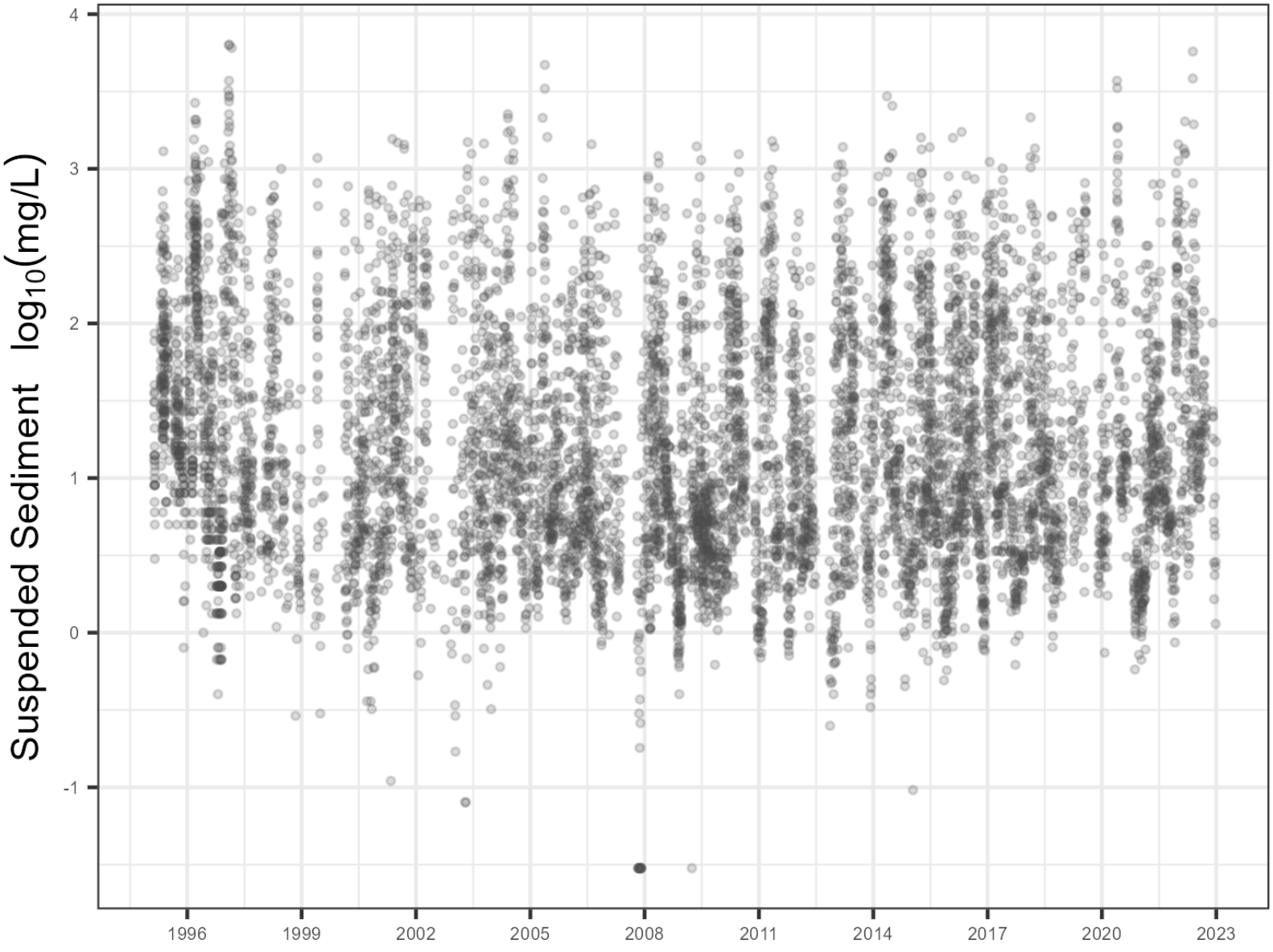
- Four Mile Creek
- Little Four Mile
- Marshall's Branch

