QAQC Overview

Definitions

QA/QC is the combination of quality assurance, the process or set of processes used to measure and assure the quality of a product, and quality control, the process of ensuring products and services meet consumer expectations.

Quality assurance is process oriented and focuses on defect prevention, while quality control is product oriented and focuses on defect identification.

This section of the wiki describes the tools, processes, application and scripts that support both quality assurance and quality control related to RTDF data. These compliment practices and procedures of the MWQI Field Unit that are generally documented elsewhere, including the Station Manual section of the wiki.

Instrument Validation

MWQI field data are obtained directly from field sensors operated and maintained by MWQI staff. The field instruments are routinely checked and calibrated by comparisons with known standards. Samples are also submitted to the Bryte Lab for analysis and comparison.

A copy of the Field Manual is located here: \\mrsbmapp20932\db\Data_Quality\Field Manual Current.pdf.

An Excel spreadsheet has been developed to guide and capture data from routine field station QC procedures. A copy of the spreadsheet is in place at each field station.

A copy of the spreadsheet is created each time the QC procedures are followed. The copy is saved in the folder C:\RTDF\QC with the date appended to the file name. The dated report is retrieved to the RTDF computer as part of the routine, hourly polling process and the original is also moved to the subfolder C:\RTDF\QC\Archive. The reports are placed on the RTDF server in the folder \db\Import\QC.

Data Quality Codes

Occasionally, the sensors or the upstream sampling system malfunction and erroneous data is reported. Data is examined after it is imported to identify and isolate questionable data values in the database through both automated and manual procedures. “Isolation” means the data is not used in further analyses or reports: the “reviewed” data value, which is used in subsequent evaluations, is set to Null, which is equivalent to “no value.” The raw data value is retained but

not used. The isolation of questionable values at the source (i.e., in the database) eliminates the need to screen the data in each procedure using the data.

Each data observation is assigned a quality code:

Value & Meaning

0,1 = Acceptable (set to 1 for Water Data Library data for consistency with Hydstra)
2 = Questionable
3 = Unacceptable
4 = Estimated (set to 10 for Water Data Library data for consistency with Hydstra)
5 = Below RL

In addition, each data observation contains both a raw value and a value. The raw value is the value reported by the instrument. The value is used for display and analysis. If the data quality is unacceptable, the value is set to null (empty). For estimated values, raw value and value are implicitly different.

Real-time field station data is typically checked soon after it is entered into the RTDF Database with an automated data screening procedure. The automated procedure identifies only unacceptable values. Automated screening is used constantly to review new data and has been in use since 2009. Later, the data is reviewed by staff to further refine the identification and flagging of both outliers and questionable points. The Data Management Utility is used for this review. The Data Management Utility can show grab sample data from the QAQC database to aid in reviewing the data. The final review occurs before the data is exported to the Water Data Library.

WDL Discrete Samples

Grab sample data for MWQI RTDF sensors is analyzed routinely by Bryte Lab. This data is published in the DWR Water Data Library (WDL) as "discrete samples". In some cases, grab samples were collected many years before the "continuous" RTDF instruments came on line. It could be considered the "gold standard" of analyses.

The WDL data for all RTDF stations is downloaded and added to the QAQC tables periodically. The values can be viewed in the Data Management Utility.

Purpose
Data filters are processes that remove or flag data outliers.

Outlier Detection Overview

This page reviews the outlier detection schemes now in use and considers both improvement and possible alternatives to those schemes.

Automated Data Screening

Purpose

Data from field instruments operated by the MWQI, including organic carbon and anion concentrations and physical parameters such as temperature and electroconductivity are screened after they are imported to flag data values that are outliers. The flags are recorded in the RTDF database.

Overview

The database contains both a raw field (*Raw Value*) , a value field (*Value*) and a quality value (*Quality*) for each observation reported by the field sensor. *Raw Value* is never changed. *Value* is initially equal to *Raw Value* but may be changed through editing.

All raw values are assumed valid before screening. If a data value fails a screening test, *Value* is set to "Null", a database value which is equivalent to "no data" and *Quality* is set to Unacceptable. Null values are excluded from subsequent computations, such as daily averages and from data dissemination products.

RTFilter

Automated data screening is encapsulated in an application named "RTFilter."

RTFilter is incorporated in the regular cycle of field data imports. An input file (\\mrsbmapp20932\scripts\RTFilter.txt) directs its operations. It writes to a log file (\\mrsbmapp20932\scripts\RTFilter.log). These file assignments are made in the batch file that launches the application.

