Controlling the Chaos: An Environmentally-informed, automated Quality-Assurance and Quality-Control Model for Continuous, Hydrological Data

McGauley Matthew[1], Jacko Brian[1], Estes Sarah[1], Smith Virginia[1], and Wadzuk Bridget[1]

[1]Villanova Center for Resilient Water Systems

November 16, 2022

Abstract

While more hydrological data is being generated than ever before, the power of modelling this collected information is not fully realized unless it is of high quality, especially considering hydrological data from sensor networks, which is often errant due to the possibility of malfunction or non-conducive environmental conditions. Fluctuations or errors are difficult to predict, identify, and interpret. Manual models of quality assurance are not designed for managing datasets with continuous timeseries or spatially extensive coverage, resulting in time- consuming models that rely on humanmade decision making and lack statistical inference. This research hypothesizes that the stochasticity of rainfall and deterministic properties of flow can be used in concert to create a more characteristic quality assurance model for high-resolution environmental data. An automated implementation of this model is presented herein with the application of two use-cases, which maintains statistical integrity and circumvents biases and potential for user error of manual frameworks.

Controlling the Chaos: An Environmentally-informed, automated Quality-Assurance and Quality-Control Model for Continuous, Hydrological Data.

M. W. McGauley[a], B. E. Jacko[a], S. Estes[a], V. B. Smith[a], B. M. Wadzuk[a]

[a]Villanova Center for Resilient Water Systems, Department of Civil and Environmental Engineering, Villanova University, 800 Lancaster Ave, Villanova, PA 19085.

Corresponding author: Matthew McGauley (Villanova Center for Resilient Water Systems, Department of Civil and Environmental Engineering, Villanova University, 800 Lancaster Ave, Villanova, PA 19085, +1-610-639-3396, mmcgau01@villanova.edu)

This manuscript is a pre-print of an article submitted to the journal Environmental Modelling and Software and has not been peer reviewed.

Abstract:

While more hydrological data is being generated than ever before, the power of modelling this collected information is not fully realized unless it is of high quality, especially considering hydrological data from sensor networks, which is often errant due to the possibility of malfunction or non-conducive environmental conditions. Fluctuations or errors are difficult to predict, identify, and interpret. Manual models of quality assurance are not designed for managing datasets with continuous timeseries or spatially extensive coverage, resulting in time-consuming models that rely on humanmade decision making and lack statistical inference. This research hypothesizes that the stochasticity of rainfall and deterministic properties of flow can be used in concert to create a more characteristic quality assurance model for high-resolution environmental data. An automated implementation of this model is presented herein with the application of two use-cases, which maintains statistical integrity and circumvents biases and potential for user error of manual frameworks.

Highlights:

- The stochasticity of rainfall and deterministic properties of flow are harnessed.
- Anomalies are corrected with high confidence and no data volume constraints.
- Risk of committing unintended questionable research practices and bias is reduced.

Keywords:

Quality assurance and control, sensor networks, hydrological data, data quality, anomalous data

Software and Data Availability:

Name of software: QAQC_nonRegression.py and QAQC_Regression.py

Developers: Matthew McGauley

Contact: mmcgau01@villanova.edu

Year first available: 2022

Hardware required: A personal computer

Software required: Microsoft Windows or MacOS

Software availability: https://github.com/mwm021/QAQC

Cost: Free. Code released under the MIT License.

1 Introduction

Recent innovations in technology and available data have created exciting opportunities for advancements across the spectrum of scientific disciplines (Philbeck and Davis 2018). The overwhelming benefits of large datasets, however, can be weighed down by the conjugate increase in error potential. This is true of all data, regardless of the source or method of derivation. The present work seeks to explore methodological means by which a model to resolve such issues and its potential was demonstrated by use cases on hydrological data. Data of this type is subject to a slew of external variables that cannot be controlled, allowing for instances of chaotic values to arise within the larger set. Despite the characteristic predictability of hydrologic systems (as they are often well modeled within the environmental context prior to institution), complex interactions with local climate and landscape introduce inherent complications (Sivakumar 2017). The insurance of quality data is an integral aspect of making informed decisions about modeled data in the field of stormwater resources management, as in other fields (Refsgaard et al. 2005; Wadzuk et al. 2021). For example, water quality and quantity data collected from stormwater control measures (SCMs) are utilized to determine system performance (Liu et al. 2017). Timely, resource-effective, and information-driven decisions about the design and maintenance of SCMs therefore require performance-based conclusions that are derived from high-quality data, i.e., that which is correct, consistent, and complete (Chao et al. 2015). As these systems become increasing pervasive collecting and managing high quality data for aiding in decision making and design becomes increasingly difficult. Further, retrospective identification of error in sensor function or quality assurance practices become increasingly onerous with larger volumes. Ultimately, these constraints limit the sustainable function of these systems.

In-situ sensors can seamlessly capture increasingly large amounts of hydrological data in high-resolution at continuous time scales. However, they are prone to malfunction for a myriad of environmental, internal, and physical reasons (e.g., the loss of site power, reading drift, lack or loss of adequate calibration, or sensor fouling) (Figure 1) that are difficult to detect due to conjugate system stochasticity in terms of cause and timing (Campbell et al. 2013). As a result, sensor-derived data typically requires additional steps to assure its quality. When sensor-derived data is uncharacteristic of the conditions of its collection, there is inherent bias and incorrect inference (McCausland 2021) that may become difficult to detect at scale.

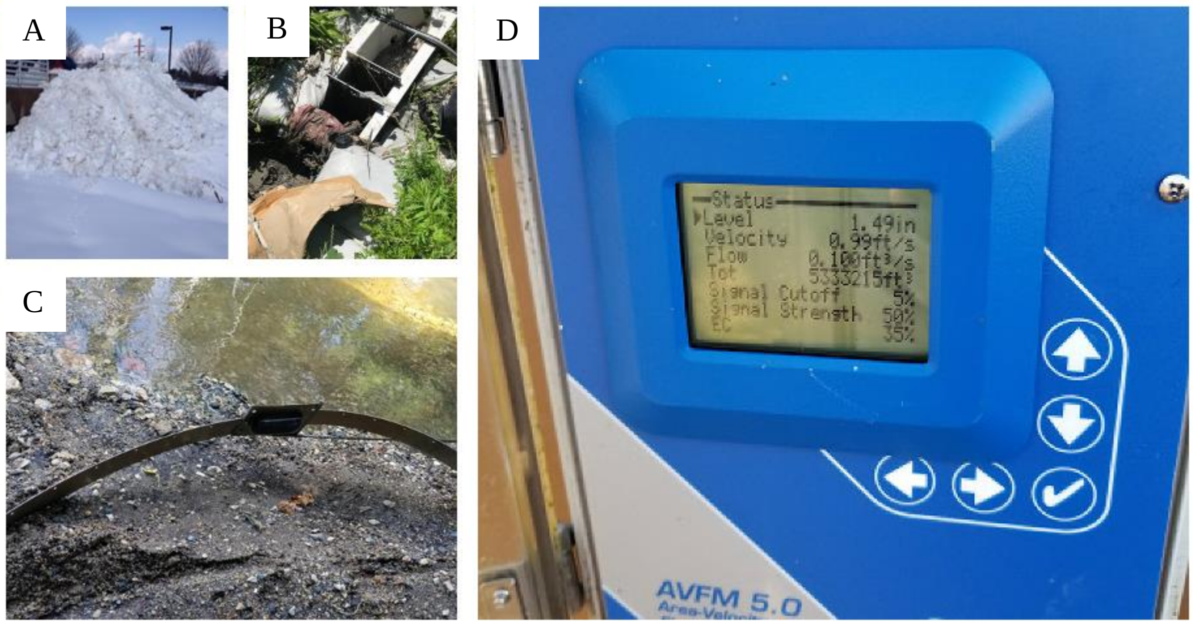


Figure 1 Sensor issues identified and documented at Stormwater Control Measures (SCMs) managed by the Villanova Center for Resilient Water Systems (VCRWS). **A.** an entire SCM (outfitted with sensors) that was buried in snow after a storm, **B.** a sensor that was clogged with debris, **C.** a sensor and conjugate pipe band torn from sample location due to extreme weather, **D.** a datalogger which reported issues with poor signal strength and electrical conductivity. Photos were provided courtesy of Dr. Gerald Zaremba.

Relying on manual models for quality assurance introduces a window for human bias and can therefore raise questions regarding research practices unless the researcher can report and justify standard, repeatable reasoning for the exclusion or augmentation of their outliers (Banks et al. 2016). Particularly with manual models on large, long-term data sets, many different people assess the data over time, employing inconsistent internal decision-making behavior that remains procedurally consistent with internally and externally accepted practices (Hertwig and Gigerenzer 2011; Hsee et al. 2004). Even with Quality Assurance and Quality Control (QAQC) protocols in place, there is always an opportunity for human-imposed bias. Manual models are often implemented without statistical foundation, are neither efficient nor feasible with large datasets, and exacerbate the issue of bias that uncharacteristic data inherently possesses (Kumari and Kennedy 2017).

Quality assurance models are well-documented in the field of hydrology. The United States Environmental Protection Agency (EPA) outlines standards for retrieving quality data from environmental systems in a Quality Assurance Project Plan ("Guidance for Developing Quality Systems for Environmental Programs" 2002), but leaves the implementation of quality assurance models to management bodies themselves. Other institutions and research groups have developed computer software to detect and correct anomalies in sensor-derived environmental data (Asquith et al. 2020; Díaz et al. 2021; Faybishenko et al. 2022; Horsburgh et al. 2015), but none have done so on an event-by-event basis, in which environmental observations data is contextualized by the unique climatic conditions of its collection. Considering observations on an event-by-event basis provides insight into the intricacies of the time, space, and climatic conditions of captured data (Blaen et al. 2017). The example of event-specific flow velocity in a constructed wetland, provided in Figure 2, demonstrates the need for this approach. Flow velocities are flashy as rainfall intensity fluctuates throughout a storm (a period of stormflow, Figure 2A) and remain more constant during dry conditions (a period of baseflow, Figure 2C). Choosing to examine flow velocities occurring during an arbitrary period, such as within a 10-hour window, could result in a capturing a distribution of values that encapsulates opposing environmental conditions (portion of periods of stormflow in Figure 2A and baseflow in Figure 2C that are captured together in the arbitrary 10-hour window in Figure 2B).

