Developing Benchmark Datasets for Testing Automated Sensor Data Quality Control Algorithm Performance

Ehsan Kahrizi, Jeffery S. Horsburgh

Utah Water Research Laboratory
UtahStateUniversity

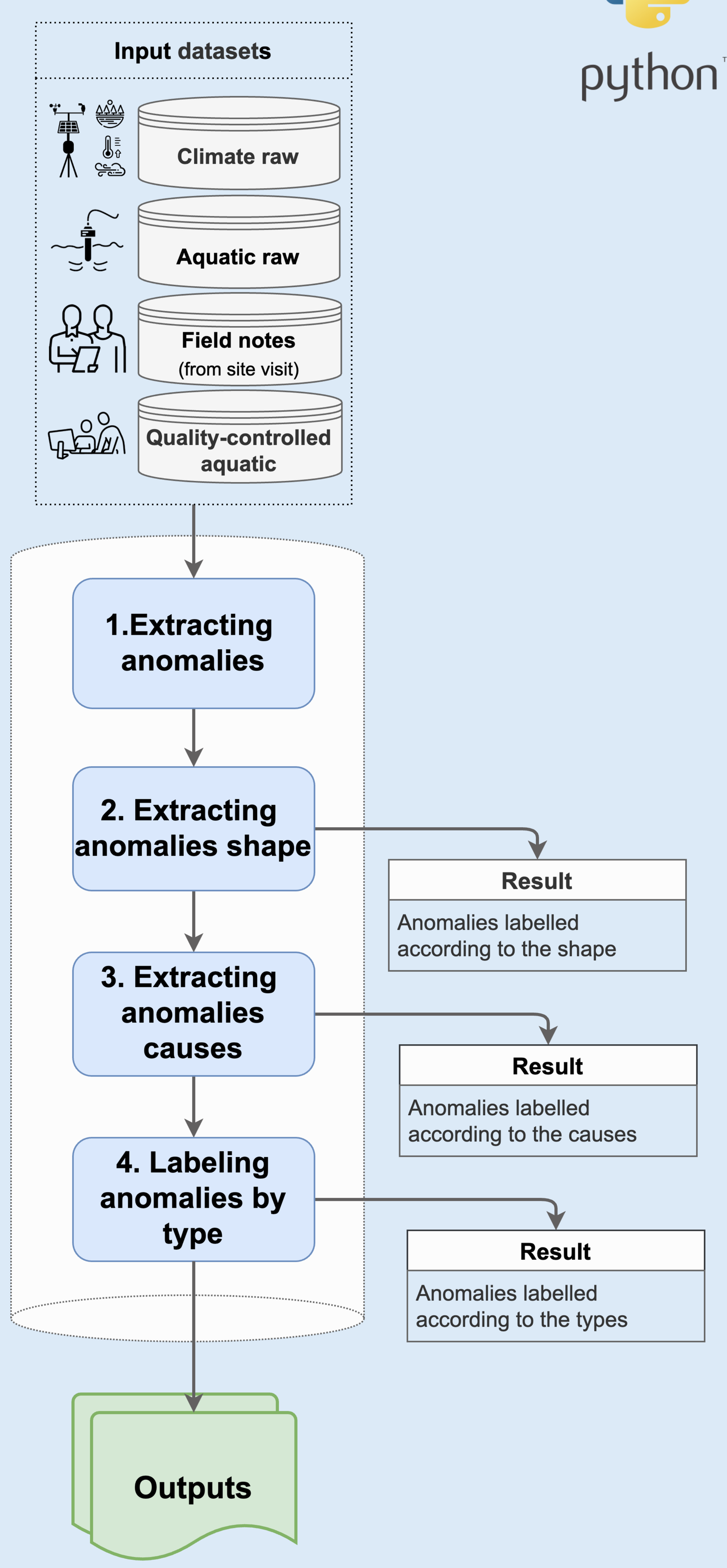
Support: Funding for this project was provided by the Utah Water Research Laboratory at Utah State University.

Introduction and Motivation

- Increasing sensor deployments have led to vast amounts of raw environmental data requiring quality control (QC) before use.
- Sensor anomalies arise from various sources, challenging anomaly detection, including environmental conditions, sensor issues, or transmission issues.
- Manual QC is costly and inconsistent, requiring expert intervention, which introduces subjectivity and delays in data availability.
- Automating QC can improve efficiency, reduce subjectivity, and enhance access to reliable, corrected data.
- Standardized benchmark datasets are needed to evaluate existing automated anomaly detection algorithms consistently.
- Very few labeled datasets exist for testing automated algorithms, which limits their testing and application.

