Developing Data Cleaning and Familiarization Tools for StreamPULSE Users

Jin Cho, Yuval Medina, Vivek Sahukar
Project Managers: Alice Carter, Mike Vlah
Faculty Leads: Dr. Emily Bernhardt, Dr. Jim Heffernan

INTRODUCTION

Metabolism is the measure of energy within a stream ecosystem and is measured as the balance between gross primary production (GPP) and ecosystem respiration (ER) using oxygen or carbon dioxide. At StreamPULSE, where metabolism is modeled using oxygen, using a clean and complete dissolved oxygen curve is essential for accurate models.

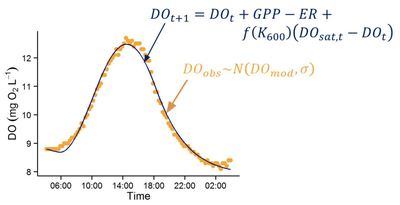
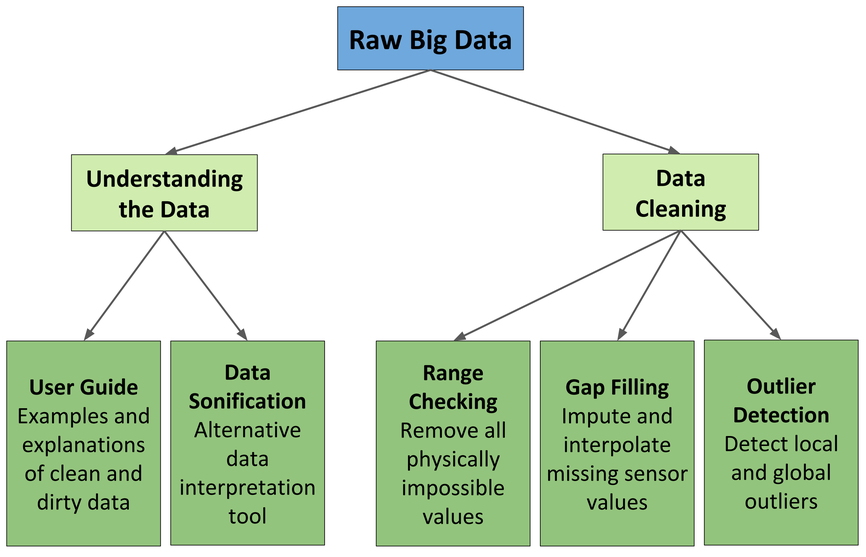


Figure 1. Raw and corrected dissolved oxygen curve from streamMetabolizer

However, raw sensor data is often incomplete or dirty and cannot be used to accurately model metabolism. Additionally, the interpretation and cleaning of large data sets from worldwide can be very difficult and creates the risk of misuse and misinterpretation. It is one of StreamPULSE's goal to create tools to familiarize users with their data and metabolism modeling, and to facilitate data munging.

OBJECTIVES



METABOLISM USER GUIDE

Not all users of StreamPULSE are familiar with the complexities of dealing with raw stream sensor data and modeling stream metabolism. Therefore, this guide was created to cover the significance and complexities of modeling metabolism and obtaining clean data to create those models. With examples and explanations, the guide shows users what dirty data they may encounter and the clean data they should look for.

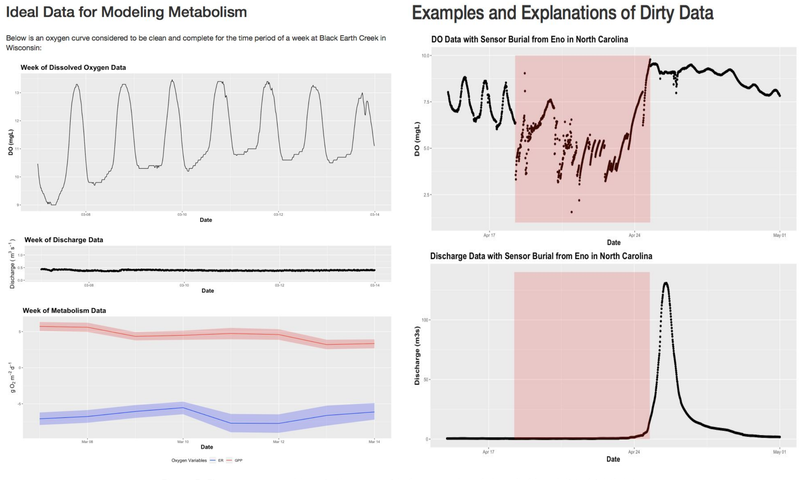


Figure 2. Figures users would see in the user guide for examples of clean (left) and dirty (right) data.