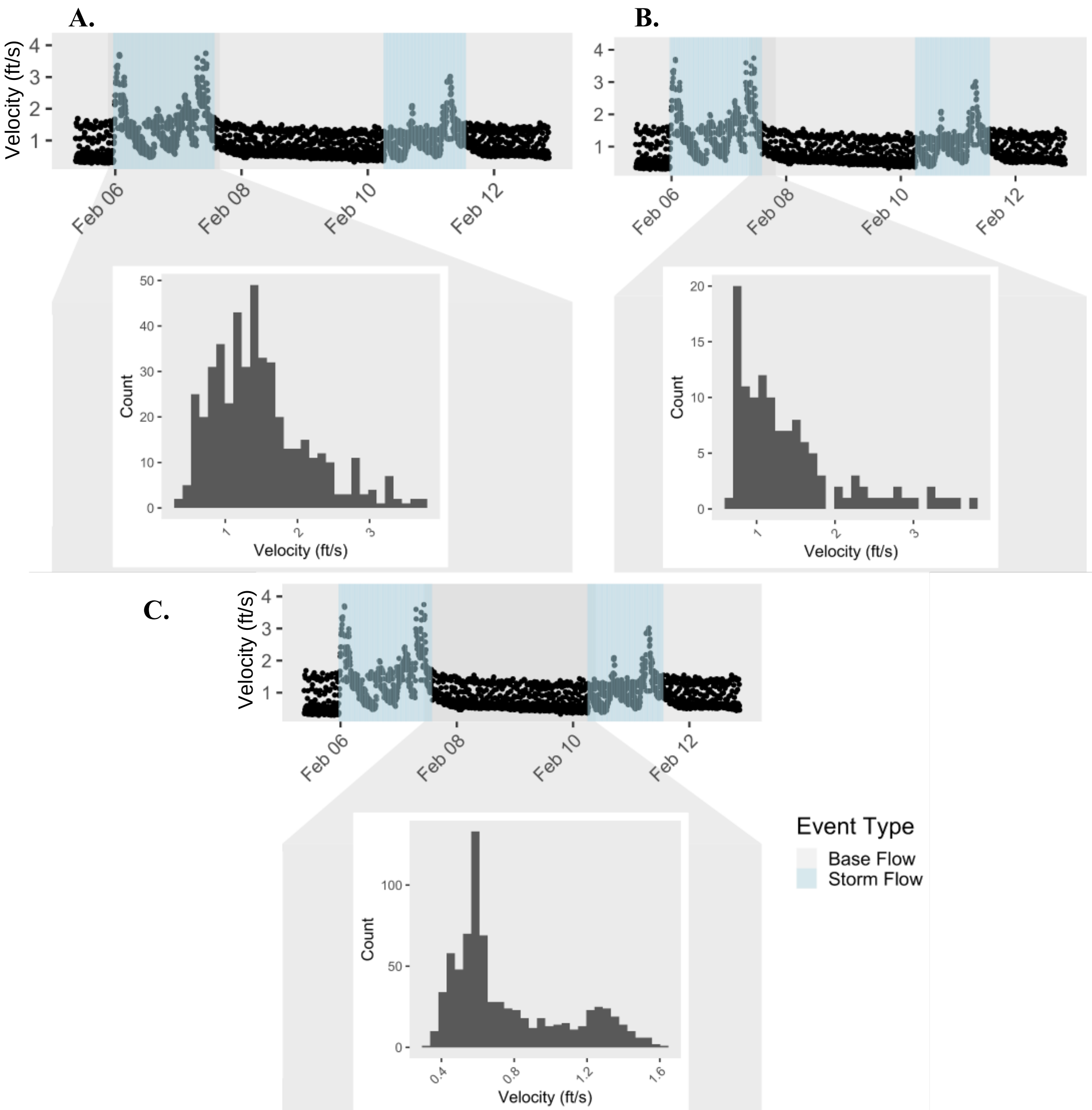


Figure 2 Distribution of measured velocity values from a constructed stormwater wetland in February 2020 **A**. during a period of stormflow, **B**. during an arbitrary 10-hour period including portions of periods storm and baseflow, and **C**. during a period of baseflow.

An event-by-event approach requires little input beyond a local precipitation and flow record to contextualize concurrent hydrological observations for the conditions during which they were collected. Considering periods of storm and baseflow over time also provide a means of remaining robust to the effects of climate change, shifting weather patterns, and watershed development. The frequency of more intense rainfall has increased over the past seventy years in the United States (Wright et al. 2019), implying that current storm characteristics are inherently different from those of the past, and that storm and baseflow periods and an SCMs response to them should be considered unique through time.

Employing statistical methods for quality assurance within stormflow (defined as coincident precipitation) and baseflow (defined as between precipitation events) periods provides an opportunity to limit potential false identification of anomalies in hydrological data that would occur due to the comparison of stormflow and baseflow periods, which are characteristically incomparable (Poff et al. 1997).

An automated QAQC model, presented here, solves these limitations by harnessing the stochastic nature of hydrological data to make efficient, data-informed, and statistically rooted inferences for handling poor quality data on a retrospective, event-by-event basis. This model is controlled, efficient, and repeatable with any continuous dataset containing observations for hydrological data with concurrent rainfall and flow data. By applying this model to sensor-derived examples using hydrological data, its capability of correcting such data without generalizing about the time and space in which it was collected is demonstrated. The resulting product is free of bias and the parameters for correction are mutable, which organizations can specify additional conditions for the correction of their data as deemed necessary. In all, applying this framework aids in mitigating the deleterious effects that poor-quality data can have on modeled data. The resulting product of this model, in which the quality of the observations within is assured, can be used to derive inference in modeled systems with higher confidence, reducing the monetary and humanitarian cost associated with systems that could be developed based on inherently erroneous data.

2 Methods

Employing this automated model framework for the correction of uncharacteristic hydrological data is achieved in a sequence of five steps. In the first step, a rainfall series for the location of collection of the uncharacteristic hydrological data to be corrected is used to assign periods of rainstorm (stormflow) and dry (baseflow) conditions. A flow record can be used in conjunction with the rainfall record as an added means of assigning storm and baseflow conditions at the end of rainstorms but is not required. A rainfall record is required at minimum. An example application using a rainfall and flow record together will be considered here. In the second step, unique periods of storm and baseflow are assigned. A distribution of values within each unique period of storm and baseflow is considered for the purpose of identifying outliers. Before outliers are to be corrected, a regression of the cumulative rainfall of each period of stormflow is mapped against the concurrent hydrological observations as a means of validation. Upon validation, unexplained outliers are corrected using a baseflow-period mean or stormflow-period median (due to skew during storm events) for each event identified in the series of data. A full illustration of the forthcoming steps in this model is outlined in (Figure 3).

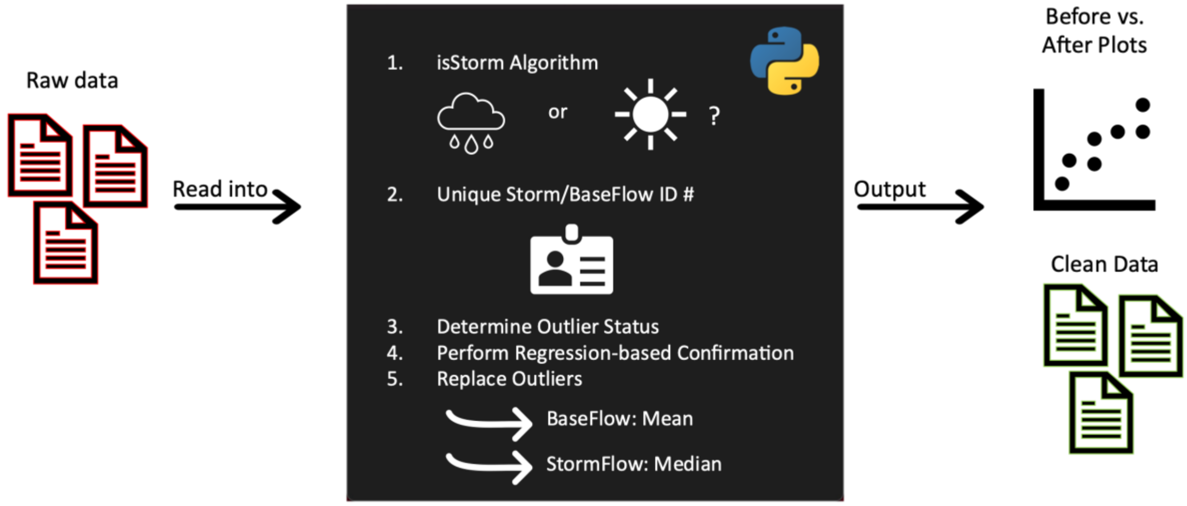


Figure 3 QAQC model overview. First, raw data is read into a python script in the proper format. Then, storm and baseflow events are assigned using the rainfall data in the input (steps 1 and 2). Outliers are determined by assigning z-scores to the distribution of values for each of the hydrological variables in the input on an event-by-event basis (step 3). Then, regression-based confirmation is performed for each storm event by fitting a 1st through 7th order polynomial between the cumulative rainfall throughout each storm and the concurrent hydrological variable's observed value, returning the original data if more than 95% of the observed data can be explained by one of the fitted curves (step 4). For periods of stormflow not explained by regression-based confirmation, outliers are replaced with the median value for such events. For periods of baseflow, outliers are replaced with the mean value for each event (step 5). Plots illustrating the changes made to the input data are produced, along with output data.

This automated QAQC model involves the identification and correction of poor-quality water quantity data aided by data transformation with the Pandas python package. The Pandas package provides a framework for data analysis and manipulation using dataframes which, facilitated by the design of the python language, streamlines large-scale analysis of data with mixed types. Pandas also provides functionality to parallelize operations applied to dataframes, which enables this model to function efficiently on a personal machine.

2.1 Data Formatting

Columns specify the datetime with observations of rainfall and flow at each timestamp. Rainfall and flow must have complementary SI or empirical units. Concurrent data for any number and type of water quantity variables are then included with observations at each timestamp, leaving missing values blank. Columns have a header in the first row of the file detailing the variable each column represents. Three mandatory columns detail the Datetime, Flow, and Rainfall observation used to assign periods of storm and baseflow. Datetime should be in a standard format detailing the date and time of the observation.

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. Pandas provides flexibility for the input file type and format of the data via packaged methods. Demonstrated here, Excel workbooks were used to read uncharacteristic data into the script used for this model. The format mentioned is not required but suggested for the formatting of dataframes within the python environment, as it is the most conducive to the operations performed in the following steps.