Four different nutrients measured

- Ammonium (NH_4)
- Nitrate (NO_3)
- Soluble Reactive Phosphorus (SRP)
- Suspended Sediment (SS)



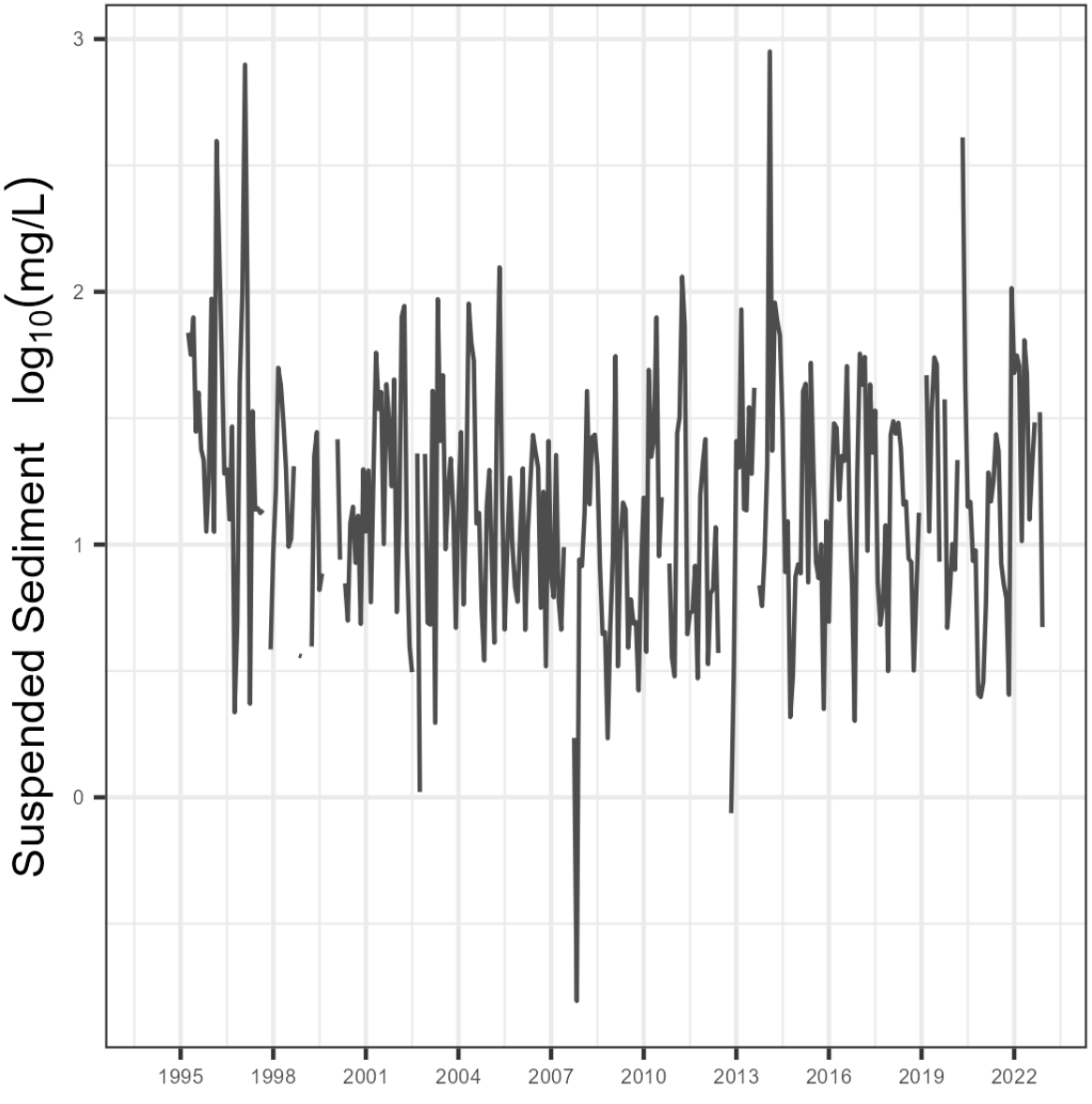
Example Data Set - Four Mile Creek Suspended Sediment