Pipeline Architecture

- A reproducible pipeline designed to create benchmark datasets from raw and manually corrected data using Logan River Observatory data.
- It could be reused for other locations' data, creating a larger set of benchmark testing datasets.
- Automates input data retrieval and archives outputs to HydroShare or other repositories.
- Categorizes anomalies by shape, source, and type for improved anomaly detection.



Contact:

Jeff Horsburgh (jeff.horsburgh@usu.edu)
Ehsan Kahrizi (Ehsan.kahrizi@usu.edu)

Benchmark datasets created by this work will be shared in the HydroShare repository.

Source code for the data extraction pipeline is shared in GitHub.

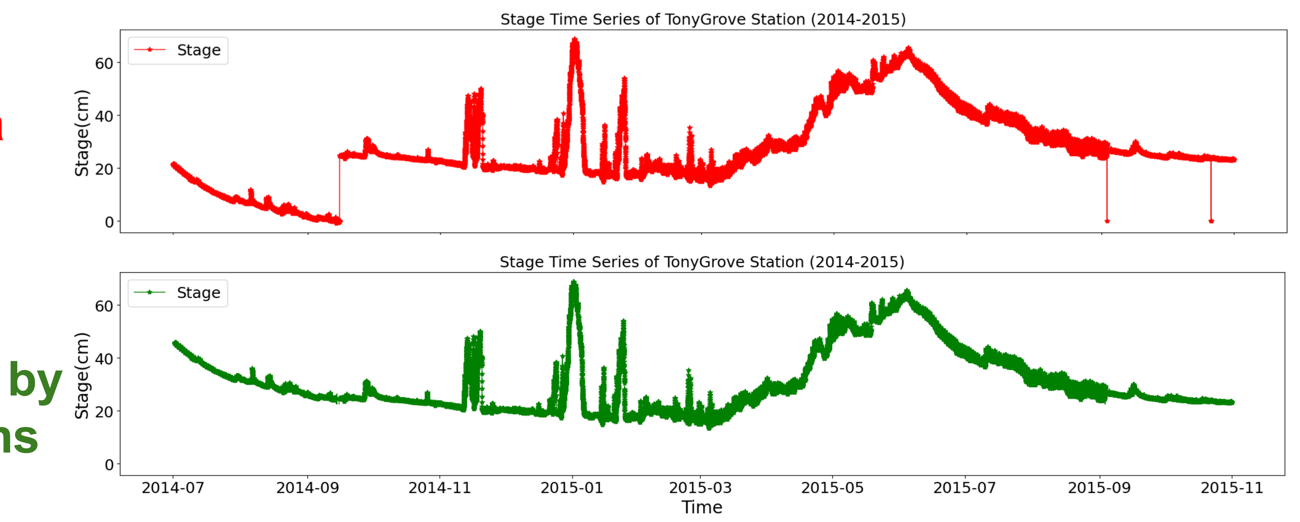
https://github.com/Ehsankahrizi/AutomatedAnomalyDetectionLabeling
https://www.hydroshare.org/resource/61a71043bc5240bea4baf3ec18872e9d/