Once this pre-processed data has been formatted appropriately, it can be read into Pandas dataframes via the python script. Timestamps in the Datetime column are formatted and set to serve as the index for the dataframe. Rainfall data is transformed into a Boolean value that is true during a precipitation event and false otherwise. The same process is repeated for flow data, with nonzero flow as true and zero flow as false.

2.2 Retrospective Storm and Baseflow Period Assignment

Periods of stormflow and baseflow are then assigned using the following logic (Figure 4): A unique period of stormflow begins when there is any amount of instantaneous rainfall, with subsequent periods of rainfall with internal dry periods lasting less than six hours, which began at least six hours after the final instance of rainfall of the previous storm. Similarly, a unique period of stormflow ends when there are at least six hours until the next instance of rainfall. Criteria can require that flow return to below a threshold (e.g., 0.1 cfs) after the final instance of rainfall of the current storm before ending stormflow conditions. Otherwise, there is a unique period of baseflow.

While this logic is specific to the climatic conditions adopted by the Villanova Center for Resilient Water Systems (VCRWS) based on nationally accepted storm criteria (Driscoll et al. 1989), this model provides flexibility for the specification of other location-specific conditions. For instance, stormflow conditions can specify a required rainfall depth, duration, and/or maximum allowable dry time throughout the course of the identified event. This model can be performed without the need of a flow record, for SCM's that do not collect such information, by removing the additional flow criteria. Alterations to the decision criteria (yellow diamonds in Figure 4) can be made for an investigator's location-specific parameters. The decision for flow can be bypassed in the instance of an SCM that does not collect such data or altered to better suit climatic conditions.

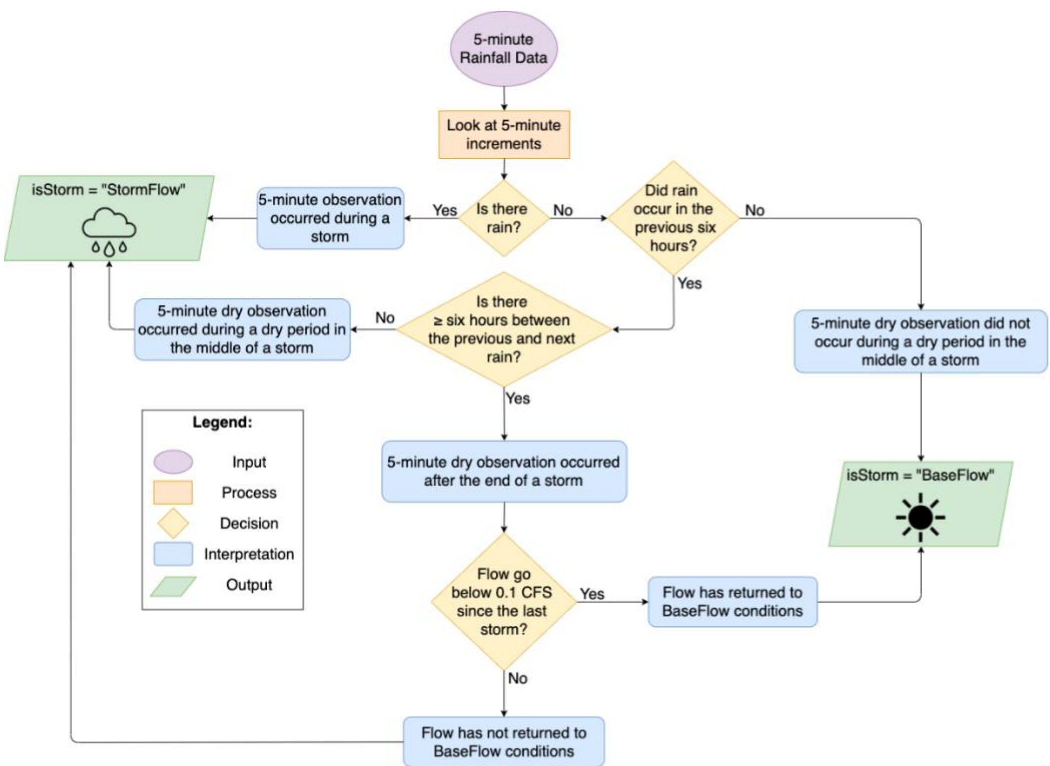


Figure 4 Logical decision-making process for determining when periods of storm and baseflow occur using flow and rainfall data at each timestamp.

2.3 Outlier Detection

Unique periods of storm and baseflow are assigned a unique identifier by appending an integer to the storm or baseflow designation assigned in the previous step. A z-score is calculated for the values of each variable of interest within the identified periods of storm and baseflow. Values with a z-score with an absolute value of greater than three are outliers (Saleem et al. 2014). The median of each period of stormflow (Figure 5A) and the mean of each period of baseflow (Figure 5B) is then used to impute missing values and detected outliers (outlined in Figure 6). Rainfall data is often stochastic in nature (Sivakumar 2017), skewing observations made during periods of stormflow. It is therefore better to impute values using the period-specific median in periods of stormflow, as it is better suited for skewed data (Hadeed et al. 2020). Rows of data containing any remaining missing data are removed, as they occur as a lack of available data to determine a mean or median to fill or replace values with for that period.

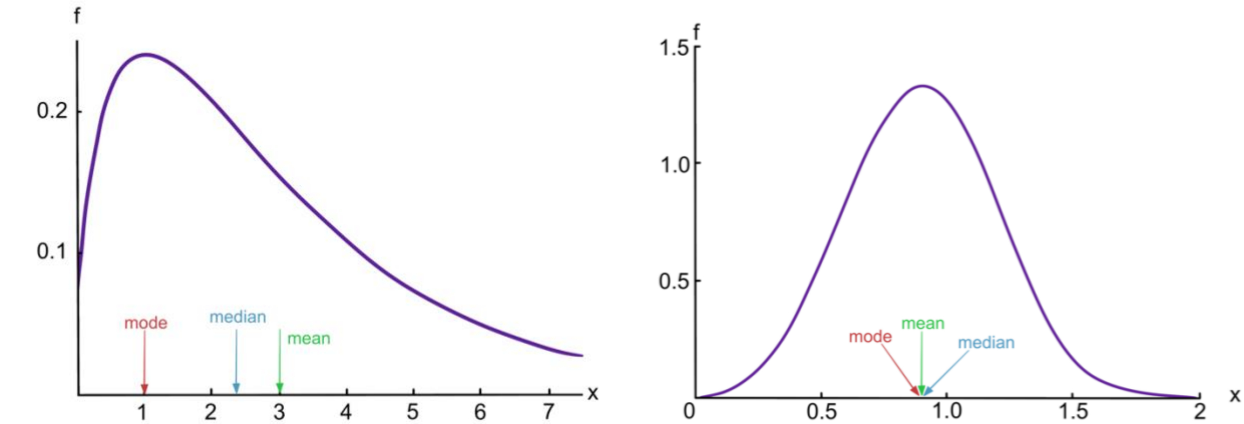


Figure 5 Illustration of mean, median, and mode for **A.** skewed data (typical of the distribution of hydrological data during stormflow) and **B.** normally distributed data (typical of hydrological data during baseflow) (adapted from von Hippel 2005).

The conditions for outlier detection are mutable. For instance, the standard threshold for outlier designation using a z-score is a z-score greater-than an absolute value of three (three standard deviations from the mean). Sensitivity to variability in the dataset, by either increasing or decreasing the threshold z-score for outlier detection, is allowable. Additionally, in climates where rainfall patterns are more consistent (i.e., rainfall intensity is less variable), a mean can be used for imputation during stormflow conditions. Additional parameters can be specified to constrain values within a known physical limit of measurement using upper and lower limits of detection.

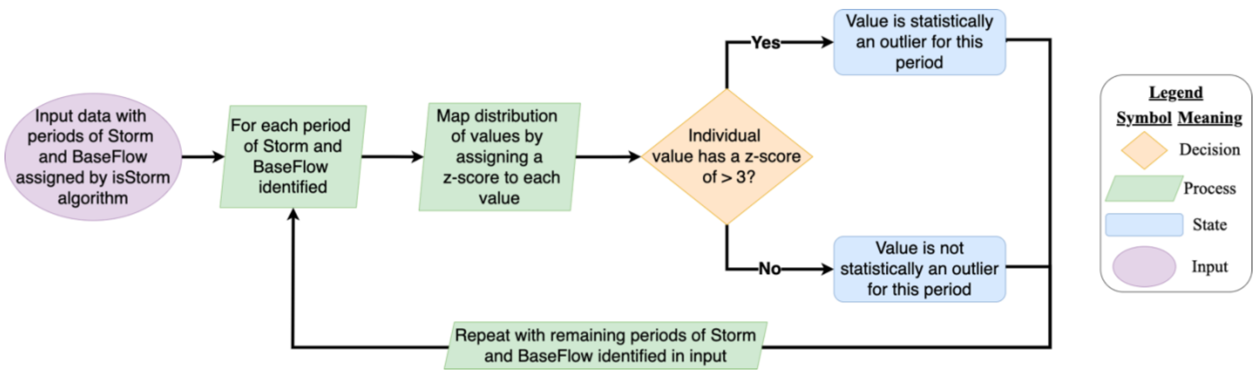


Figure 6 Overview of QAQC outlier detection step.

2.4 Regression-based Confirmation

Sensor-derived data often contains excess noise, which can lead to false outlier detection. As these values provide important insight into system function, it is imperative to avoid instances of improper exclusion (Basu and Meckesheimer 2007). This can impact not only the observation of primary data, but also subsequent calculations that characterize a site.

For example, the relationship between cumulative rainfall for a given storm event and conjugate generated runoff is well understood by the National Resources Conservation Service Curve Number (NRCS CN) method, which relies on several primary data points (Aron et al. 1977). Despite its limitations, the NRCS CN method highlights an important feature of climatic conditions and hydrology through its development: developing empirical models to explain physical processes (Baiamonte 2019). This analysis builds on the framework of characterizing physical processes using concurrent climatic conditions by building a regression between the cumulative rainfall of stormflow events and concurrent observations of physical processes.

In an additional analysis, the process previously illustrated for stormflow and baseflow period identification was performed with a regression fitting step prior to outlier identification (outlined in Figure 7). Regressions (first through seventh order) were fit to each period of stormflow to assess whether the water quantity variable of interest could be explained by cumulative rainfall. If 95% of the observed data (R-Squared value of >0.95) for a stormflow period could be explained by one of the nonlinear regressions mapped to that period, outliers were not replaced with outlier removal steps, as an empirical model explains the observed data with high confidence.