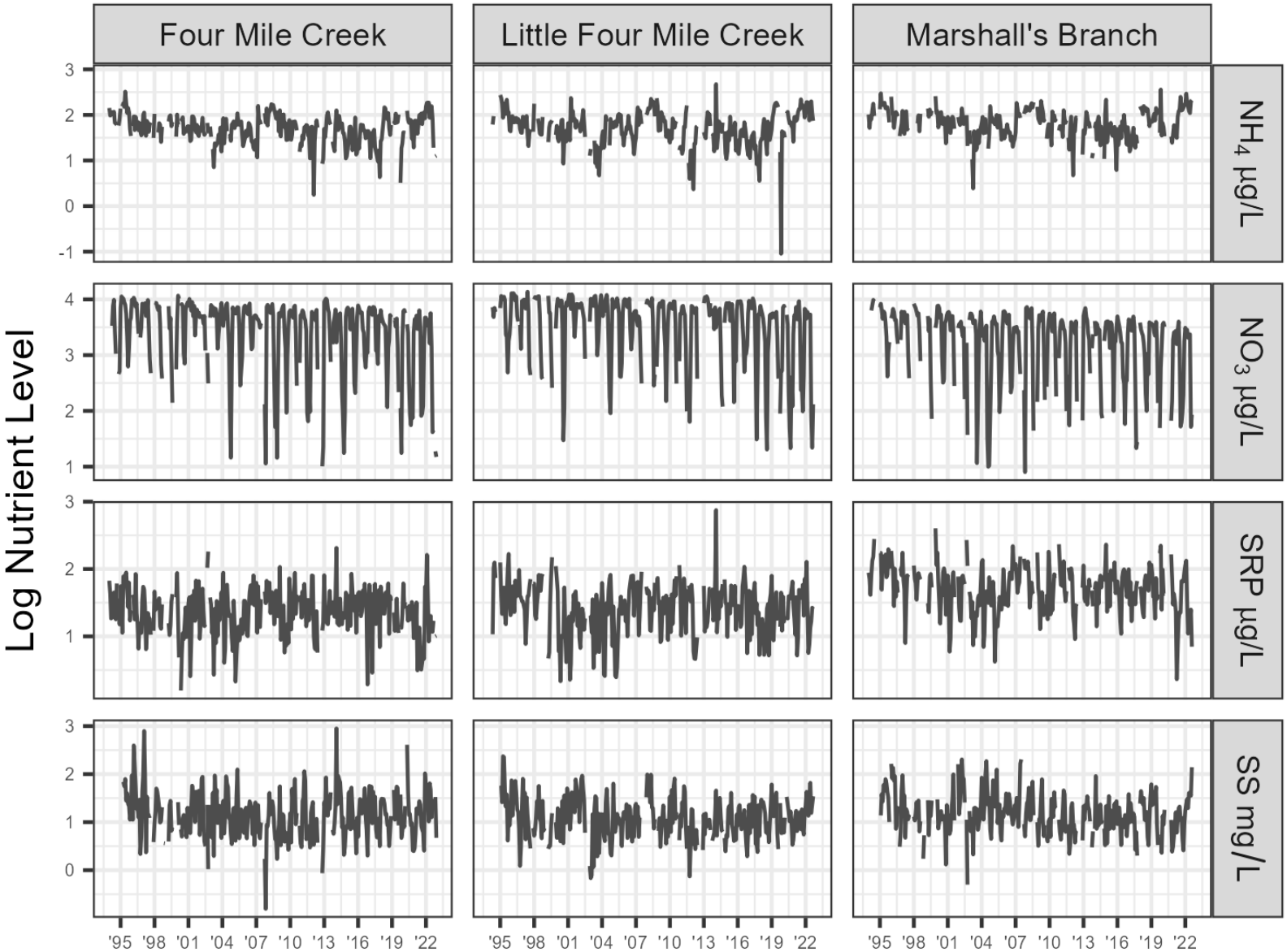
Example Aggregated Data Set

Monthly averages on a log scale

- $\log_{10}$ on each observation
- For each month-year, compute the mean of all observations