RTFilter employs a combination of outlier detection algorithms in sequence. Screening parameters are specified as described elsewhere under "screening parameters".

1. Values below the reporting limit for the sensor are flagged. The relevant reporting limit comes from the screening parameters and are summarized. If a point that is less than the reporting limit may later designated as an outlier, overriding the reporting limit flag.

2. Check standard values are flagged as Unacceptable. Check standard values apply only to Banks anions and are specified in application code.

3. A Min-Max screen is applied. The relevant Minimum and maximum values come from the screening parameters and are summarized. Values that are outside the min-max range are identified as outliers.

4. The autoregression lag 1 (serial correlation) method is first applied. Outliers identified previously are ignored. If the computed R^2 is greater than 0.85 and no outliers are found, no further analysis is performed.

5. If the serial correlation R^2 is less than 0.85 or the AR lag 1 analysis indicates outliers, a joint outlier detection analysis is applied. Both the k-Nearest Neighbors and Periodic Medians approaches are evaluated. Both methods must agree for an outlier to be designated. Values previously identified as outliers are ignored.

6. An outlier threshold must be greater than 5% of the median series value to eliminate the identification of outliers based on negligible differences.

Outlier Detection Algorithms

Outlier Discriminators

This page lists and reviews algorithms examined for outlier detection for the record. This includes those that did not work well and those that are still being evaluated.

Minimum and Maximum Values

Minimum and maximum expected values, if identifiable, are very effective in eliminating outliers without further, more complex analyses. In the RTDF it is used as a first step in sequence with other automated procedures for that reason.

Minimum and maximum values were established for each sensor by computing long-term statistics and by examining histograms by the review of raw observed data. Values were compared among stations with the same sensor type for validation. Values were also compared between the WDL and RTDF datasets.

Many of the sensors in the RTDF show a skew to the right in the period of record data: median values are slightly less than mean values and there is more variation above the mean than below it. The range below the mean to zero is constrained, but the range above is not.

Minimum values and reporting limits are important, related parameters. The reporting limit is based on the analyzer's capability to measure small values: below the reporting limit, the measured (observed) value is unreliable. However, it is important to know that a measurement was taken and that the

measured value was very low. The minimum value must for this reason be lower than the reporting limit, else the measurement would simply be rejected and not reported. The minimum value, then is more of a logical limit: concentrations, for example, cannot be less than zero. As noted, the Water Data Library is considered the authoritative version of the data. By convention, values below the reporting limit are not reported. Instead, a note is inserted for the value.

Minimum and Maximum Allowable Values for MWQI Stations							
Analyte	Minimum	Reporting Limit	Maximum by Station				
			Banks	Gianelli	Hood	Jones	Vernalis
DOC & TOC mg/L	0	0.5	15	15	15	15	15
Bromide mg/L	0	0.05	0.75	0.65	na	0.75	1.00
Chloride mg/L	0	10	150	130	na	150	200
Nitrate mg/L	0	1	15	13	na	15	20
Sulfate mg/L	0	10	150	130	na	150	200
EC mS/Cm	100	50		10,000			
Turbidity NTU	0	0		50			
Dissolved Oxygen mg/L	0	1		15			
Temperature deg C	0	5		30			
pH	4	4		11			

Some sensors notoriously contain more outliers than others. Minimum and maximum values are useful with these for removing outliers from graphics and reports.

General Minimum and Maximum Allowable Values by Sensor		
Analyte	Minimum	Maximum
EC mS/Cm	100	10,000

Turbidity NTU	0	250

Forward Weighted Moving Average

A forward-weighted moving average (FMA) scheme has been used since 2011 in routine, automated screening of MWQI data.

This scheme relies on pre-defined parameters rather than a statistical analysis of the series.

A min/max scan precedes the FMA analysis.

A forward-weighted moving average is computed over a specified time window prior to the value in question. If the subject value exceeds the specified allowable deviation from the computed moving average, it is rejected.

Values in the computation are weighted in proportion to their time difference from the end of the time window: a value at the end of the time window is given a weight of 1.0, a value at the beginning a weight of 0.0.

Observations with a quality of *Unacceptable* or *Questionable* are not included in the average computation.

A minimum number of values is specified as a threshold for performing the averaging computation. If the minimum number of acceptable values within the computation time window is less than the threshold, the average is not computed and no comparison is done.

If the difference in time between the last value in the computation and the value being examined is greater than the specified maximum allowable, the deviation test is not made.

This test relies on earlier data that has undergone screening. An alternative test or tests may be required if a sufficient number of valid data points is not available.