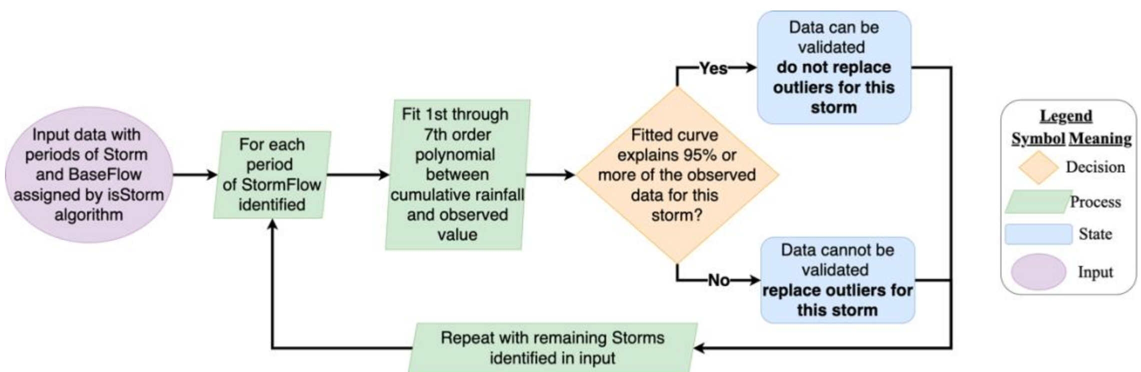


Figure 7 Process diagram for the regression-based confirmation step of QAQC. Once periods of storm and baseflow have been assigned in the input, this process is performed.

2.5 Model Performance Analysis

Plots are constructed using the matplotlib package to show the distribution of each water quantity variable in the dataset before and after processing with this model. These figures also delineate stormflow and baseflow periods during the observation window and display changes made to initial reported values. Summary statistics are computed to determine the effect size of the quality assurance model on key data quality metrics using both pre- and post-processed datasets in R. Data quality is assessed for each water quantity variable before and after processing with this model by assessing the following parameters: number of missing and extreme values, minimum, maximum, standard deviation, and average of each quantity value. These results were compared for datasets with and without regression verification for outliers. Before correction, population medians of measured values during periods of storm and base flow were compared to those of sample of medians of arbitrary time intervals of 1-day, 12-hours, 6-hours, 3-hours and 1-hour using a Mann-Whitney test. The number of intervals not significantly different from storm and baseflow, and the number of groups significantly different from both storm and baseflow was reported. This analysis was performed to assess whether arbitrary time intervals for grouping values can capture the distribution of values that occur when grouping by storm and baseflow periods as employed by the model described herein.

3 Results of Use Cases

The capability of this model at improving the correctness, completeness, and consistency of poor-quality data was demonstrated with two case studies. In both case studies, data from SCMs managed by VCRWS, each equipped with sensors producing 5-minute hydrological data, was QAQC'd. The data from each SCM contained uncharacteristic values because of sensor malfunctions and power issues at their respective sites. Each case study considered correcting values with and without a regression-based confirmation for comparison.

3.1 Case Study 1: Flow velocity data from a Constructed Stormwater Wetland

There is monitored water quality and quantity data from the Constructed Stormwater Wetland (CSW) located on Villanova University's campus. The CSW receives runoff from an 18.2 ha (45 ac) watershed with 53% impervious surfaces. Flow enters through two merging inlets, then passes through a series of three meanders, before exiting the system at an outlet preceded by a finishing pond. Several sensors have been installed to monitor the flowrate (via a depth-velocity sensor) of runoff in each of the two pipes. Data detailing the velocity of flow from one of the inlet pipes is used in this example. The flow meter is a BlueSiren® FlowSIREN sensor ("FLOWSiren® PRO" n.d.). Several challenges emerge when attempting to collect quality data from this sensor. As flowrate is derived from two separate measurements (depth and velocity, related by area-velocity), each value's error is propagated in the calculated result. This sensor is also physically difficult to reach, as it is located inside an underground pipe that leads directly into the CSW's inlet forebay, adding challenges to sensor maintenance. Additionally, there have been issues routing and maintaining power to this sensor, which has historically led to negative impacts on sensor function. Previously, as a workaround, flow data derived from a SWMM (Stormwater Management Model) model for the CSW was used to supplement questionable and/or missing data from this sensor. However, while useful, modeled data lacks the same inferential power that real-time data can provide. A use case of velocity data from this sensor collected between January and June 2020 will be used to explore this automated quality assurance and control to overcome these challenges.

Table 1 Parameters for the correction of velocity data from the Inlet of the CSW. All parameters were retained as their default as stated in the Methodology, aside from the addition of a parameter specifying the correction of negative velocity values.

Parameter	**Specification**
Storm and Baseflow Designation	Default conditions described in (Section 2.2, *Retrospective Storm and Baseflow Period Assignment*)
Outlier Status	• Z-score of >3 or <-3 • Negative measured velocity
Value Replacement	• Median stormflow velocity when not validated by regression-based confirmation • Mean baseflow velocity

The parameters used in the correction of velocity data from the Inlet of the CSW are summarized in (Table 1). Default parameters, described in the Methodology were used, aside from an additional parameter specifying the correction of negative velocity values, which this sensor has been known to record under low-power conditions.

Table 2 Assessment of difference between population median velocity for storm and baseflow in comparison to sample medians of velocity values from arbitrary time intervals before processing with this model using Mann-Whitney test (α = 0.05). Number of intervals in each category is reported.

Time Interval	**No Significant Difference from Storm Flow**	**No Significant Difference from Base Flow**	**Significantly Different from both Storm and Base Flow**
1-day	2 (1.1%)	33 (18.1%)	147 (80.8%)
12-hours	7 (1.9%)	92 (25.3%)	265 (72.8%)
6-hours	20 (2.7%)	268 (36.8%)	440 (60.4%)
3-hours	59 (4.0%)	699 (48.0%)	698 (48.0%)
1-hour	308 (7.0%)	3035 (69.5%)	1025 (23.5%)

For each of the arbitrary time interval groupings identified in (Table 2) except the 1-hour interval, the groupings have median depth values that are inconsistent with the population median storm and base flow depths. The proportion of time intervals falling under the "No Significant Difference" increases with shorter arbitrary intervals, with roughly tenfold more intervals consistent baseflow than with storm flow for their respective time interval. The proportion of time intervals falling under the "Significantly Different from both Storm and Base Flow" category decreases with shorter time intervals.

The result of identifying and correcting velocity data from this sensor is shown in Figure 8. The model mostly corrects observations that occur during rainstorms (gray-shaded portions Figure 8) and occasionally during baseflow periods (unshaded portions of Figure 8). Where corrections are made, velocity values are typically reduced by less than 10 ft/s with changes of between 10 and 100 or more ft/s occurring with much lower frequency (blue stars in Figure 8, on right-hand axis, "Velocity Change (Before – After)").

The moving 1-day average velocity of runoff entering the CSW at Inlet West is typically less than 10 ft/s both before and after processing with this model (red and black lines in Figure 8, on left-hand axis, "Velocity (ft/s)"). There are instances of instantaneous velocity measurements of greater than 100 ft/s in June 2020 before processing with this model (orange, dotted line in

Figure 8, June 2020). This identification and correction model typically makes no changes to velocity during periods of baseflow, with only two instances of alteration during baseflow conditions in May and June of 2020 (blue stars in Figure 8, on left-hand axis, "Velocity (ft/s)" in unshaded portions for baseflow). During stormflow, alterations to velocity measurements flagged as outliers have a greater effect on the moving 1-day average velocity, typically reducing the moving average velocity by more than 1 ft/s (black and red lines in Figure 8, on left-hand axis, "Velocity (ft/s)"). One such instance with significant reduction in this value can be seen in the beginning of June 2020 during a period of baseflow that preceded a storm, where instantaneous velocity was reduced by more than 100 ft/s (blue stars in Figure 8, on left-hand axis, "Velocity (ft/s)", at the beginning of June 2020)

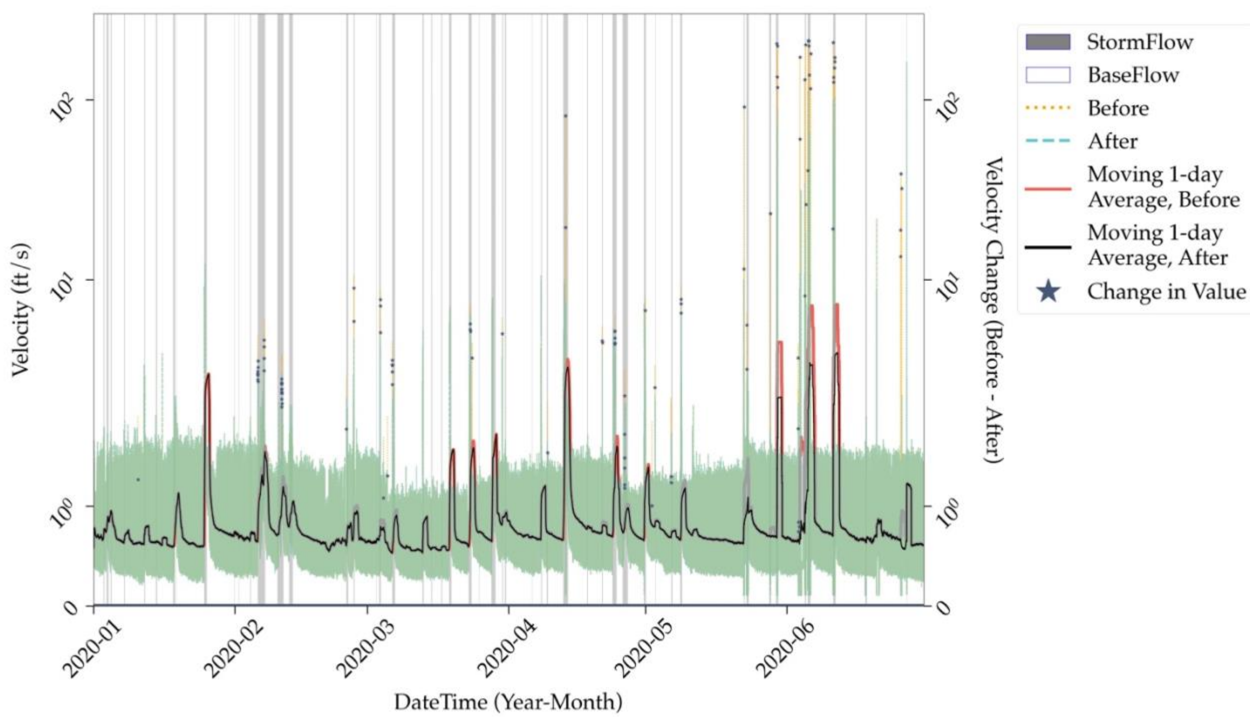


Figure 8 Distributions of the instantaneous and rolling, 1-day average Velocity before and after processing with regresssion confirmation. Instantaneous velocity values are shown by the orange (Before) and green-blue (After) lines with values on the left axis. Moving 1-day average before and after processing are shown with red and black lines, respectively. The change in value Before and After processing is shown with blue, star-shaped markers on the right axis. Periods of Storm (shaded) and BaseFlow (unshaded) are filled.