All Series


Four Mile Creek
Little Four Mile Creek
Marshall's Branch
NH4 µg/L
NO3 µg/L
SRP µg/L
SS mg/L
Log Nutrient Level
'95 '98 '01 '04 '07 '10 '13 '16 '19 '22

Features of the Data

Influence by stream flow (discharge)

- More water flow - more nutrients
- Known to be nonlinear relationship

Seasonal

- Weather patterns
- Seasonal Agricultural practices

Temporal correlation

- Underlying AutoRegressive type process

Possible long-term trends

Model Form

$$Y_t = \mu + \alpha t + \gamma_1 Q_t + \gamma_2 Q_t^2 + \sum_{j=1}^{5} \beta_j \sin\left(\frac{2\pi jt}{12}\right) + \sum_{j=1}^{6} \beta_j^* \cos\left(\frac{2\pi jt}{12}\right) + \varepsilon_t$$

- The term ε_t follows an AR(p) process
- The term Q_t models the effects of discharge
- The sin and cos terms model seasonality

Model Form

$$Y_t = \mu + \alpha t + \gamma_1 Q_t + \gamma_2 Q_t^2 + \sum_{j=1}^{5} \beta_j \sin\left(\frac{2\pi j t}{12}\right) + \sum_{j=1}^{6} \beta_j^* \cos\left(\frac{2\pi j t}{12}\right) + \varepsilon_t$$

- The term ε_t follows an AR(p) process
- The term Q_t models the effects of discharge
- The $\sin$ and $\cos$ terms model seasonality

Previous work looked at long term trends (importance of α); see Renwick et al. [2008] or Renwick et al. [2018].

Model Form

$$Y_t = \mu + \alpha t + \gamma_1 Q_t + \gamma_2 Q_t^2 + \sum_{j=1}^{5} \beta_j \sin\left(\frac{2\pi jt}{12}\right) + \sum_{j=1}^{6} \beta_j^* \cos\left(\frac{2\pi jt}{12}\right) + \varepsilon_t$$

- The term ε_t follows an AR(p) process
- The term Q_t models the effects of discharge
- The sin and cos terms model seasonality

Previous work looked at long term trends (importance of α); see Renwick et al. [2008] or Renwick et al. [2018].

We are looking for potential changes in μ and α

Model form (with change points)

$$Y_t = \mu + \delta_t + (\alpha + \Delta_t)t \; + \; \gamma_1 Q_t + \gamma_2 Q_t^2 + \sum_{j=1}^{5} \beta_j \sin\left(\frac{2\pi jt}{12}\right) + \sum_{j=1}^{6} \beta_j^* \cos\left(\frac{2\pi jt}{12}\right) + \varepsilon_t$$

where

$$\delta_t = \begin{cases} 0 & t < \tau_1 \\ \mu_1 & \tau_1 \le t < \tau_2 \\ \mu_2 & \tau_2 \le t < \tau_3 \\ \vdots & \vdots \\ \mu_m & \tau_m \le t \le N \end{cases}, \quad \Delta_t = \begin{cases} 0 & t < \tau_1 \\ \alpha_1 & \tau_1 \le t < \tau_2 \\ \alpha_2 & \tau_2 \le t < \tau_3 \\ \vdots & \vdots \\ \alpha_m & \tau_m \le t \le N \end{cases}$$

for change points

$$1 < \tau_1 < \tau_2 < \ldots < \tau_m \le N$$

Model form (with change points)

$$Y_t = \boxed{\mu + \delta_t + (\alpha + \Delta_t)t} + \gamma_1 Q_t + \gamma_2 Q_t^2 + \sum_{j=1}^{5} \beta_j \sin\left(\frac{2\pi jt}{12}\right) + \sum_{j=1}^{6} \beta_j^* \cos\left(\frac{2\pi jt}{12}\right) + \varepsilon_t$$

