Wayback Machine: Analyzing Specific Conductance

Manatee County Natural Resources Department: Environmental Protection Division

Standard Operating Procedures

7/7/2025

1. Purpose of project

The purpose of this project is to examine temporal trends of specific conductance levels in the Upper Manatee River. Similar studies have been conducted in surrounding watersheds, such as Horse Creek, a tributary of the Peace River in Hardee County, FL. Comparing trends between these watersheds may offer insight into broader regional patterns in water quality.

2. Quality Assurance

I. Data Integrity

All raw data were checked for completeness, consistency, and validity prior to analysis. Outlier values (e.g., pH > 9) were flagged and reviewed based on known environmental thresholds and patterns in surrounding years. Missing values were explicitly handled using LOESS interpolation only when surrounding data supported stable local trends.

II. Model Validation

The LOESS model was selected for its non-parametric flexibility and suitability for environmental time-series data. Model parameters (e.g., span) were selected using a combination of K-fold cross-validation and visual inspection to balance model fit and smoothness. Model fits were reviewed for smoothness and consistency with observed data on both sides of the interpolated gap (1986–1994).

III. Reproducibility

All analyses were conducted using R and fully documented in Quarto notebooks. All code was written using the Tidyverse framework, following consistent data manipulation and visualization practices for readability and reproducibility. The scripts are version-controlled and can be re-run using the same data and settings to reproduce all tables, figures, and interpolations.

IV. Cross-Validation with External Studies

Results were compared against similar studies from the Peace River Cumulative Impact Study (2007), particularly those from Horse Creek and Joshua Creek, to verify consistency of long-term trends across watersheds.

3. Procedure

I. Data Mining

i. Sift through documents and spreadsheets found in Wayback Machine, looking for suitable data with conductance values. Utilize the EPD “Correspondence Historical – ESTECH” reports. The data is organized by date and sampling site. Obtain data from “site 5,” which is the most comparable to the UM4 site. A map of the sampling sites is in the documents.

ii. Append conductance and pH measurements to an Excel spreadsheet. The spreadsheet should have three (3) columns: Year (or data), Conductance, and pH.
iii. Download contemporary dataset that contains data post-1995 from digital records. Ensure conductance levels at the UM4 sampling site are retrieved
iv. Combine old data with current data into one large Excel spreadsheet.

II. Data Cleaning

i. Export Excel spreadsheet as a .csv file.
ii. Import the .csv file to your IDE of choice, mine being RStudio.
iii. Read the file into your working directory.
iv. Reformat the "Year" column if the original data was in M/D/Y.
v. Filter out any years before 1980, as we are just looking for 1980 to now.
vi. Filter out any observations where the pH is over 9. These are outliers.
vii. Create a "Source" factor column that labels the data as either "archived" for the Wayback data or "current" for the digitized, downloaded data.
viii. Omit any rows with missing values

III. Data Modeling

There are missing values from 1986-1994, since the Wayback Machine does not have reports for this window. The downloaded data only includes data post-1995, thus leaving a gap. I needed to employ a model to interpolate these missing values. Given the time-series nature of the data and the presence of noisy observations, I chose a LOESS (Locally Estimated Scatterplot Smoothing) model to interpolate a smooth trend across the entire time range.

i. Decide values for the model's parameters. In my case, I really only needed to fine-tune the span parameter. I used K-fold cross-validation to identify the optimal span value, which controls the number, k, of neighboring data points used in each local fit. However, the cross-validation results varied across runs, yielding different span values each time. To address this, I repeated the tuning process multiple times to observe convergence patterns, but the selected span remained inconsistent. Lower spans allowed the earlier years (before 1986) to overly influence the interpolated region, since their weights are so large. Higher spans just captured the overall trend, which I did not want. Ultimately, I chose a span of 0.5, as it provided a good balance between the local trend of very small specific conductance levels and the broader trend of large increases in specific conductance levels.
ii. Fit the model using the "Year" column as the predictor variable and "Conductance" as the response. Set the data to be the imported dataframe, and set span to whatever you decide it to be, in my case 0.5. I set the family parameter to "symmetric" to reduce effect by extreme values.

iii. Use the fitted model to interpolate values for years where no data exists, specifically 1986–1994. Use predict(model, data= years 1986:1994), and set SE to True for standard error for visualizations. This will output a vector of predicted values for each desired year.

IV. Statistical Analysis

i. Add interpolated mean values to a tibble, along with mean values for each observed year. The result is a tibble with the mean conductance values for every year 1980-2024, with interpolated values filling the missing gap.

ii. Group values into 5-year bins, and find the 5 year averages for each bin. Creates a bar graph of the bins, color-coding if the bin is made up of interpolated values or not.

iii. Add standard error bars to the plot.

iv. Find percent increase and raw total increase between the first 5 years and the last 5 years.