DATA CLEANING

Gap Filling by Using Imputation

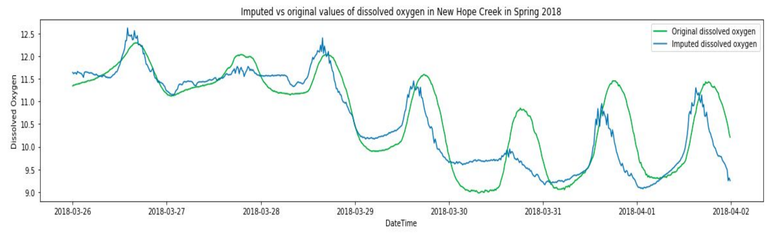


Figure 3. Comparison of original and imputed dissolved oxygen values by using IterativeImputer in scikit-learn

Simple imputation methods are better than recurrent neural networks for gap filling, even when there are periodic trends in the time series. Variables highly correlation with dissolved oxygen are used to fill missing values in the raw dissolved oxygen data.

Outlier Detection - Robust Random Cut Forest (RRCF)

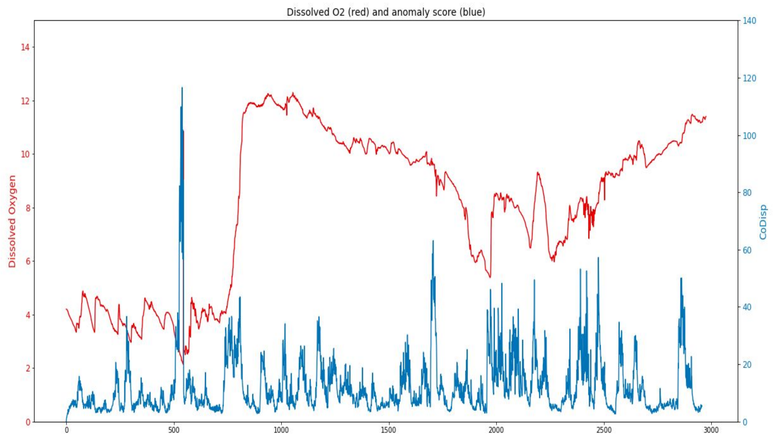


Figure 4. Dissolved oxygen values and their respective anomaly scores reported by RRCF algorithm

RRCF detects both local and global outliers in streaming data. Each data point is assigned an anomaly score with significantly higher scores indicative of outliers. A threshold determining which scores are flagged as outliers can then be manually set.

Outlier Detection Overview

The current method, although very fast, detects several false positives and negatives. In order to reinforce the current system, multiple new approaches and algorithms were explored, each with varying speeds and accuracy. With the synthesis of these methods, more outliers will be detected accurately with speed.

Outlier Detection - Moving Standard Deviation Window

Outliers were defined as points lying outside a standard deviation range of 3.1 within a sliding window of 900 points.