where

$$\delta_t = \begin{cases} 0 & t < \tau_1 \\ \mu_1 & \tau_1 \le t < \tau_2 \\ \mu_2 & \tau_2 \le t < \tau_3 \\ \vdots & \vdots \\ \mu_m & \tau_m \le t \le N \end{cases}, \quad \Delta_t = \begin{cases} 0 & t < \tau_1 \\ \alpha_1 & \tau_1 \le t < \tau_2 \\ \alpha_2 & \tau_2 \le t < \tau_3 \\ \vdots & \vdots \\ \alpha_m & \tau_m \le t \le N \end{cases}$$

for change points

$$1 < \tau_1 < \tau_2 < \ldots < \tau_m \le N$$

Model Selection

Minimum Description Length

- Based on coding and information theory
  - How much storage is required to describe a model
- Better models = models with minimal storage
- Each parameter can uniquely be penalized

Similar to the well-known AIC or BIC, but penalty structure is different.

The MDL approach has been shown to be effective in the change point problem; see Shi et al. [2022].

MDL Objective Function

$$
\begin{aligned}
-2\log L \;+\; 2\frac{\log(N)}{2} + k\frac{\log(N)}{2} \;+\; & \\
\sum_{j=1}^{m+1}\log(\tau_j - \tau_{j-1}) + \log(m) + \sum_{j=2}^{m}\log(\tau_j) + & \\
\frac{\log(N)}{2} + p\frac{\log(2N)}{2} + \log(p) &
\end{aligned}
$$

with $\tau_0 = 0$ and $\tau_{m+1} = N$.

MDL Objective Function

$$\boxed{-2\log L} \;+\; 2\frac{\log(N)}{2} + k\frac{\log(N)}{2} \;+\; \sum_{j=1}^{m+1}\log(\tau_j - \tau_{j-1}) + \log(m) + \sum_{j=2}^{m}\log(\tau_j) + \frac{\log(N)}{2} + p\frac{\log(2N)}{2} + \log(p)$$

with $\tau_0 = 0$ and $\tau_{m+1} = N$.

Log-Likelihood...

MDL Objective Function

$$-2\log L + \boxed{2\frac{\log(N)}{2} + k\frac{\log(N)}{2}} + \sum_{j=1}^{m+1}\log(\tau_j - \tau_{j-1}) + \log(m) + \sum_{j=2}^{m}\log(\tau_j) + \frac{\log(N)}{2} + p\frac{\log(2N)}{2} + \log(p)$$

with $\tau_0 = 0$ and $\tau_{m+1} = N$.

Cost for estimating two discharge coefficients and k sin and cos terms.

MDL Objective Function

$$-2\log L \;+\; 2\frac{\log(N)}{2} + k\frac{\log(N)}{2} \;+\; \sum_{j=1}^{m+1}\log(\tau_j - \tau_{j-1}) + \log(m) + \sum_{j=2}^{m}\log(\tau_j) + \boxed{\frac{\log(N)}{2} + p\frac{\log(2N)}{2} + \log(p)}$$

with $\tau_0 = 0$ and $\tau_{m+1} = N$.

Cost for determining autoregressive order and modeling AR(p).

MDL Objective Function

$$-2\log L + 2\frac{\log(N)}{2} + k\frac{\log(N)}{2} + \boxed{\sum_{j=1}^{m+1}\log(\tau_j - \tau_{j-1})} + \log(m) + \sum_{j=2}^{m}\log(\tau_j) + \frac{\log(N)}{2} + p\frac{\log(2N)}{2} + \log(p)$$

with $\tau_0 = 0$ and $\tau_{m+1} = N$.

Cost for fitting μ_j and α_j in $m+1$ regimes, each of length $\tau_j - \tau_{j-1}$.

MDL Objective Function

$$-2\log L + 2\frac{\log(N)}{2} + k\frac{\log(N)}{2} + \sum_{j=1}^{m+1}\log(\tau_j - \tau_{j-1}) + \boxed{\log(m) + \sum_{j=2}^{m}\log(\tau_j)} + \frac{\log(N)}{2} + p\frac{\log(2N)}{2} + \log(p)$$

with $\tau_0 = 0$ and $\tau_{m+1} = N$.

Cost for determining the number of change points and their locations.