Variable	**Description**
Averaging interval	Length of time interval in days over which the moving average is computed. The end of this interval is the date-time value of the observation prior to the value being examined.
Maximum deviation	Maximum allowable deviation of the data value from the weighted moving average.

Minimum points	Minimum number of valid points to use in computing the weighted average. If the minimum number is not available, the test is not applied.
Max gap	The maximum difference in days between the value being examined and the end of the averaging interval, ie, the time value of the prior observation.

The FMA analysis moves forward in time, examining each point in turn for which an acceptable average can be computed. Points identified as outliers are not included in the moving average. The moving average weights points in inverse proportion to their displacement in time from the point being examined.

The FMA assumes a stochastic process but tolerates a large random difference between values.

The weaknesses of this scheme include:

1. Parameters are required, and they tend to be used over long periods of time and a wide variation of conditions. The parameters are necessarily based largely on judgement rather than the local statistical properties of the series. The maximum allowable deviation, for example, does not consider the actual series variability. Relatively quiescent periods exhibit a small variability, while rapidly changing conditions exhibit a much higher variability.

2. The procedure assumes a strong correlation between subsequent values, i.e., a stochastic process. It is not robust when series have a low degree of serial correlation.

Other approaches that reflect the specific statistical properties of the series are recommended.

Sequential Differences

This discriminator identifies outliers from the analysis of differences between successive values. Outliers are those values with differences exceeding the computed threshold. The computation of sequential differences and the discriminating criteria are straightforward, but the results require interpretation: not all values exceeding the threshold are outliers. The difference may identify a shift or jump, for example. Secondary tests, described below, to resolve over-identification of outliers.

The sequential differences approach assumes a stochastic process but tolerates a large random difference between values.

Decomposition is not applied prior to the application of the sequential differences test. It is assumed that a trend would be reflected in the distribution of differences.

Auto-Regression Lag 1

The Auto Regression Lag 1 (AR1) discriminator uses a series of estimates based on a linear regression of values on their previous values. Outliers are identified using the statistics of the differences between the actual and model values.

The regression procedure computes the covariance of the previous with the current value. The availability of the covariance provides a basis for indicating the degree to which the previous value is a determinant in the stochastic process: the ratio of their covariance to the product of the variances of each of the separate variances computed independently is the R^2 value. As noted above, many or most RTDF water quality series are serially correlated relatively high (>.80) R^2 values: the prior value is a good predictor of the subsequent value.

The AR1 discriminator is very reliable for R^2 values > 0.9 and even > 0.8. It works well even with a strongly tidal signal. However, like the sequential difference, it identifies any change in value that exceeds its criteria as an outlier, including shifts, jumps and values following outliers. These problems are rectified to the extent feasible by the secondary tests described below.

Kalman Filter

The Kalman filter was developed in the mid-1900s as a basis for compact and efficient automated system control. It was used successfully in the navigation of manned space vehicles and is widely used in current control systems.

The Kalman filter finds the optimum linear estimate of the state of a system by statistically combining a modeled value with a measured value, using a Bayesian analysis that considers uncertainties in both (Q and R, respectively). The uncertainties are variances treated as “white noise”.

The Kalman filter does not require a history of values: it may be used for any single pair of model and measurement values, especially when it recursively adjusts its state variables. A reasonable (approximate) estimates of both model and measurement uncertainties is required, however. It also requires an initial (approximate) estimate of the joint probability which it can adjust recursively to converge on a stable value as well as an optimal weighting between the model and measured values.

Most applications of the Kalman filter model several variables in a system of linear equations. For the RTDF, the adopted model consists simply of the previous value plus a random value, i.e., model uncertainty: a simple “black box” “model”.

The model uncertainty (variance) needs to be sufficient to allow the filter estimates to follow measurements. It is estimated as the standard error of estimate for a serial correlation (autoregressive model).

The measurement variance, Q, is the total conventional variance of the differences with the model variance removed. If the model variance is greater than measurement variance, the total variance is split between the two. The covariance between the model estimate error and the measured value error is assumed to be negligible.

The error in the Kalman prediction, the difference between the measurement and the initial (prior) model estimate, can be subjected to a Chi-Square test to assess the measurement’s validity with the standard error computed using the filter’s intermediate variables. A 99.9 % level of confidence with one

degree of freedom is the criterion for validation of measured values.

The Kalman filter is unique among all the outlier detection algorithms in several ways:

• It basically works with a single prediction if recursion is allowed and the model and measurement variances are reasonable. The variances need not be precise and could be determined in a separate analysis and brought into the process as long-term parameters.