Regardless of whether regression-based confirmation is used, this model reduces the number of outliers and missing values in the processed dataset (Figure 9, Number of Outliers and Number of Missing Values, “After” and “Regression”). There is an apparent reduction in the average, maximum, and standard deviation from the original dataset’s velocity values for both processing methods (Figure 9, Average, Maximum, and Standard Deviation, “After” compared to “Regression”). Regardless of processing method, there is no alteration in minimum observed velocity (Figure 9, Minimum, “Before,” “After,” and “Regression”). Regression-based confirmation does not result in any further reduction in outliers when compared to regression-exclusionary iterations (Figure 9, “Regression” distribution in comparison to “After” and “Before” distributions). Outliers are corrected in the same manner as those in non-regression-based iterations (when < 95% is explained by the regression) and explainable outliers are retained, producing a more characteristic illustration of site performance (Figure 9, “Regression” distribution in comparison to “After” and “Before” distributions).

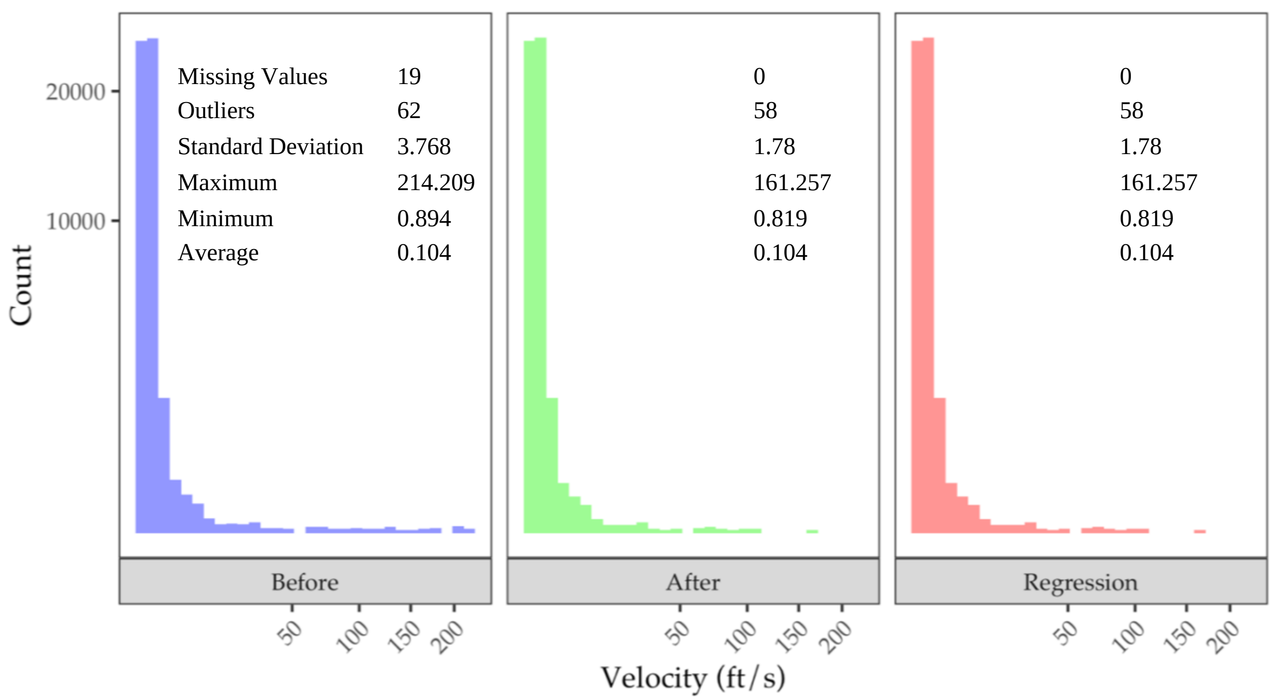


Figure 9 Frequency of Velocity values Before processing, labeled with key statistics.

3.2 Case Study 2: Depth data from a Redeveloped Site

Villanova University recently (2019) completed the redevelopment of a 100% impervious parking lot with no stormwater management into a student housing facility known as The Commons. The Commons redevelopment project reduced the impervious surface area by roughly 40% and added 15 SCMs both above ground (rain gardens) and underground (water harvesting cisterns and infiltration beds). Most of the precipitation that falls on the site is conveyed to at least one stormwater SCM to promote infiltration and ultimately attenuate peak flow rates and volumes leaving the site. Runoff depth and velocity was monitored at the downstream portion of a re-developed site at a pipe to understand how the hydraulics and hydrology have changed with the newly implemented stormwater control measures.

A Greyline Instruments® OCF 5.0 area velocity flow meter ("Greyline Instruments OCF 5.0 Open Channel Flow Monitor | Ultrasonic Flow Meters | Instrumart" n.d.) is used to collect the data, and the data stored using a Campbell Scientific® CR 800 datalogger. The monitoring system is powered by a 12-V DC battery with a 20-W and subsequently 100-W solar panel. When the system loses power and starts back up again, the readings can default to the system maximums until the system is manually reset; in this case, 18 in (45.7 cm) pipe diameter and 12 ft/s (3.66 m/s) velocity. Monitoring precipitation within the vicinity of the flow meter allows for data validation through this QAQC model. During times of baseflow, depths greater than four inches and less than three inches were not realistic due to the dam placed behind the sensor. A shallow concrete dam was constructed behind the AV sensor which requires a minimum depth of at least one in for accurate readings and to maintain a more laminar flow. Unlike the long-established dataset from the CSW, the dataset from The Commons is relatively new because post-construction monitoring was only started in June 2020.

To demonstrate the capability of this automated QAQC model at compensating for the issues with the depth sensor at The Commons, 5-minute water depth from June through December 2020 were analyzed. During this time, the sensor was known to have issues related to losing power, resulting in the sensor reading a depth of 0 in when there was known to have been a measurable depth of water in the pipe.

Table 3 Parameters for the correction of depth data from the 18 in pipe of the reconstructed site. All parameters were retained as their default as stated in the Methodology, aside from the addition of a parameter specifying the correction of negative and zero depth values.

Parameter	**Specification**
Storm and Baseflow Designation	Default conditions described in (Section 2.2, *Retrospective Storm and Baseflow Period Assignment*)
Outlier Status	• Z-score of >3 or <-3 • Negative and zero measured depth
Value Replacement	• Median stormflow depth when not validated by regression-based confirmation • Mean baseflow depth

The parameters used in the correction of depth data from the 18 in pipe at The Commons are summarized in (Table 3). Default parameters described in the Methodology were used, aside from an additional parameter specifying the correction of negative and zero depth values.

Table 4 Assessment of difference between population median depth for storm and baseflow in comparison to sample medians of depth values from arbitrary time intervals before processing with this model using Mann-Whitney test (α = 0.05). Number of intervals in each category is reported.

Time Interval	**No Significant Difference from Storm Flow**	**No Significant Difference from Base Flow**	**Significantly Different from both Storm and Base Flow**
1-day	8 (4.7%)	9 (5.4%)	151 (89.9%)
12-hours	27 (8.1%)	27 (8.1%)	281 (83.9%)
6-hours	58 (8.7%)	62 (9.3%)	548 (82.0%)
3-hours	147 (11.0%)	187 (14.0%)	1001 (74.9%)
1-hour	695 (17.3%)	892 (22.3%)	2414 (60.3%)

For each of the arbitrary time interval groupings identified in (Table 4), most of the groupings have median depth values that are inconsistent with the population median storm and base flow depths. The proportion of time intervals falling under the “No Significant Difference” increases with shorter arbitrary intervals, with roughly equal proportions of intervals consistent with either storm or base flow for their respective time interval. The proportion of time intervals falling under the “Significantly Different from both Storm and Base Flow” category decreases with shorter time intervals.

The result of identifying and correcting depth data for the 18 in pipe at the redeveloped site is shown in Figure 10. The model mostly corrects observations that occur during rainstorms (gray-shaded portions Figure 10) and occasionally during baseflow periods (unshaded portions of Figure 10). Where corrections are made, depth values are typically changed by less than 1 inch with changes of between one and ten inches occurring with much lower frequency (blue stars in Figure 10, on right-hand axis, “Depth Change (Before – After)”).

The moving one-day average depth of water in the 18-inch pipe at the redeveloped site is typically less than 10 inches both before and after processing with this model (red and black lines in Figure 10, on left-hand axis, “Velocity (ft/s)”). This model typically makes changes in the moving, 1-day average depth during periods of baseflow, with many instances in both September and November 2020 where corrections lead to an increase in the 1-day moving average depth (red and black lines in Figure 10, on left-hand axis, “Depth (in)” in unshaded portions during September and November 2020). In instances where there was an alteration in the moving one-day average depth, the average was increased to within the expected depth range of between three and four inches. During stormflow, alterations to depth measurements flagged as outliers have a larger effect on the moving, one-day average depth, increasing the moving average depth by more than one inch (red and black lines in Figure 10, on left-hand axis, “Depth (in)”, in shaded portions in August, September, and October 2020). In most instances of missing data, erroneous instances where there is a measured depth of zero (Figure 10, red line “Before” on left-hand axis “Depth (in)”), this model compensated by filling with the appropriate mean for baseflow periods or median for stormflow periods. Only one instance could not be compensated for, when a large swath of missing data coincided with an extended period of baseflow (Figure 10, red and green lines on left-hand axis, “Depth (in), mid-July 2020”) and a representative mean could not be determined to replace the missing data.