MDL Objective Function

$$-2\log L + 2\frac{\log(N)}{2} + k\frac{\log(N)}{2} + \sum_{j=1}^{m+1}\log(\tau_j - \tau_{j-1}) + \log(m) + \sum_{j=2}^{m}\log(\tau_j) + \frac{\log(N)}{2} + p\frac{\log(2N)}{2} + \log(p)$$

with $\tau_0 = 0$ and $\tau_{m+1} = N$.

The details can be found in Lu et al. [2010].

Implementation

How to find the optimal MDL objective function?

Implementation

How to find the optimal MDL objective function?

Genetic Algorithm

- Binary search on $N - 1$ potential change point locations
  - Max *iterations* 5000, a *run* of 500 for convergence
  - Mutation probability: 0.1; Crossover probability: 0.8
- *Suggestions* for initial populations
  - No change point configuration
  - All single change point configurations
  - Random selection of 2 change point configurations
- Additional regime penalty (minimum regime length of 60 months or 5 years)

See Li and Lund [2012] for more on genetic algorithms and MDL.

Implementation

How to find the optimal MDL objective function?

Genetic Algorithm

- Binary search on $N - 1$ potential change point locations
  - Max *iterations* 5000, a *run* of 500 for convergence
  - Mutation probability: 0.1; Crossover probability: 0.8
- *Suggestions* for initial populations
  - No change point configuration
  - All single change point configurations
  - Random selection of 2 change point configurations
- Additional regime penalty (minimum regime length of 60 months or 5 years)

See Li and Lund [2012] for more on genetic algorithms and MDL.

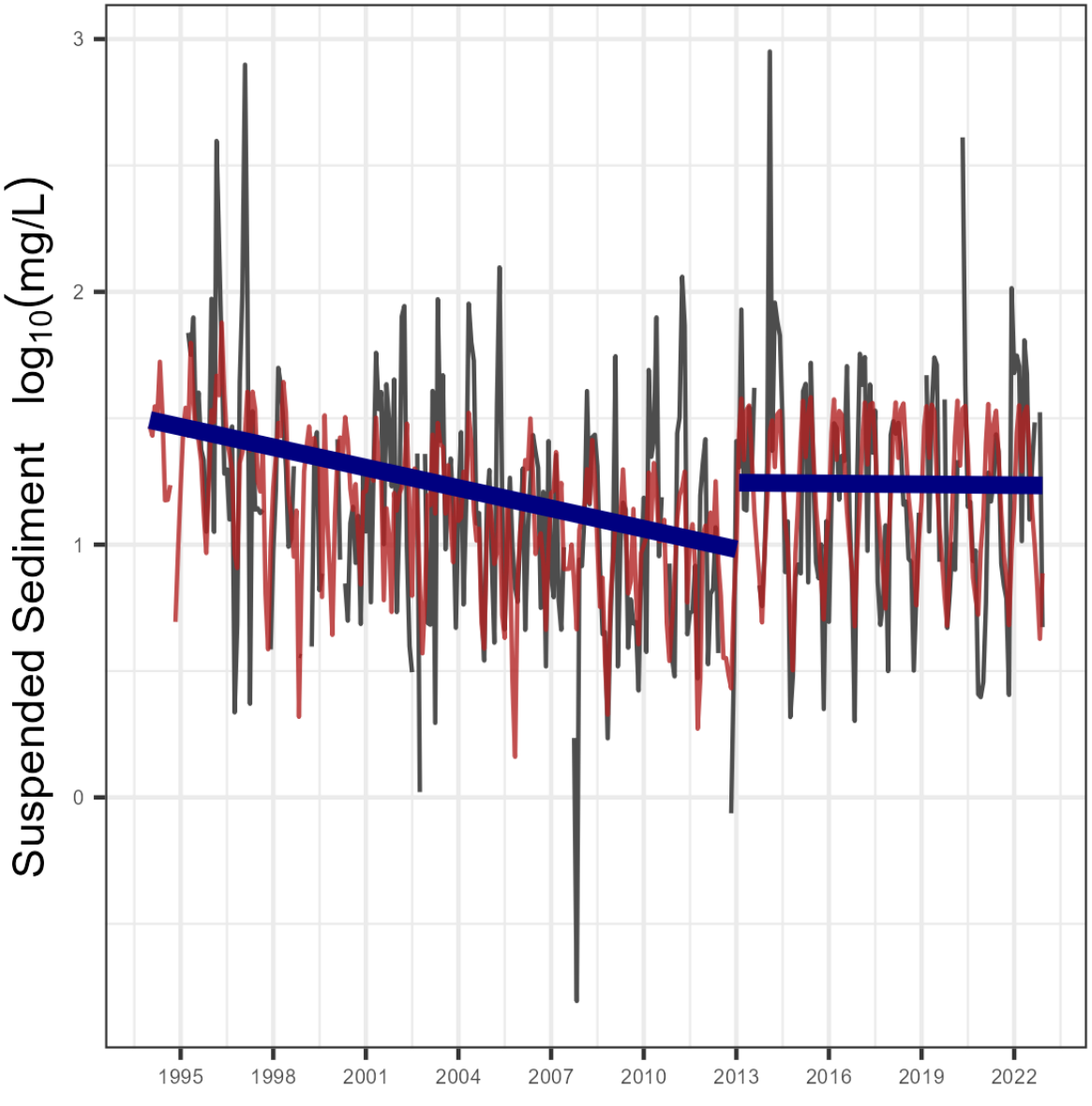
Repeated on all 4×3 combinations of streams & nutrients.