Anomaly Detection, Categorization, and Benchmarking Datasets Approach

1. Anomaly extraction criteria

Raw data

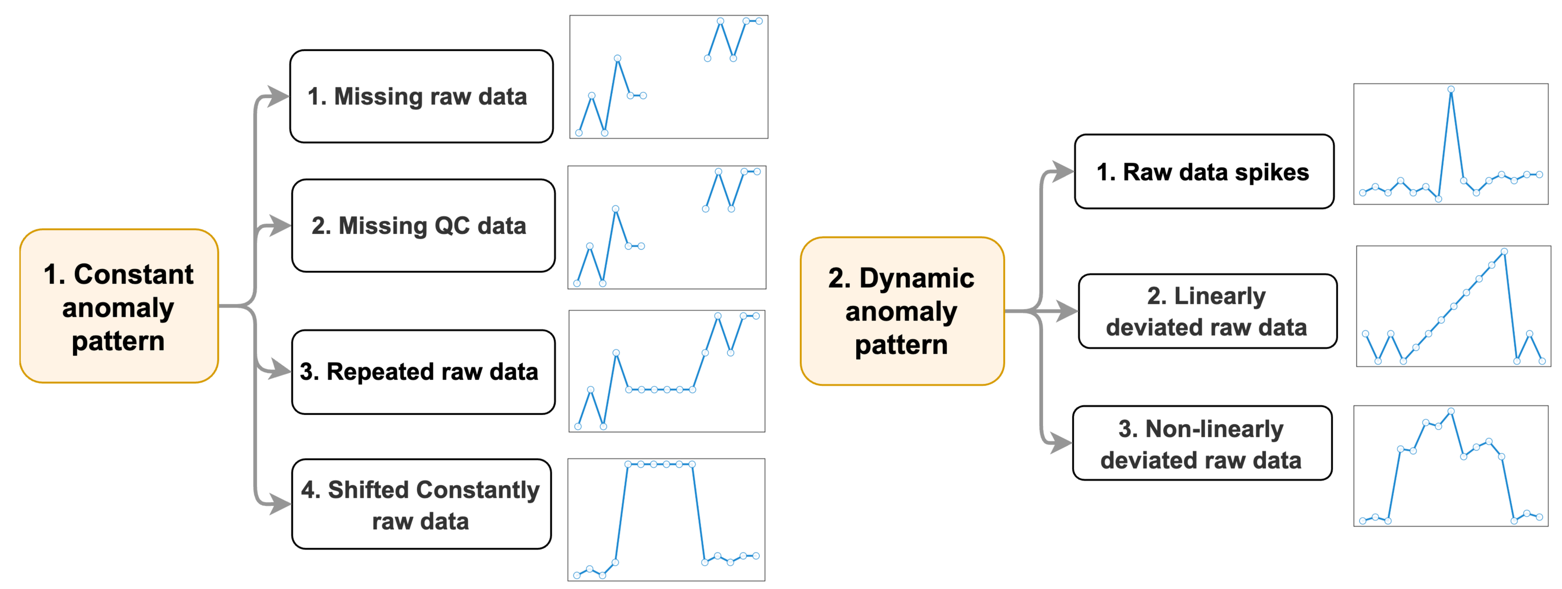
Manually corrected by technicians



- Anomalies consist of errors in the sensor data that need to be corrected.
- Differencing raw and manually corrected data identifies anomalies corrected by a technician.
- The anomaly domain using this logic was extracted, providing key features of each anomaly, such as duration, range, and additional features.

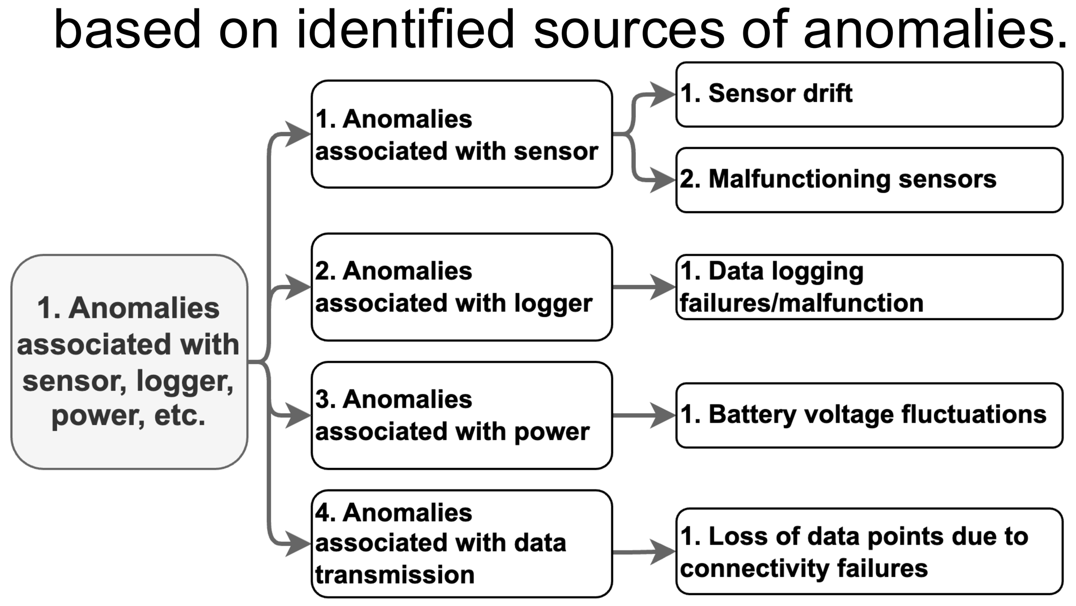
2. Extracting and labeling anomaly shapes

- Categorizing anomalies based on their shape and pattern.
- Automatically identifying visual features of anomalies.