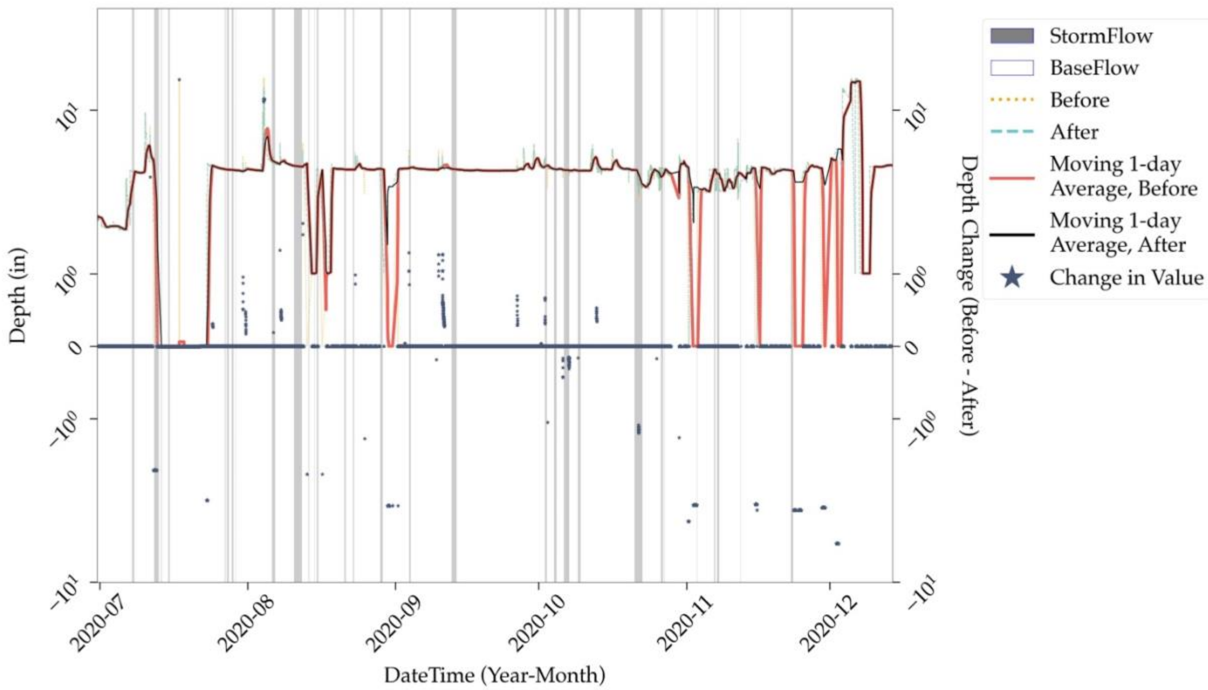


Figure 10 Distributions of the rolling, 1-day average Depth before and after processing with regresssion confirmation. Instantaneous Depth values are shown by the orange (Before) and green-blue (After) lines with values on the left axis Moving 1-day average Depth values are shown by the red (Before) and black (After) lines with values on the left axis. The change in value Before and After processing is shown with blue, star-shaped markers on the right axis. Periods of Storm (gray) and BaseFlow (white) are filled.

Regardless of whether regression-based confirmation is used, there is no reduction in the number of missing values, as the logger recorded a value of zero in place of missing data (Figure 11, Number of Missing Values, “After” and “Regression”). Both iterations remove outliers from the distribution, with each iteration having the same number of outliers in its distribution (Figure 11, Number of Outliers, “Before,” “After,” and “Regression”). Regardless of method, there is no alteration in the average, minimum, and maximum depth, and the standard deviation of the distribution (Figure 11, Average, Minimum, Maximum, and Standard Deviation, “After” and “Regression” with respect to “Before).

Outliers are corrected in the same manner as those in non-regression-based iterations (when < 95% is explained by the regression) and explainable outliers are retained, producing a more characteristic illustration of site performance (Figure 11, “Regression” distribution in comparison to “After” and “Before” distributions). In this scenario, there is no difference between the non-regression and regression-based confirmation iterations (Figure 11, Average, “After” and “Regression”).

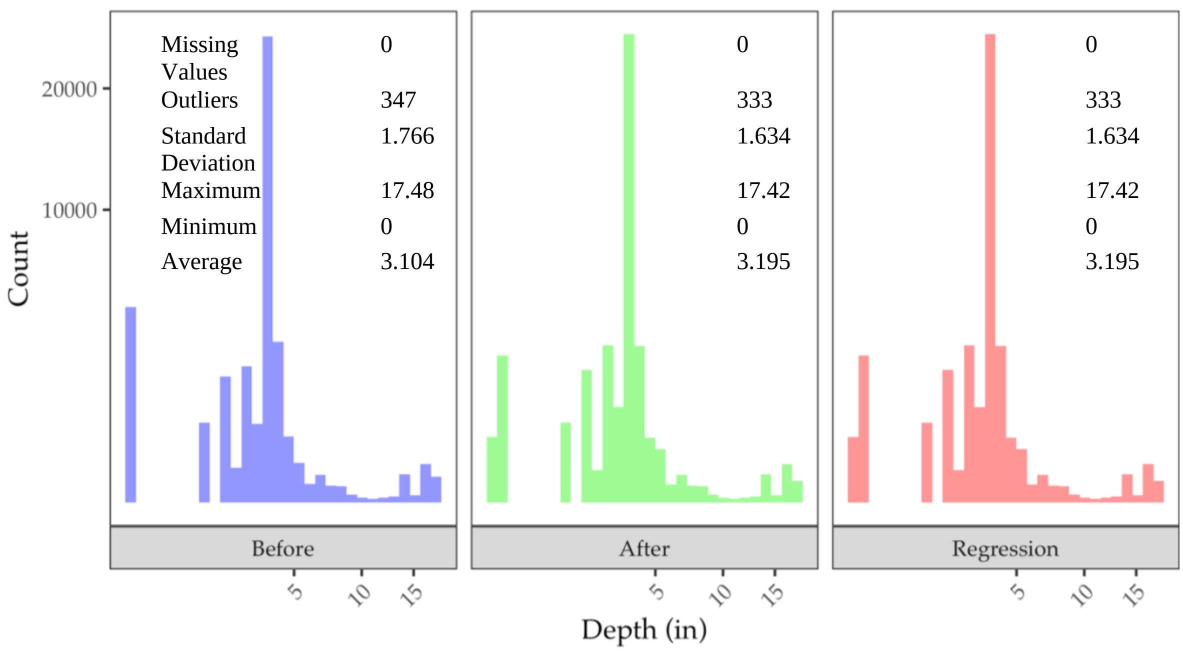


Figure 11 Frequency of depth values before processing, labeled with key statistics.

4 Discussion

This automated QAQC model proves useful in many situations, not the least of which is sensor malfunction. Errant operation allows for the corruption of otherwise reliable datasets through the introduction of incorrect, negative, or even missing values. Though traditional statistical methods or even manual data manipulation can be utilized to remove these as outliers, reduction in the size of a dataset provides a conjugate reduction in inference power. With the aforementioned quality assurance model, there is improved maintenance of large datasets by reducing the number of excluded points through the understanding of reasonable extremes (i.e., those that likely hold significance rather than improper skew) and the capacity for selective interpolation. While most methodologies treat points in a given dataset without discrimination, this model considers external environmental factors that impact sensor readings of interest. As a result, output values can be categorized, allowing for individualized interpolation where necessary. The processed data, therefore, is unique to the system at that point in time.

In the two use-case studies outlined for testing the power of this model at improving data quality, the forthcomings and limitations of this model were demonstrated. In the case where velocity data from the CSW was corrected, this model demonstrated the ability to compensate for poor quality, while remaining consistent with the original dataset. In the iteration where regression-based confirmation was used, extreme, but not necessarily incorrect, data points were maintained from the original dataset. In comparison to the iterations where regression-based confirmation was not used, there was no difference in the resulting dataset. This is largely due to extreme, non-characteristic values having occurred during periods of baseflow, and being flagged as outliers and removed as such. In one such instance, a velocity reading of 181.157 ft/s that occurred hours after a previous storm ended was changed to the average of that period of baseflow. Regression-based confirmation is only employed in periods of stormflow where the relationship between the observed rainfall and concurrent hydrological observations can be determined, therefore there is no way to confirm if the extreme values during periods of baseflow are valid. In instances where values were changed during periods of stormflow (i.e., when the observation was an outlier and <95% of the observations for that storm could be explained by the concurrent rainfall), there is added confidence for the value that was changed.

The resulting dataset from correcting velocity data from the CSW demonstrates good performance under consideration of the resulting values of correction that occurred during rainstorms of the early summertime, which are characterized by short, flashy frontal rainstorms that are being more intense with shifts in the climate paradigm (Schlef et al. 2021). This model accounts for observations made during such storms, as demonstrated by the more frequent, large corrections that occurred during the storms of late May and June 2020 at the CSW (Figure 8). Regression-based confirmation could not determine a relationship between the observed velocity and the concurrent cumulative rainfall with great enough confidence (Figure 9), leading to the outliers observed during these storms to be reduced to a more characteristic value.

Contextually, based on the physical limitations of the depth of water that can be measured in the pipe, the resulting dataset produced by the depth in the 18-inch pipe at the redeveloped site demonstrates improved data quality. There is a decrease in the frequency of measured depths of zero inches in the eighteen-inch pipe regardless of whether regression-based confirmation is performed, indicating compensation for a known impossibility (Figure 10). There is an increase in the consequential average depth over the duration of the scenario (Figure 11). However, this scenario demonstrates another shortcoming of this model; correcting data where this is no aggregable mean or median for a period of storm or baseflow. While this is a rare occurrence, having only occurred once over the six-month period examined here, it presents a potential issue for arid climates, where sensor malfunction could occur concurrently with extended periods of no rainfall. Using the known physical limitation for the depth in the pipe (being that the pipe is 18 inches in diameter) adds confidence to the resulting dataset. In future iterations, it is proposed that this model incorporates measurable limits for sensors where possible as an added outlier detection mechanism. This way, the known possible values of data from the sensor can help to identify true anomalies.

From the use cases considered above, several inferences can be made regarding model performance. When regression-confirmation is applied, there is a greater reported instance of maximum-value retention. In datasets where this value is likely a result of extreme weather, it is less likely to be considered errant or outlying relative to data in a temporal locality. There is additional evidence that this method tends towards the preservation of average and standard deviation values of initial, raw datasets, though there is not enough support to conclude that with great confidence (Figure 9, Figure 11). As is often the case, these trends can be seen with increasing magnitude as the datasets to which they apply grow large. The confidence of changes made by the quality assurance model, therefore, can be understood to be at its highest when data input (from both sensors of interest and environmental considerations) is robust.

