Quality over Quantity: A Data-Driven, Automated Quality Assurance and Control Process for Continuous, Hydrological Data.

M.W. McGauley[1*], V.B. Smith, Ph.D.[1], B.M. Wadzuk, Ph.D.[1]

[1]*Villanova Center for Resilient Water Systems, Villanova University, Villanova, Pennsylvania*

**Corresponding author email:* mmcgau01@villanova.edu

Highlights

- Sensor networks increase data resolution and quantity, but are prone to error, requiring automated methods to produce uniformly processed, timely data.
- This process harnesses the stochasticity of rainfall and flow data to make informed decisions that assure the quality of sensor-derived data in a repeatable manner.

Introduction

Hydrological data is inherently chaotic and modern methods of collection exacerbate this issue. Patterns in hydrologic systems are often predictable based on their characteristics, but are inherently random as a result of complex interactions between their climate and landscape (Sivakumar 2017). Sensors can seamlessly capture increasingly massive amounts of hydrological data in high-resolution, continuous time scales. However, they are prone to malfunction for a myriad of reasons that are just as stochastic as patterns in hydrologic system in terms of their cause and timing (Campbell et al. 2013). As a result, sensor-derived data typically requires additional steps to assure its quality.

Ensuring data quality is an integral aspect to making informed decisions in the field of water resources management. For example, determining the performance of stormwater control measures (SCM's) is assessed with water quality and quantity data (Liu et al. 2017). Timely and resource-effective decisions about the design and maintenance of SCM's therefore requires that performance-based conclusions are derived from high-quality data, defined as being correct, consistent, and complete (Chao et al. 2015). Large amounts of potentially poor hydrological data require quality assurance processes to transform them into high-quality datasets. Relying on manual methods for quality assurance introduces human bias, is not based on statistical measure, and is inefficient, and potentially infeasible, with large datasets.

An automated quality assurance process, presented here, solves these limitations by harnessing the stochastic nature of hydrological data to make efficient, data-driven, and statistically based inferences for handling poor quality data. The process is controlled, efficient, and repeatable with any continuous dataset containing observations for water quantity data with concurrent rainfall and flow data.

Methodology

This process involves the correction of poor-quality water quantity data stored in Excel workbooks (.xlsx) aided by data transformation with the Pandas python package. Columns specify the datetime (a combination of a date and timestamp in 24-hour format) with observations of rainfall (depth in inches since last timestep) and flow (instantaneous reading or average in cubic feet/second, cfs, over time since previous timestep) at each timestep are required. Concurrent data for any number and type of water quantity variables are then included with observations at each timestamp, leaving missing values blank. Columns are required to have a header in the first row of the Excel workbook detailing the variable each column represents. Datetime, Flow, and Rainfall columns should be named as stated here and appear in this order followed by the columns with water quantity data to be quality assured.

Excel workbooks containing properly formatted data are read into Pandas dataframes in the python script used for this analysis. Timestamps in the datetime column are formatted and set to serve as the index for the dataframe. Periods of stormflow and baseflow are then assigned using the following logic: A unique period of stormflow occurs when there is any amount of instantaneous rain that begins after at least six hours of no rain, followed by any amount of rain with dry periods in between them that last less than six hours, and is followed by at least a six-hour dry period after the last instance of rainfall or whenever flow returned to below 0.1 cfs after the last instance of rainfall. Otherwise, there is a unique period of baseflow.

Unique periods of storm and baseflow are then assigned a unique ID by appending an integer to the storm or baseflow designation assigned in the previous step. A z-score is then calculated for each value of each water quantity variables within each unique period of stormflow and baseflow. The mean of each period of baseflow and the median of each period of stormflow is then used to fill missing values and replace values that are beyond three standard deviations (left or right) from each period's mean. Plots of the change in value before and after the quality assurance process are then created. The quality assured dataset is exported to an excel workbook file. Summary statistics are computed to determine the effect size of the quality assurance process on key data quality metrics.

Results and discussion

The Villanova Center for Resilient Water Systems (VCRWS) monitors velocity of stormwater runoff entering a constructed stormwater wetland (CSW) SCM at its inlet. This sensor is known to produce unrealistically large values, especially during the peak flows of storm events. Figure 1 shows six months of CSW velocity data where 52,416 values were processed in 49.26 seconds to change extreme values.