Results - Four Mile Suspended Sediment

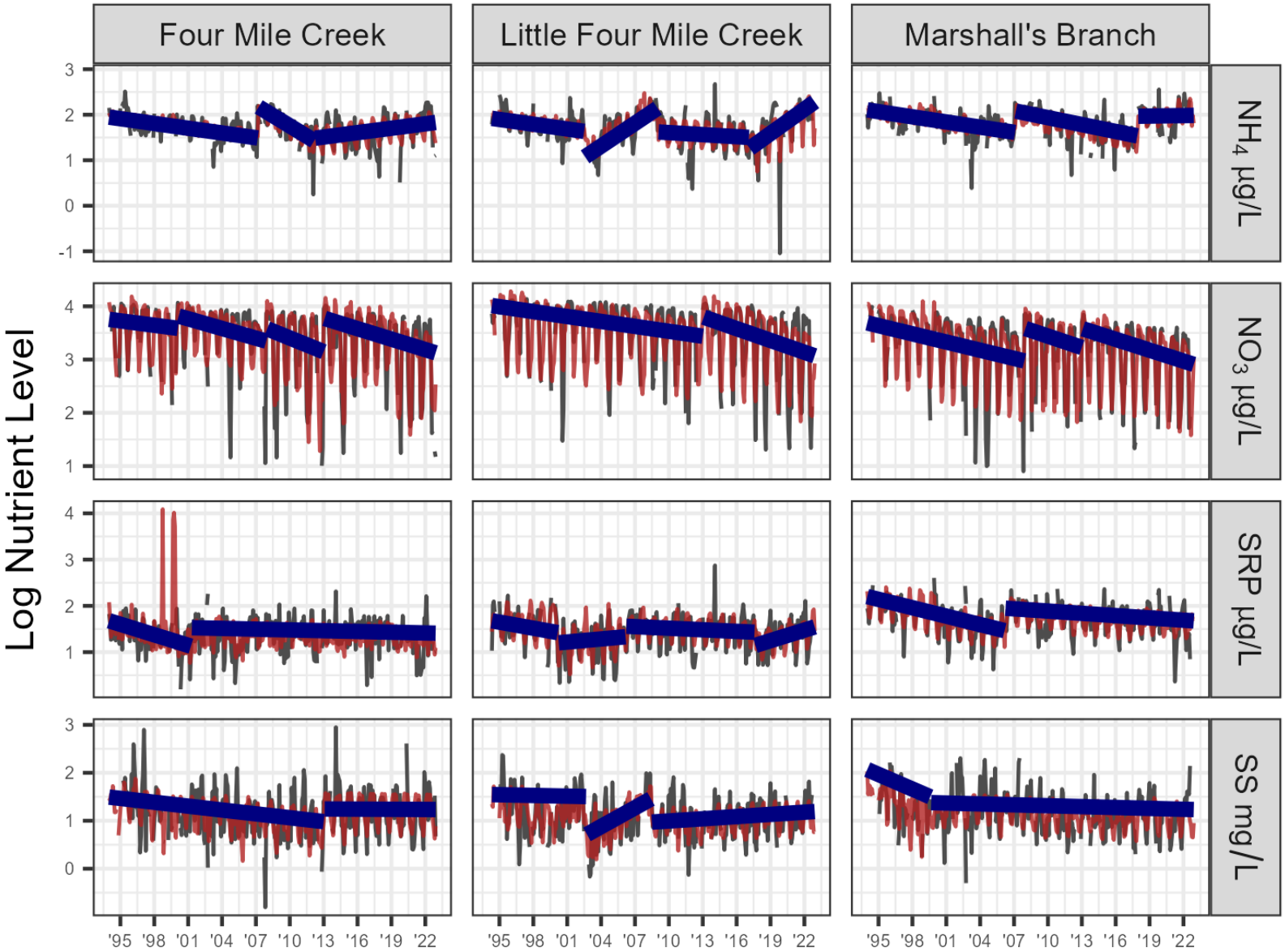
Original series

Fitted Model

Underlying Trend



Results - All stream-nutrient combinations



Conclusions

Some contextual conclusions

Ammonium (NH_4)

- Decreasing until ~2007, increasing since? (*squirrely*)

Nitrate (NO_3)

- Several decreasing regimes, with intermittent jumps

Soluble Reactive Phosphorus (SRP)

- Relatively stable after an initial decrease

Suspended Sediment (SS)

- Early decrease, now pretty stable

References

Shanghong Li and Robert Lund. Multiple changepoint detection via genetic algorithms. *Journal of Climate*, 25(2):674–686, 2012. ISSN 08948755, 15200442. URL https://doi.org/10.1175/2011JCLI4055.1.

Q. Q. Lu, R. Lund, and T. Lee. An MDL approach to the climate segmentation problem. *The Annals of Applied Statistics*, 4(1):299–319, 2010. URL http://dx.doi.org/10.1214/09-AOAS289.

William H. Renwick, Michael J. Vanni, Qianyi Zhang, and Jon Patton. Water quality trends and changing agricultural practices in a midwest u.s. watershed, 1994–2006. *Journal of Environmental Quality*, 37(5):1862–1874, 2008. URL https://doi.org/10.2134/jeq2007.0401.