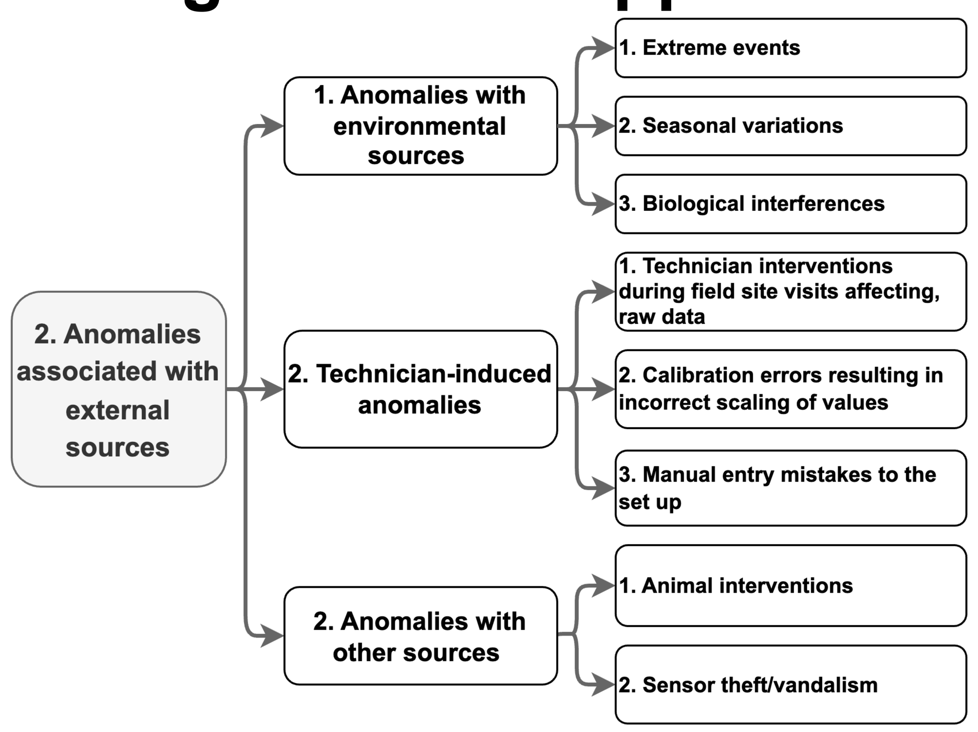


3. Labeling anomalies by cause

- Using multivariate analysis of aquatic and climate data, along with documented field notes from site visits, anomaly causes were identified.
- A generic categorization of anomaly causes was developed based on identified sources of anomalies.



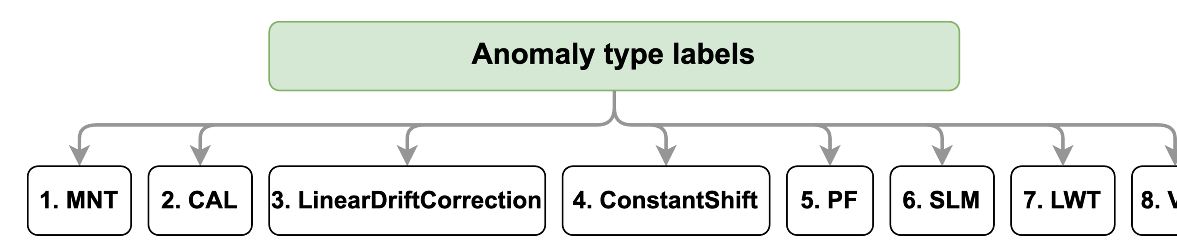
Identifying possible anomaly causes associated with infrastructure.



Recognizing possible anomalies caused by external sources, such as environmental conditions, technician-induced, or vandalism.

4. Labeling anomalies by type

- Considering field conditions, infrastructure, and site visit notes, a set of diagnostic questions was designed to link anomaly shape labels with cause labels.
- This process resulted in a generic categorization of anomaly types, summarizing patterns, features, and potential causes.



- **CAL:** Calibration performed
- **MNT:** Maintenance performed
- **PF:** Power failure issues
- **LinearDriftCorrection:** Corrected using linear regression
- **ConstantShift:** Corrected using a constant shift
- **SLM:** Sensor/logger malfunction
- **LWT:** Sensor was out of water
- **VIN:** Visual inspection needed

5. Output: Benchmark datasets shared in HydroShare

- Outputs include raw data, corrected data, and identified anomalies, along with calculated features and labels.
- Outputs are ready to serve as inputs to train new algorithms and methods for anomaly detection and correction.
- Example of outputs: