Validating data quality for water quality monitoring: Objective comparison of different data quality assessment approaches

Janelcy Alferes[1,4*], John Copp[2], Stefan R. Weijers[3], Guillaume Cussonneau[4], Gilles Fay[4], Abel Dembele[5] and Peter A. Vanrolleghem[1]

[1]modelEAU, Université Laval, Dép. de génie civil et de génie des eaux, Québec, QC G1V 0A6, Canada

[2]Primodal Inc., Hamilton, ON L8S 3A4, Canada

[3]Waterschap De Dommel, P.O. Box 10.001, 5280 DA Boxtel, The Netherlands

[4]CIRSEE, Suez, 38, rue du Président Wilson, 78230 Le Pecq, France

[5]Smart Solutions, Suez, 38, rue du Président Wilson, 78230 Le Pecq, France

*Corresponding author: *Janelcy.alferes@suez.com*

Abstract: Important advances have been made recently regarding several monitoring tasks and measurement applications within the global water system. However, besides the huge amount of real-time data being collected by such measurement set-ups, one of the most important steps forward have been made in the field of data quality evaluation. Challenging measurements conditions that characterise the monitored water environments make that data is frequently affected by different kinds of faults that degrade the quality of the data and compromise its use in further applications. Data validation becomes then critical to guarantee an efficient and reliable monitoring strategy. Nevertheless, the decision about when a value can be considered as “valid” or “not valid” is not straightforward and the assessment of the data validation approach itself should be evaluated, especially in a practical setting. In this paper, practical data validation developments from different teams are presented and evaluated using a common framework that allows an objective comparison of the methods and their usefulness for water quality time series validation. The proposed methodology will allow the unbiased comparison of data validation methods with the final objective of improving and applying suitable data validation tools for practice.

Keywords: Data quality assessment, fault detection, process monitoring, water quality

INTRODUCTION

Well-operated continuous water quality measurement systems have the ability to capture dynamics in water and wastewater systems, which allows for the identification of critical events, the evaluation of discharge impacts on receiving water bodies and the identification of cause and effect relationships. For many activities water authorities and water utilities are increasing the number of continuous measurements at different locations within their water and wastewater systems. As a main challenge, water quality sensors are subjected to different kinds of faults, leading to bias, drift, precision degradation or total failure, all causing a deterioration of the reliability of measurements (Branisavljevic et al., 2010; Garcia et al., 2014). Common errors observed in the increasingly long raw data series include missing values, NaN values, outliers, measurement values out of range or being constant, and varying noise.

To be able to effectively use the information that is in principle present in those continuous measurement data sets, the data should be accurate, verifiable, and properly validated. Given the size of the data sets produced nowadays in online water quality monitoring schemes, automated data validation is the only practicable option (Alferes et al., 2013). However, existing “standard” methods for data validation can only be implemented with difficulty in the water sector due to the properties of the water quality measurements and the water environments they are subjected to. As a result, common practice still today is visual inspection and the application of manual procedures.

In the framework of practical water quality monitoring applications different approaches for automatic data quality assessment developed by different teams are presented and compared. The approaches are each intended to detect doubtful and unreliable data by evaluating different features extracted from the data. However, the decision about when a value can be considered as “valid” or “not valid” is not simple. For complex time series, as in case of water quality parameters, an extra challenge is the risk of fault under-estimation and over-estimation. Several criteria such as variable characteristics, type of sensor, and sensor location should be considered. Expert knowledge about expected data variability and sources of faulty situations should be combined to set the methods’ parameters for each application.

Given the differences in methodologies and objectives of the different approaches it is also not evident to objectively evaluate the performance of the data validation processes itself in a practical setting. Thus, in this paper a methodology will be presented to evaluate and compare the usefulness or appropriateness of each approach based on data validation performance indices that are calculated from actually collected water quality time series. The final objective of this collaborative effort among the different teams is to come up with better tools and improved approaches for implementing successful automatic data quality validation procedures. All this will be illustrated with extensive data sets collected at different locations in water systems, including sewers, wastewater treatment plants (WWTP) and receiving waters.

COMPARISON OF DATA VALIDATION APPROACHES

Different expert teams (from the research, utility and consultant side) have developed actual implementations that are currently being used in measurement systems with data quality validation schemes. In general, different fault types can be detected at different complexity levels with those algorithms where evaluated data features include, among others: (i) Measurements outside physically acceptable limits, (ii) constant value or regular oscillations, (iii) constant difference between consecutive points, (iv) abnormal residual standard deviation between an estimated and an observed value, (v) difference between sequential points outside an acceptable tolerance, (vi) autocorrelation outside an acceptable tolerance, (vii) abnormal slope values (rate of change between two data points), (viii) high percentage of outliers or non-valid data.

Table 1. Results of three data validation approaches applied to four water quality time series

	Accepted data values (%)		
Time series	**Method 1**	**Method 2**	**Method 3**
Ammonia-WWTP	82	61	44
Flow-WWTP	86	95	94
Conductivity-WWTP	88	94	42
Level-Sewer	92	95	91

Table 1 summarizes some of the results of applying three different data validation approaches (detailed in the next subsection *Data validation methods*) to different water quality time series collected at a WWTP and a sewer location. Important differences concerning the percentage of accepted data were obtained for example for conductivity and ammonia time series. Such differences are due to different types and frequencies of faults, validations based on different criteria, no clear distinction between correct and faulty measurements, tuning of the methods and subjectivity in the interpretation and validation task.