• The resulting value is the best estimate of the value of the observed variable given the measured and modeled values and their respective uncertainties (noise).

Testing revealed that the Chi-Square prediction test resulted in outlier criteria that were too restrictive and lead to the over-discrimination of outliers. Although the Kalman Filter is very useful in control systems where the model is stable and well-conditioned, the multiplicity of underlying variables influencing stochastic behavior of water quality analytes appears to render it much less effective in the RTDF setting.

Like the sequential differences and serial correlation, the underlying stochastic process needs to be strongly dependent on the prior values. However, given reasonable estimates of model and measurement noise and a recursive implementation, its best role may be in quickly evaluating a new value without evaluating in detail a complete series. The model and measurement noise estimates need not be precise: the challenge is finding a reasonable way to estimate the imprecise values.

Discrete Cosine Transform (DCT)

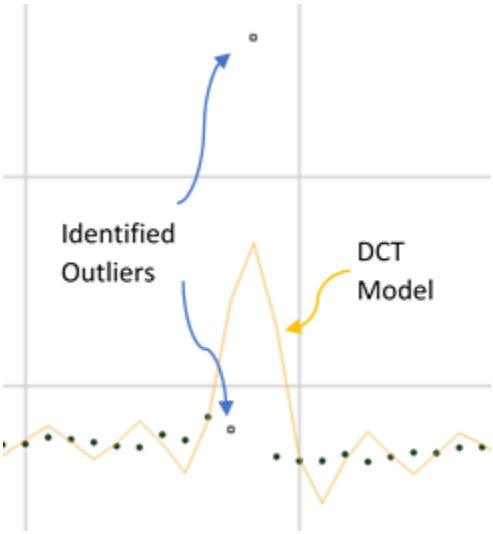


Figure 5 DCT Erroneous Outlier

The discrete cosine transform (DCT) is a variant of the general Fourier transform that employs the cosine function alone. It is widely used in digital signal processing for compression of audio and images where high frequency information can be filtered (4). It is used with a "low pass filter" to create an

approximate and smooth, continuous function through the data points as a model of the underlying data. Individual data points that deviate from the function by a threshold amount are considered outliers. Given the assumed filter, it applies to series with an average of less than an hour between observations. The assumed “low pass filter” ignores transform coefficients for frequencies greater than once per half-hour, and coefficients for frequencies between once per hour and two per hour (one per half-hour) are reduced proportionally.

The DCT first computes all the transform coefficients that optimally fit the data points. The number of coefficients is equal to the number of data points. The coefficients are ordered in the frequency domain in the order of increasing frequency. The first coefficient is essentially a constant since the frequency is constant (no variation or periodicity). The nth coefficient corresponds to a period of (2*T/n) with a frequency of n/(2*T) where T is the overall duration of the data series.

The filtered DCT model reliably identifies singular outliers in time series reporting more frequently than once per hour. It performs well at the boundaries of the time window. It does not require a stochastic process. However, distortions in the filtered model near an outlier can cause other points to be erroneously identified as outliers as shown nearby.

Since few of the MWQI sensors report frequently, the DCT filter has limited usefulness in quality control. However, many series at the California Data Exchange Center, CDEC, do report at regular intervals. The DCT can be very useful in filtering these series for graphic presentations.

Smoothed Periodic Medians

This filter constructs an approximate smooth line as a model. The smooth line intercepts periodic medians computed over the data series at regular periods, centered in each period. The computational period depends on the average point frequency of the data, the overall duration of the series divided by the number of points. The chosen computation period is reported in the results.

The nearest median values are added at the boundaries of the series of medians to guide the cubic spline approximation through those values.

Model values for each observed point are computed by interpolation in the local cubic spline function. The outlier threshold is computed from the distribution of errors: differences between the observed value and the interpolated value. The threshold computation is described elsewhere.

Since the cubic spline model is forced to the value of the nearest median at the boundaries, it is possible that erroneous outliers may be found at or near the boundary. This is less likely with a higher density of points.

Smoothed periodic medians do not require a stochastic series. This is a significant advantage with sensors that report less frequently but with values confined to a range, such as is typical of the tidally influenced Jones pumping plant.

As a model it is not precise, but it provides a reasonable, time-varying reference from which errors can be computed and analyzed for the presence of outliers.

Nearest Neighbors

A nearest-neighbor approach examines each point in the context of a selected number of nearby points. In the specific implementation for RTDF screening, a sample of points before and after the point of interest are examined and an average of differences between the point and its n nearest neighbors is computed to become the criterion for discriminating outliers. Normally, distances to n points before and after are computed and an average of the nearest n points is used for discrimination. For points near the beginning and end of a series, some of the points normally considered are obviously not available.