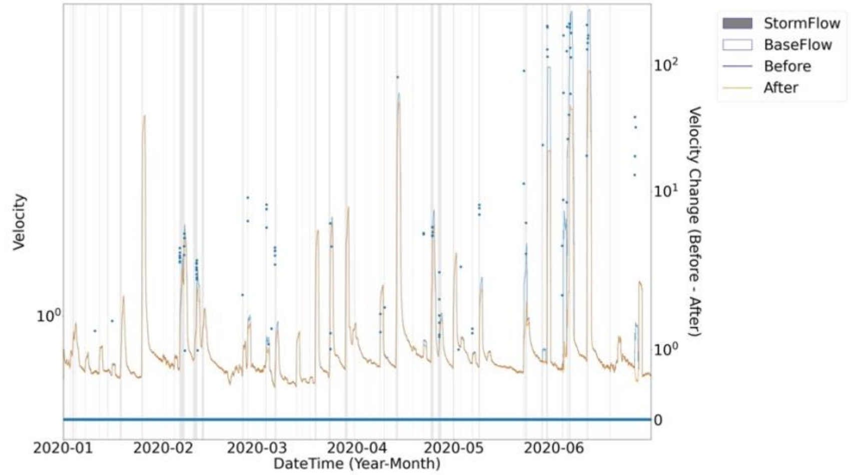
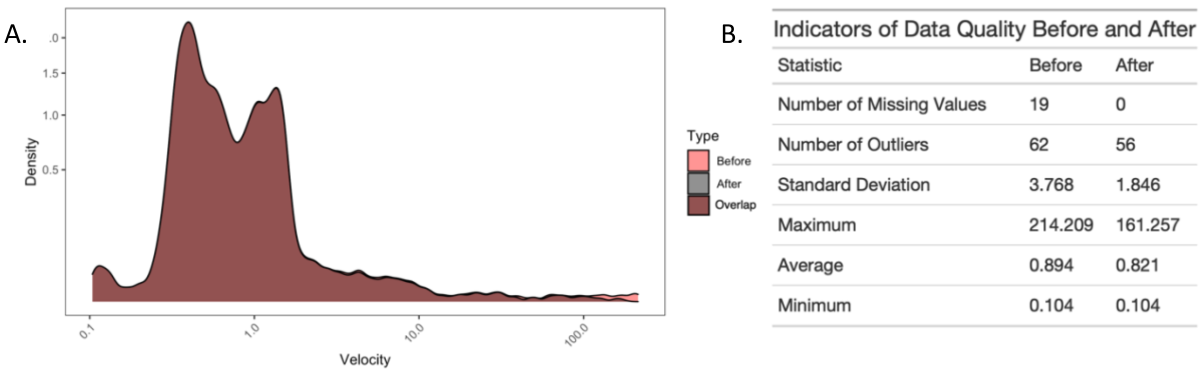


Figure 1 Change in moving one-day average velocity (feet/second) (orange and blue lines on left scale) and velocity change before and after (blue points on right scale) processing with labelled periods of storm (grey) and baseflow (white).

Most value changes occur during periods of stormflow, with some values changing during periods of baseflow. This indicates that faulty readings with this sensor typically occur during periods of high velocity because of stormwater runoff entering the CSW, but that faulty readings can still occur randomly when there is no rainfall occurring.

This process is effective at replacing extreme values with more-representative values of true conditions at the time of recording. The right-most portion of Figure 2.A. in solid red shows there is a high concentration of velocity readings of 150 feet/second or greater before processing. This process reduces the maximum reading to about 161 feet/second from over 200 feet/second, reducing the number of outliers that are more than six standard deviations from the mean (Figure 2.B., Maximum and Number of Outliers). Data completeness improves by filling missing values and data correctness improves from a decrease in the standard deviation while keeping the average relatively unchanged (Figure 2.A., dark red “Overlap” below 100 ft/s), largely due to altering the highest values (Figure 2.B., Number of Missing Values, Standard Deviation, Average).



Indicators of Data Quality Before and After

Statistic	Before	After
Number of Missing Values	19	0
Number of Outliers	62	56
Standard Deviation	3.768	1.846
Maximum	214.209	161.257
Average	0.894	0.821
Minimum	0.104	0.104

Figure 2 A. Density of Velocity Values (feet/second) and **B.** Indicators of Data Quality Before and After Processing.

Conclusions and future work

The process described here is an effective and efficient means of harnessing the power of the stochasticity of hydrological data to improve data quality for any continuous water quantity data. It makes informed poor-quality data detection and replacement decisions that do not cause departure greatly from the overall characteristics of the original dataset. This process can be expanded by improving the poor-data replacement aspect with the use of value imputation and machine learning techniques, which can make more intelligent decisions beyond using a mean or median value.

References

Campbell, J. L., Rustad, L. E., Porter, J. H., Taylor, J. R., Dereszynski, E. W., Shanley, J. B., Gries, C., Henshaw, D. L., Martin, M. E., Sheldon, W. M., and Boose, E. R. (2013). “Quantity is Nothing without Quality: Automated QA/QC for Streaming Environmental Sensor Data.” *BioScience*, 63(7), 574–585.

Chao, L., Hui, Z., and Xiaofeng, Z. (2015). “Data quality assessment in hydrological information systems.” *Journal of Hydroinformatics*, 17(4), 640–661.

Liu, Y., Engel, B. A., Flanagan, D. C., Gitau, M. W., McMillan, S. K., and Chaubey, I. (2017). “A review on effectiveness of best management practices in improving hydrology and water quality: Needs and opportunities.” *Science of The Total Environment*, 601–602, 580–593.

Sivakumar, B. (2017). *Chaos in Hydrology*. Springer Netherlands, Dordrecht.

Stewart, B. (2015). “Measuring what we manage – the importance of hydrological data to water resources management.” *Proceedings of the International Association of Hydrological Sciences*, 366, 80–85.