Regression-based confirmation and event-by-event consideration demonstrate two often overlooked aspects of existing quality assurance models. As shown here, simply detecting outliers and replacing their values is not enough to determine that they are invalid. Performing a regression-based confirmation step provides validity to the observed without making assumptions about the underlying data being invalid is automatically incorrect for being extreme. No outliers that were identified in the iteration without regression-based confirmation were then retained by regression-based confirmation, but it is expected that this is a possibility. In instances of extremely high intense rainfall, this possibility is likely to occur, as the sudden change in the cumulative rainfall should correlate well with flashes in observed hydrological processes, thus leading to these values being retained. This underscores the importance of event-specific conditions. While rainstorms can be classified by their recurrence interval, observed intensity, and depth over a specified period of time (McGraw et al. 2019), this does not capture the intricacies of time-specific considerations that a sequential event-by-event approach accounts for (Table 2, Table 4). As demonstrated, choosing an arbitrary time interval, and extracting the values in that window, often does not capture the observed characteristics of both periods of storm and baseflow. Choosing smaller arbitrary time intervals to group values together (of 1-hour of data at a time) for correction more often captures the conditions of periods of storm and baseflow than larger intervals of 3-hours or more, but at the expense of having fewer values to correct from. Taking a smaller sample size reduces the statistical power of the z-score test, increasing its sensitivity to outliers. At the same time, it could result in the misidentification of outliers. An outlier, which is repeated and not considered an outlier in a preceding or forthcoming interval, could be erroneously removed simply because of the time interval it falls under. Using an event-by-event by approach ensures that both a representative sample is considered for correction and that it captures the conditions representative of all observations in an event. Taking an event-by-event approach also allows for further development using artificial intelligence and machine learning techniques. Classification algorithms and deep neural network regression techniques could be used to learn the conditions where value correction is needed and apply changes when such conditions occur.

The limited requirements for performing correction of hydrological data with this model is far-reaching. This model requires only that a rainfall and (optionally) a flow record be provided for the correction of any number of hydrological variables. Amid the data revolution, where hydrological data management is moving to highly organized data structures, the simplistic data format requirements for this model make it executable with little to no need for data translation. This model also utilizes the capabilities of multi-thread processing and modern technology to provide results in an efficient manner. Processing the two scenarios, each with roughly 50,000 observations, on a machine with a 3.2 GHz 8-core CPU and 8GB of memory provided results within five to ten seconds. Processing the same two scenarios on an older, 2.7 GHz dual-core CPU with 8GB of memory provided results in one to two minutes. The interoperability between web-based applications and database backends provides an avenue for seamless interaction between stored, ready-to-correct hydrological data and this model. Python provides a simple to use, packaged interface for interacting with SQL-based databases (Joyce 2022). This model has the potential to retrieve and return data to such databases, providing the capability for executing this model without the need for command line scripts.

5 Conclusions

As the present age of research strives toward the collection and maintenance of increasingly large volumes of environmental data, it grows significantly more important and difficult to assure data quality. Though data quality is considered at length in in-situ, front-end solutions, they can only go so far to assist in the achievement of the goal of consistently high-quality data. Innumerable and often unavoidable sources of error may only be detected once datasets have already been collected, providing research bodies with the challenging task of justifiably augmenting that which is meant to appropriately characterize a site’s performance without the introduction of personal bias. This is not only difficult and time consuming, but also introduces recurring possibilities for poor research practices and unintended data manipulation. Time and space are also considered causal rather than coincidental in this model. Each period of retrospective storm and baseflow is treated as being distinctive, rather than painting broad generalizations about observations that occurred during an arbitrary period. It is based on a known rainfall record, directly related to the observed processes to be corrected in the input, rather than arbitrary climatic observations, maintaining simplicity of use.

The quality assurance framework described here is a novel method to efficiently overcome these hurdles while maintaining confidence in explainable outliers. This not only corrects for inaccurate values due to in situ sensor malfunction but allows designated leeway for realistic extreme values that would otherwise be improperly excluded by statistical methods. Additionally, the level of user input is significantly reduced when compared to manual data quality adjustment. This facet of this model is perhaps highly beneficial to a research body as it instantaneously treats all datapoints within a set (reducing temporal investment) and statistically justifies all changes. Bias is therefore removed as a possible source of error in subsequent calculations, site characterization, and further decision making, especially in the consideration of remediating natural data drift that was erroneously identified as error. This model both harnesses the power of traditional statistical methods and circumvents their shortcomings resulting in a comprehensive method by which to ensure that each dataset, regardless of size, is trustworthy, characteristic, and statistically reliable. While this model provides a framework for the automation of correcting poor-quality data, it should not be. considered a standalone replacement for traditional methods. The power of this model is fueled by and dependent upon the volume of data it serves; corrective data processes must be iterative and rely on a baseline of quality assurance metrics (i.e., sensor calibration, determination of fitness within reasonable limits, etc.). The process of correcting data should be iterative and consider calibration of additional parameters were deemed necessary until a satisfactory outcome for the correction of uncharacteristic observations is achieved.

Through utilization of data derived from the Villanova CSW and Commons, the model outlined through this paper is capable of successfully replacing and interpolating errant and nonexistent hydrologic data. In concert with traditional monitoring and data collection, this framework has the capacity to improve data resolution and confidence, enabling a dramatic shift in the approach to the collection and maintenance of environmental data.

6 Acknowledgements

We acknowledge the Villanova Center for Resilient Systems and the Villanova University College of Engineering for providing the funds to make this research possible. Author contributions: M.W.M., B.E.J., S.E.E., V.B.S., and B.M.W. wrote and edited this paper; M.W.M. designed, developed, and tested the implementation for this QAQC model; B.E.J. and S.E.E. identified and reported on sensor errors.

7 References

Aron, G., D. F. Lakatos, and A. C. Miller. 1977. "Infiltration Formula Based on SCS Curve Number." *J. Irrig. Drain. Div.*, 103 (4): 419–427. https://doi.org/10.1061/JRCEA4.0001169.

Asquith, W. H., R. C. Seanor, V. L. McGuire, and W. H. Kress. 2020. "Methods to quality assure, plot, summarize, interpolate, and extend groundwater-level information—examples for the Mississippi River Valley alluvial aquifer." *Environ. Model. Softw.*, 134: 104758. https://doi.org/10.1016/j.envsoft.2020.104758.

Baiamonte, G. 2019. "SCS Curve Number and Green-Ampt Infiltration Models." *J. Hydrol. Eng.*, 24 (10): 04019034. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001838.

Banks, G. C., S. G. Rogelberg, H. M. Woznyj, R. S. Landis, and D. E. Rupp. 2016. "Editorial: Evidence on Questionable Research Practices: The Good, the Bad, and the Ugly." *J. Bus. Psychol.*, 31 (3): 323–338. https://doi.org/10.1007/s10869-016-9456-7.

Basu, S., and M. Meckesheimer. 2007. "Automatic outlier detection for time series: an application to sensor data." *Knowl. Inf. Syst.*, 11 (2): 137–154. https://doi.org/10.1007/s10115-006-0026-6.

Blaen, P. J., K. Khamis, C. Lloyd, S. Comer-Warner, F. Ciocca, R. M. Thomas, A. R. MacKenzie, and S. Krause. 2017. "High-frequency monitoring of catchment nutrient exports reveals highly variable storm event responses and dynamic source zone activation." *J. Geophys. Res. Biogeosciences*, 122 (9): 2265–2281. https://doi.org/10.1002/2017JG003904.

Campbell, J. L., L. E. Rustad, J. H. Porter, J. R. Taylor, E. W. Dereszynski, J. B. Shanley, C. Gries, D. L. Henshaw, M. E. Martin, W. M. Sheldon, and E. R. Boose. 2013. "Quantity is Nothing without Quality: Automated QA/QC for Streaming Environmental Sensor Data." *BioScience*, 63 (7): 574–585. https://doi.org/10.1525/bio.2013.63.7.10.

Chao, L., Z. Hui, and Z. Xiaofeng. 2015. "Data quality assessment in hydrological information systems." *J. Hydroinformatics*, 17 (4): 640–661. https://doi.org/10.2166/hydro.2015.042.

Díaz, J. J., I. Mura, J. F. Franco, and R. Akhavan-Tabatabaei. 2021. “aiRe - A web-based R application for simple, accessible and repeatable analysis of urban air quality data.” *Environ. Model. Softw.*, 138: 104976. https://doi.org/10.1016/j.envsoft.2021.104976.

Driscoll, E., G. Palhegyi, E. Strecker, and P. Shelley. 1989. “Analysis of storm events characteristics for selected rainfall gauges throughout the United States.” *US Environ. Prot. Agency Wash. DC*.

Faybishenko, B., R. Versteeg, G. Pastorello, D. Dwivedi, C. Varadharajan, and D. Agarwal. 2022. “Challenging problems of quality assurance and quality control (QA/QC) of meteorological time series data.” *Stoch. Environ. Res. Risk Assess.*, 36 (4): 1049–1062. https://doi.org/10.1007/s00477-021-02106-w.

“FLOWSiren® PRO.” n.d. *Blue-Siren Inc*. Accessed April 28, 2022. https://www.blue-siren.com/products/copy-of-b-big-big-ecosiren%c2%ae-b-big-big-br-b-big-wireless-environmental-monitor-big-b-br-small-10-data-channels-with-vision-port-small.

“Greyline Instruments OCF 5.0 Open Channel Flow Monitor | Ultrasonic Flow Meters | Instrumart.” n.d. Accessed April 28, 2022. https://www.instrumart.com/products/41619/greyline-instruments-ocf-50-open-channel-flow-monitor.

“Guidance for Developing Quality Systems for Environmental Programs.” 2002. 114.

Hadeed, S. J., M. K. O’Rourke, J. L. Burgess, R. B. Harris, and R. A. Canales. 2020. “Imputation methods for addressing missing data in short-term monitoring of air pollutants.” *Sci. Total Environ.*, 730: 139140. https://doi.org/10.1016/j.scitotenv.2020.139140.

Hertwig, R., and G. Gigerenzer. 2011. “Behavioral Inconsistencies Do Not Imply Inconsistent Strategies.” *Front. Psychol.*, 2. https://doi.org/10.3389/fpsyg.2011.00292.

Horsburgh, J. S., S. L. Reeder, A. S. Jones, and J. Meline. 2015. “Open source software for visualization and quality control of continuous hydrologic and water quality sensor data.” *Environ. Model. Softw.*, 70: 32–44. https://doi.org/10.1016/j.envsoft.2015.04.002.

Hsee, C. K., J. Zhang, and J. Chen. 2004. “Internal and substantive inconsistencies in decision making.” *Blackwell Handb. Judgm. Decis. Mak.*, 360–378. Blackwell Publishing.