Nearest-neighbor approaches are relatively simple but effective and are widely used in practice (2).

Haar Wavelet

Wavelets in general are extensively used in signal processing (5). They are used in place of the Fourier transform because they usually require fewer computation resources and are said to be localized, i.e., transforms are specific to an identifiable subsets of data points that are independent of the transform of other disjoint subsets, facilitating the identification of anomalies, jumps, changes in statistics.

The Haar (rhymes with car) wavelet utilizes a simple transform and one of the first identified as a wavelet. It is very easy to understand and has proven effective in many applications (e.g., the JPEG 2000 compression algorithm). The transform is essentially based on recursive levels of differences between successive pairs of data values into an average and difference pair. The inverse is also recursive and simply splits the average values into pairs based on the corresponding differences.

The Haar wavelet has been used for outlier detection. A simple, understandable methodology (5) for using the wavelet transform to identify outliers was tested and found it to be effective. However, it offered no advantage beyond the simpler sequential differences analysis.

The Haar transform can also be used to identify jumps and periods in the series with different variances as described in several papers (6) (7). These approaches are being examined for possible consideration for future enhancements.

Alternatives Eliminated

Several potential discriminators were tested and found to be problematic.

Invariant Sequence

A series of values that are reasonably identical can indicate a sensor or reporting failure. A small relative change is ignored, and a minimal number of points must be found. However, testing revealed that a sequence of nearly equal, valid values can occur regardless of the relative change tolerance, with a surprising number of repeats. Period-of-record testing might indicate reasonable values for these two

parameters, but they might be sensor-related. Further exploration of this test was judged less effective than other alternatives.

Modified Z Score

The modified Z score uses the median of the data rather than the mean and the median absolute deviation (MAD) from the median rather than the mean and standard deviation. These statistics are much less sensitive to outliers than the mean and standard deviation.

The modified Z Score is the deviation from the median times a constant that is used to render a score that is roughly equivalent to a conventional Z score.

ModZScore(x) = C *((x-median))/MAD

MAD is the median absolution deviation of data values from the median.

The constant C is used to make the modified Z-score equivalent to a conventional z-score assuming a normal distribution. Different values are used in different references. A value of .6748 is commonly cited.

The main problem in applying this approach was in defining thresholds for outliers. The possible presence of a trend was also problematic. Further refinements could overcome these problems, but other alternatives provide equally effective approaches.

ScreenDataSQL Script

This script is no longer used. It was used from about 2011 through 2017. It was replaced by RTFilter.

- Parameters that control the screening process are stored in the database and may be tested and revised using a dedicated tool in the Data Management Utility.
- Data screening is performed by the script ScreenDataSQL.vbs. The script is normally applied within the data import processes.

- Manual data screening is performed using the Data Management Utility editing tool.

Purpose

The Real-Time Data and Forecasting Project (RTDF) relies extensively on water quality and hydrologic sensors deployed in many locations, that remotely measuring and report digital hydrologic and water quality data frequently, typically many times each day. The reports are collected and stored in a digital time history, or time series, of measurements. The recent, or "real time" dataset informs operational decisions, and the historic or archival record is useful for various scientific inquiries. It is probably inevitable that outliers are encountered in the automated data stream. Outliers are exceptional values that are inconsistent with most of the data in a dataset. They appear out of place in the context of the rest of the data.

Outliers arise from many sources, including problems in the sampling, measurement or reporting systems. Their existence renders reporting and analysis difficult and unreliable. Identifying and removing outliers is required for most purpose.

The need for timely dissemination of the most recent data effectively precludes the careful visual review of data quality. Even with adequate time, the manual review and removal of outliers is time-consuming, tedious, and invites human error. Automated processes are, therefore, highly desirable. This document describes the development of processes for the automated detection and removal of outliers in the RTDF project. Several alternative algorithms have been examined as candidates for deployment in the RTDF. These are described and compared, and recommendations are presented. The algorithms described are contained in a shared module that is deployed in various MWQI RTDF software components.

The automated data stream from MWQI sensors incorporates the algorithms to eliminate outliers so that they are not included in the data that is immediately disseminated. Outliers found in these processes are also marked in the RTDF database.

The same algorithms are integrated into other applications for reviewing MWQI data for quality assurance prior to exporting the data to the Water Data Library. They are also incorporated in processes that create graphic dissemination products, but these dissemination processes do not change the underlying data and may be less precise than quality control processes.