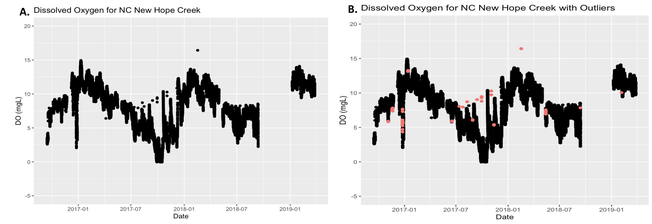
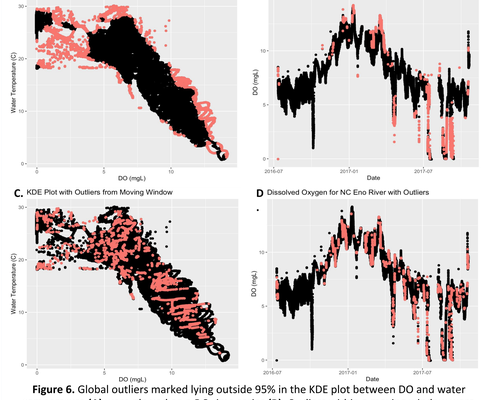


Figure 5. Method was applied to raw dissolved oxygen data from New Hope Creek in North Carolina **(A)**, resulting in flagged outliers in red **(B)**

Outlier Detection - Kernel Density Estimation (KDE)



Variables in the stream ecosystem are expected to be correlated, so KDE plots were developed between DO and water temperature to flag outliers identified as points lying outside the 95% contour line, which indicated their relationship at that time stamp was abnormal. This was applied to both the whole data set and a moving window of 2,300 points.

Figure 6. Global outliers marked lying outside 95% in the KDE plot between DO and water temperature **(A)** were plotted on a DO time series **(B)**. Outliers within a moving window were plotted on a correlation plot of all data **(C)** and on a DO time series **(D)**.

DATA SONIFICATION

This tool was developed using a Python module communicating with a SuperCollider IDE, an extremely powerful sound-synthesis programming language.

Being able to **hear** what various river systems sound like has great potential in **expanding interactions** with the StreamPulse community and beyond.

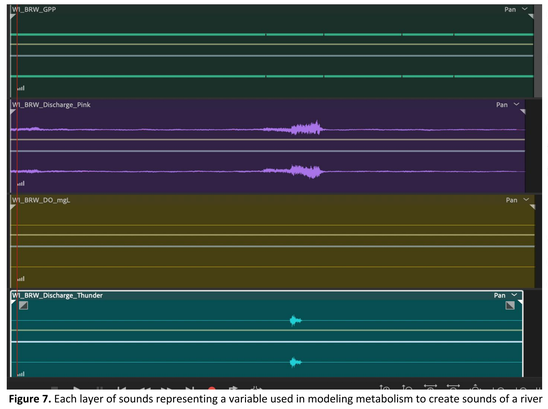


Figure 7. Each layer of sounds representing a variable used in modeling metabolism to create sounds of a river

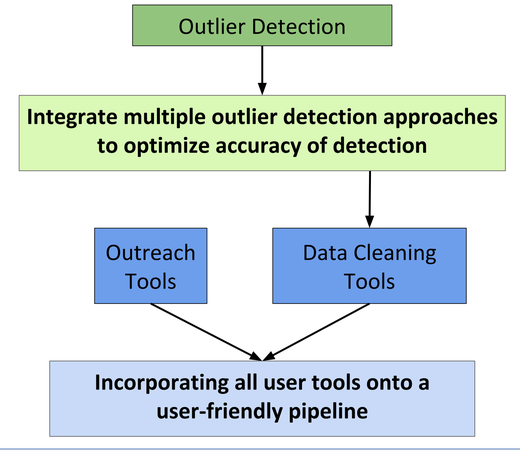
GPP – ascending and descending tones with respect to daily metabolism values calculated from DO and other data.

Discharge – daily standard deviations of the rate of water flowing through stream. Controls two tracks representing storms and other patterns that impact metabolism.

Dissolved Oxygen – standard deviation of daily windows of DO values. Being the variable most directly tied to metabolism (GPP), the two are in different sound registers to highlight this relationship.

Based on continuous feedback from river biologists, the current pipeline can be modified to create a more intuitive way of perceptualizing rivers through sound.

FUTURE STEPS



ACKNOWLEDGEMENTS

We would like to thank our project managers Alice Carter and Mike Vlah for their guidance throughout the summer project. We greatly appreciate the opportunities faculty leads Dr. Emily Bernhardt and Dr. Jim Heffernan, the members of the Duke River Center, and the Data+ Program at Duke University have provided for us.