Joyce, P. 2022. *C and Python Applications: Embedding Python Code in C Programs, SQL Methods, and Python Sockets*. Berkeley, CA: Apress.

Kumari, P., and O. Kennedy. 2017. “The good and bad data.” *Proc. VLDB Endow.*, 11 (4).

Liu, Y., B. A. Engel, D. C. Flanagan, M. W. Gitau, S. K. McMillan, and I. Chaubey. 2017. “A review on effectiveness of best management practices in improving hydrology and water quality: Needs and opportunities.” *Sci. Total Environ.*, 601–602: 580–593. https://doi.org/10.1016/j.scitotenv.2017.05.212.

McCausland, T. 2021. “The Bad Data Problem.” *Res.-Technol. Manag.*, 64 (1): 68–71. https://doi.org/10.1080/08956308.2021.1844540.

McGraw, D., E. I. Nikolopoulos, F. Marra, and E. N. Anagnostou. 2019. “Precipitation frequency analyses based on radar estimates: An evaluation over the contiguous United States.” *J. Hydrol.*, 573: 299–310. https://doi.org/10.1016/j.jhydrol.2019.03.032.

Philbeck, T., and N. Davis. 2018. “THE FOURTH INDUSTRIAL REVOLUTION: SHAPING A NEW ERA.” *J. Int. Aff.*, 72 (1): 17–22. Journal of International Affairs Editorial Board.

Poff, N. L., J. D. Allan, M. B. Bain, J. R. Karr, K. L. Prestegaard, B. D. Richter, R. E. Sparks, and J. C. Stromberg. 1997. “The Natural Flow Regime.” *BioScience*, 47 (11): 769–784. [American Institute of Biological Sciences, Oxford University Press]. https://doi.org/10.2307/1313099.

Refsgaard, J. C., H. J. Henriksen, W. G. Harrar, H. Scholten, and A. Kassahun. 2005. “Quality assurance in model based water management – review of existing practice and outline of new approaches.” *Environ. Model. Softw.*, 20 (10): 1201–1215. https://doi.org/10.1016/j.envsoft.2004.07.006.

Saleem, A., K. H. Asif, A. Ali, S. M. Awan, and M. A. Alghamdi. 2014. “Pre-processing Methods of Data Mining.” *2014 IEEEACM 7th Int. Conf. Util. Cloud Comput.*, 451–456.

Schlef, K. E., B. François, and C. Brown. 2021. “Comparing Flood Projection Approaches Across Hydro-Climatologically Diverse United States River Basins.” *Water Resour. Res.*, 57 (1): e2019WR025861. https://doi.org/10.1029/2019WR025861.

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

Wadzuk, B., B. Gile, V. Smith, A. Ebrahimian, and R. Traver. 2021. “Call for a Dynamic Approach to GSI Maintenance.” *J. Sustain. Water Built Environ.*, 7 (2): 02521001. https://doi.org/10.1061/JSWBAY.0000945.

Wright, D. B., C. D. Bosma, and T. Lopez-Cantu. 2019. “U.S. Hydrologic Design Standards Insufficient Due to Large Increases in Frequency of Rainfall Extremes.” *Geophys. Res. Lett.*, 46 (14): 8144–8153. https://doi.org/10.1029/2019GL083235.

8 Appendices

Appendix A. Regression-based Confirmation Figures

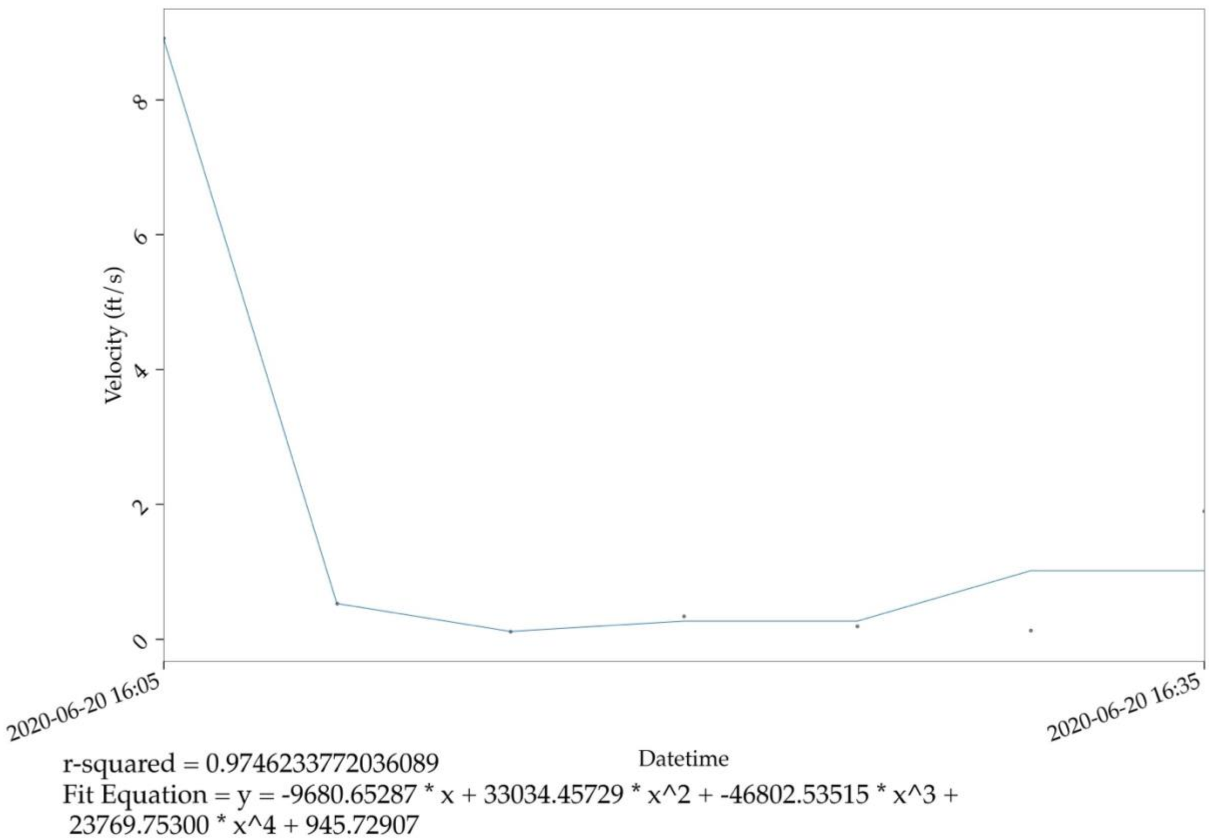
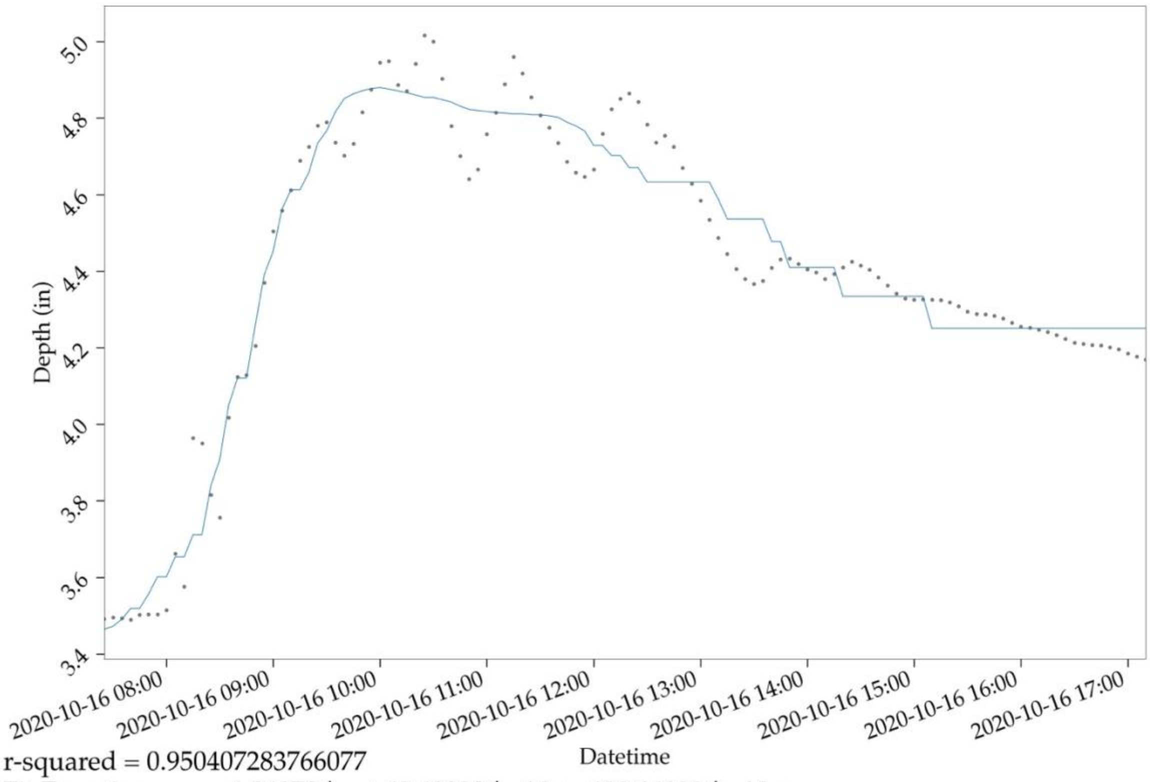


Figure A.1 Example Regression-based Confirmation Plot from the CSW. The relationship between the measure Velocity (in feet per second, on y-axis) at the CSW (Section 3.1) and concurrent cumulative rainfall depth is shown. The fourth-order polynomial confirming the relationship between the observed data and concurrent cumulative rainfall is displayed in the bottom left along with its associated coefficient of determination (R^2 value).



r-squared = 0.950407283766077

Fit Equation = y = -1.04870 * x + 63.49920 * x^2 + -132.14932 * x^3 + -438.71209 * x^4 + 1910.41471 * x^5 + -2417.33662 * x^6 + 1024.56527 x^7 + 3.46859

Figure A.2 Example Regression-based Confirmation Plot from the reconstructed site. The relationship between the measured Depth (in inches, on y-axis) at the redeveloped site (Section 3.2) and concurrent cumulative rainfall depth is shown. The seventh-order polynomial confirming the relationship between the observed data and concurrent cumulative rainfall is displayed in the bottom left along with its associated coefficient of determination (R^2 value).