Background

- A heuristic, automated screening procedure has been in use in examining RTDF data since 2009. It is described below in the section titled Forward Moving Average. It includes a range check. This procedure has worked moderately well but has several significant limitations: parameters are required; parameter values rely on users for definition; parameter values they are typically constant over long periods of time during which the characteristics of the data may vary widely. State-of-the-art alternatives with parameters computed from the data are at least likely to be more appropriate.

Although the existing may have been reasonably adequate for real-time dissemination, where the value of timely reporting out-weighed the time and labor required for a higher level of quality, the desire that the data serve as a reliable archive required a higher level of quality assurance. Beginning in 2014, in response to user (client) urging, MWQI began publishing data collected as part of the RTDF program to the Water Data Library. With that effort came a concern that the quality of the data should receive a more rigorous review.

Initially, it was assumed that a classic assumption of a normal distribution for each day's data would be adequate: data exceeding three to six standard deviations from the mean would be considered outliers. With a little testing, it was quickly apparent that this approach would not account for numerous complications:

• Data values can vary during the day

• Outliers typically distort conventional means and standard deviations.

• Actual long-term data distributions are typically skewed above the mean.

• The computed minimum acceptable value is often less than zero

• Missing data required special handling

The challenges that surfaced from the initial effort led to extensive research, development and testing over an 18-month period. This memorandum describes the results of that project to date. Some of the lessons learned have substantially improved both the real-time, automated outlier detection capability as well as enhanced the tools available for interactive review of the data.

This is not necessarily the end of outlier discrimination research and development. A search for further improvements seems advisable, though given the efficacy presently available, further research at the same intensity may be less urgent.

Research

- The technology of outlier detection has developed substantially in the past 30-40 years and continues to develop. It is being driven by enhancements in software development, in new developments in the application of statistics and by the enormous increase in the quantity of data used in all sorts of enterprises. Current names for applications in data analysis include data mining, data science, machine learning and artificial intelligence.

  The significant recent focus on data science has necessarily included attention to the problem of outliers. Many papers, books and presentations on outlier detection can be found readily.

  Internet searches were the primary means of finding information on outlier detection. Wikipedia articles were helpful for basic concepts. Many papers found were from

academia and academic research. Those were often useful in documenting research into new applications rather than in identifying workable solutions.

The textbook “The Elements of Statistical Learning: Data Mining, Inference and Predictions” (1) was especially valuable in identifying and applying basic analytical tools and concepts.

Several sources (2) (3) presented comprehensive state-of-the-art summaries of outlier detection methods in wide use. These sources helped in identifying alternatives to try and in confirming whether this project was on track or not.

The research was not exhaustive and not all reviews resulted in an implementation to test.

- # Testing
- In the earliest phases of the project, efforts focused on examining the data from MWQI sensors in their entirety. The R statistics program was used for this purpose. R readily accepts and efficiently handles much larger datasets.

  The period of record data for each MWQI sensor was exported to a text file. This avoided the need to work out the details of connecting R to the database.

  Eventually, a VB.NET application was developed to compute medians and median absolute deviations (see Robust Statistics below) for successive ten-day periods and save the results in text files. These in turn could be filtered to datasets with and without outliers that have been identified over time.

  The analyses of the two datasets resulted in the following conclusions:

  1. The difference between the long-term mean and median are small.

  2. The long-term distribution of values for nearly all values is skewed above the mean. This was especially evident in histograms of the distributions.

  3. The differences between sequential values for most sensors falls within a narrow range, with a small variance. Histograms of the differences solidly confirmed the relationships.

  While R offered robust and diverse analysis tools, connecting it with data was laborious and comparatively inflexible. Some higher order packages were difficult to understand and didn’t allow customization. Moreover, the eventual outlier detection algorithms were going to be implemented in the standard RTDF platform: VB.NET. Therefore, a widely-used statistics library for the .NET framework (Math.NET, https://www.mathdotnet.com) was obtained for statistical computations. A new form was created for the Data Management Utility (DMU) that combined graphics, data retrieval and statistical analysis for testing outlier detection components. A separate analysis library of outlier detection algorithms was also created for eventual adoption in the production environment. The

DMU provided the basic interface with the RTDF database which contains selected CDEC data records as well as the MWQI sensor data. These changes enabled development of reasonably flexible algorithms for testing and analysis.

Histograms of series, both the primary series of observations and secondary computed series, were most essential in understanding robust probability distributions for discriminating outliers.

Outlier Detection R Workflow

RTDF sensor records were exported from the RTDF database into text files using SQL queries. Only the observation time and non-null raw values were included in the exports. The R subset command was used to further filter the data for basic evaluations as shown below.

The following broad filters are useful in analyzing the various records:

TOC 0-20

DOC 0-18

Bromide 0-1

Chloride 0-200

Nitrate 0-10

Sulfate 0-200

EC 0-1000

1. Get data from text file.

```
Sievers_TOC <- read.delim("C:/Projects/QAQC/POR Analyses/Hood/TOC/Sievers
TOC/Sievers_TOC.txt")
```

2. Filter data vales into a subset by min and max

```
x <- subset(Sievers_TOC$rawvalue, Sievers_TOC$rawvalue <=20 &
Sievers_TOC$rawvalue>=0)
```

3. Create a histogram with a nested filtering, i.e., without creating a separate filtered subset as in step 2 above

```
hist(subset(Sievers_TOC$rawvalue, Sievers_TOC$rawvalue <=20 &
```

```
Sievers_TOC$rawvalue>=0), breaks=50, main="Hood Sievers_TOC", xlab="TOC,
mg/L", freq=FALSE)
```

4. Display summary statistics of the nested filtered dataset.

```
summary(subset(Sievers_TOC, Sievers_TOC$rawvalue <=20 &
Sievers_TOC$rawvalue>=0))
```

```
                  ObsTime          rawvalue
 2002-04-05 14:06:00:     1   Min.   : 0.000
 2002-04-09 06:06:00:     1   1st Qu.: 1.620
 2002-04-09 07:06:00:     1   Median : 1.888
 2002-04-09 08:06:00:     1   Mean   : 2.240
 2002-04-13 00:00:00:     1   3rd Qu.: 2.435
 2002-04-13 00:06:00:     1   Max.   :20.000
 (Other)            :294029
```

5. Create a series of differences between successive values using a nested filter

```
y <- diff(subset(Sievers_TOC$rawvalue, Sievers_TOC$rawvalue <= 20 &
Sievers_TOC$rawvalue >= 0))
```

6. Create a histogram of the successive differences using a nested filter.

```
hist(subset(y,y<=.5 & y >= -.5),breaks=50,main="Hood Sievers TOC
Differences", xlab="TOC, mg/L", freq=FALSE)
```

Time Series

DMC_EC <- readSeries("C:/Projects/QAQC/Realtime Analyses/DMC_EC.txt",header=TRUE,format='"%Y-%m-%d %H:%M:%S", zone="PST")

Data Screening Tool

A special form can be used to test and edit screening data for MWQI sensors.

The form is activated by clicking the *Screening* button on the toolbar.

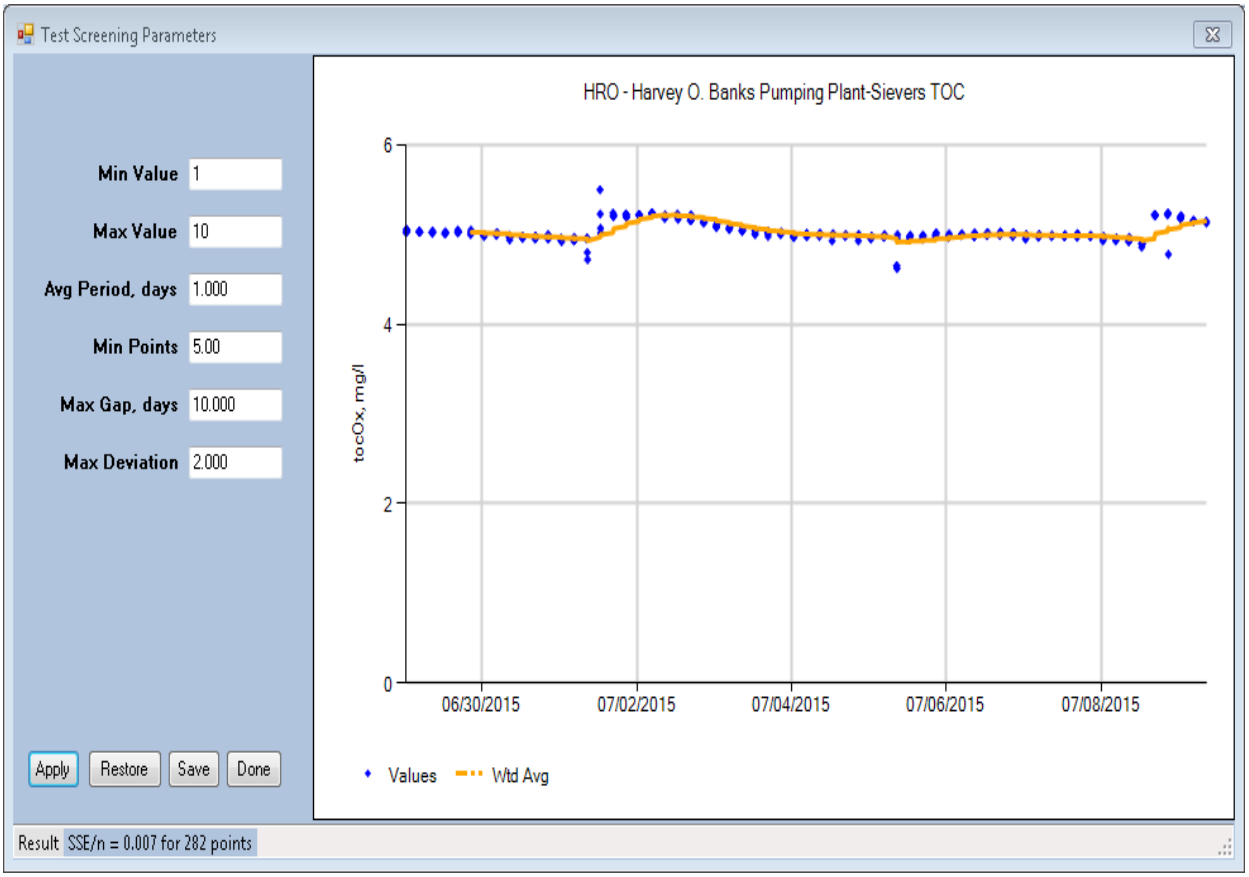
The form does not change series data.