William H. Renwick, Michael J. Vanni, Thomas J. Fisher, and Emily L. Morris. Stream nitrogen, phosphorus, and sediment concentrations show contrasting long-term trends associated with agricultural change. *Journal of Environmental Quality*, 47(6):1513–1521, 2018. URL https://doi.org/10.2134/jeq2018.04.0162.

Xuesheng Shi, Colin Gallagher, Robert Lund, and Rebecca Killick. A comparison of single and multiple changepoint techniques for time series data. *Computational Statistics & Data Analysis*, 170:107433, 2022. ISSN 0167-9473. URL https://doi.org/10.1016/j.csda.2022.107433.

Thanks!

Collaborator

- Hannah Waler - Former Masters Student
  Data Analyst at Progressive Insurance

Other contributors

- Dr. Mike Vanni - Ecologist
  Department of Biology - Miami University
- Dr. Bart Grudzinski - Hydrologist
  Department of Geography - Miami University
- Dr. Bill Renwick - Geographer
  Department of Geography - Miami University
- Dr. Emily Morris - Former Undergraduate Student
  Food & Drug Administration

Questions? Comments? Suggestions?

Shameless Self Promotion

Slides are available on my github site:
`https://tjfisher19.github.io/`

Github handle: **`tjfisher19`**

Code available: `https://github.com/tjfisher19`

Email: **`fishert4@miamioh.edu`**