The screening methods incorporated in the form are the minimum and maximum cutoffs and and the weighted forward moving average.

To test parameters, change the values and click on the *Apply* button. The plot will refresh to reflect the current parameter values.

Click the *Done* button when finished.

Use the *Save* if you want the automated screening process to use the current parameter values.



Automated Data Screening

Purpose

Data from field instruments operated by the MWQI, including organic carbon and anion concentrations and physical parameters such as temperature and electroconductivity are screened after they are imported to flag data values that are outliers. The flags are recorded in the RTDF database.

Overview

The database contains both a raw field (*RawValue*), a value field (*Value*) and a quality value (*Quality*) for each observation reported by the field sensor. *RawValue* is never changed. *Value* is initially equal to *RawValue* but may be changed through editing.

All raw values are assumed valid before screening. If a data value fails a screening test, *Value* is set to "Null", a database value which is equivalent to "no data" and *Quality* is set to Unacceptable. Null values are excluded from subsequent computations, such as daily averages and from data dissemination products.

RTFilter

Automated data screening is encapsulated in an application named "RTFilter."

RTFilter is incorporated in the regular cycle of field data imports. An input file (\\mrsbmapp20932\scripts\RTFilter.txt) directs its operations. It writes to a log file (\\mrsbmapp20932\scripts\RTFilter.log). These file assignments are made in the batch file that launches the application.

RTFilter employs a combination of outlier detection algorithms in sequence. Screening parameters are specified as described under "screening Parameters".

1. Values below the reporting limit for the sensor are flagged. The relevant reporting limit comes from the screening parameters and are summarized. If a point that is less than the reporting limit may later designated as an outlier, overriding the reporting limit flag.

2. Check standard values are flagged as Unacceptable. Check standard values apply only to Banks anions and are specified in application code.

3. A Min-Max screen is applied. The relevant Minimum and maximum values come from the screening parameters and are summarized. Values that are outside the min-max range are identified as outliers.

4. The autoregression lag 1 (serial correlation) method is first applied. Outliers identified previously are ignored. If the computed R^2 is greater than 0.85 and no outliers are found, no further analysis is performed.

5. If the serial correlation R^2 is less than 0.85 or the AR lag 1 analysis indicates outliers, a joint outlier detection analysis is applied. Both the k-Nearest Neighbors and Periodic Medians approaches are evaluated. Both methods must agree for an outlier to be designated. Values previously identified as outliers are ignored.

6. An outlier threshold must be greater than 5% of the median series value to eliminate the identification of outliers based on negligible differences.

ScreenDataSQL Script

This script is no longer used. It was used from about 2011 through 2017. It was replaced by RTFilter.

- Parameters that control the screening process are stored in the database and may be tested and revised using a dedicated tool in the Data Management Utility.
- Data screening is performed by the script ScreenDataSQL.vbs. The script is normally applied within the data import processes. See Server Scripts and batch Files.

- Manual data screening is performed using the Data Management Utility editing tool.