The development of successful on-line data quality validation schemes would benefit greatly from an unbiased comparison of different data validation methods, evaluating their performance in terms of identifying real sensor/process faults. Taking as basis the objective assessment procedure for process monitoring strategies within the benchmark simulation platform BSM1_LT (Corominas et al., 2010), a systematic procedure is proposed to compare data validation approaches on the basis of real data sets and therefore real sensor faults. Working with real rather than simulated sensor data is the real challenge of the work. Indeed, in the work of Corominas et al. (2010) a set of given ('real') sensor faults is simulated and the fault sequence given by a validation method can be assessed in a simulation benchmarking context. Here, the assessment is to be done on actual data for which the faults are not known. To replace the real fault sequence, expert evaluation of the time series itself, supported by the results of the applied data validation methods, is used to create a surrogate fault sequence to which the different validation methods can then be compared.

For illustration, Figure 1 presents an example of the construction of the surrogate fault sequence for an online sensor. The output of each data validation method is called the estimated state of the sensor. For each data validation method a value of "1" represents normal behaviour and a value of "2" represents a fault. The final surrogate sensor status (last graph) is calculated after combining the results from each data validation method and the expert evaluation.

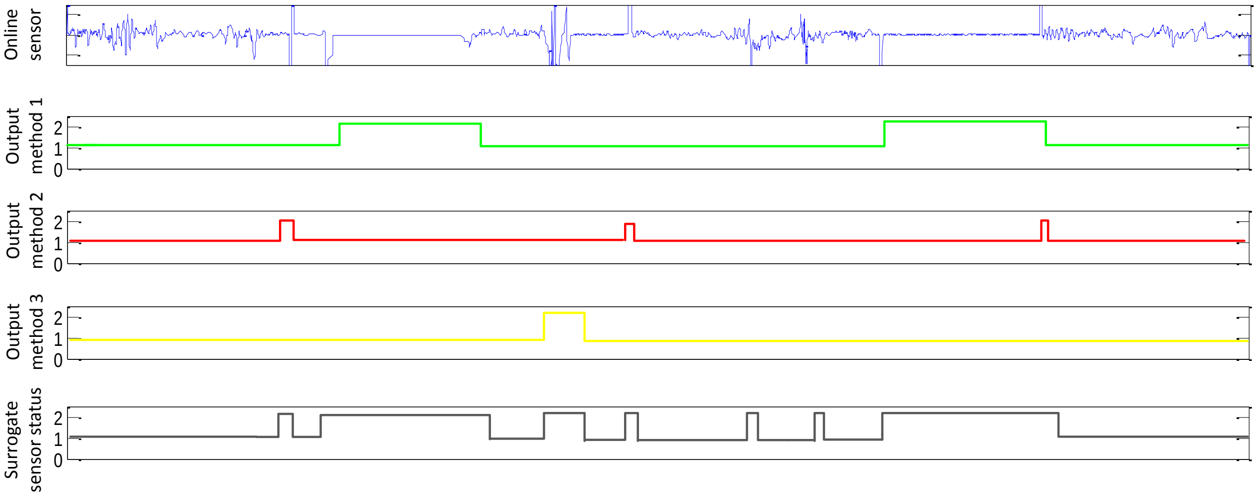

Figure 1. Example of the creation of the surrogate fault sequence for an online sensor

A global monitoring index is used for assessment of the data validation performance. Such index is calculated in such a way as to assign penalization points for inadequate performance of a data validation approach (providing an estimated fault status of the sensor) against the true (surrogate) fault status of the real sensor. Each data validation method is evaluated over the whole time series giving as output the classification of every sample point to be faulty or not. Based on Corominas et al. (2010) the global index penalizes for:

(1) The indication of a false acceptance: to evaluate the speed at which the fault is detected, as shown in Figure 2a, the more time needed to detect a fault event the more penalization points are given to the data validation method. To calculate the penalty function for false acceptance $P_{\mathrm{FAC}}(t)$ a timer is initialised and switched on at the beginning of a fault event and switched off when the fault event ends. $P_{\mathrm{FAC}}(t)$ is then calculated as follows:

$$P_{\mathrm{FAC}}(t) = P_{\mathrm{FAC,0}} + \left(P_{\mathrm{FAC,sat}} - P_{\mathrm{FAC,0}}\right).\left(1 - e^{\left(\frac{-k(t)}{\tau_{\mathrm{FAC}}}\right)}\right) \quad (1)$$

Where $\mathrm{k(t)}$ is the timer function, $\mathrm{P_{FAC,sat}}$ a saturation value for the penalty function and τ_{FAC} the time at which the penalty function reaches its maximum value (to account for the urgency in the detection).

(2) The intermittent detection of a fault: to evaluate the trustworthiness of the method, extra penalization points are given to indicate an intermittent detection of a fault within the duration of a longer fault. Based on equation (1) the time from the start of the fault is increased by $\mathrm{k_{switch}}$ whenever the alarm switches from correct detection to false acceptance:

$$P_{\mathrm{FAC}}(t) = P_{\mathrm{FAC,0}} + \left(P_{\mathrm{FAC,sat}} - P_{\mathrm{FAC,0}}\right).\left(1 - e^{\left(\frac{-\left(k(t)+\mathrm{k_{switch}}\right)}{\tau_{\mathrm{FAC}}}\right)}\right) \quad (2)$$

(3) The indication of a false alarm. In this case (Figure 2b) a constant penalty value $\mathrm{P_{FAL,0}}$ is given and the penalty function for false alarm $P_{\mathrm{FAL}}(t)$ is calculated as $P_{\mathrm{FAL}}(t) = P_{\mathrm{FAL,0}}$

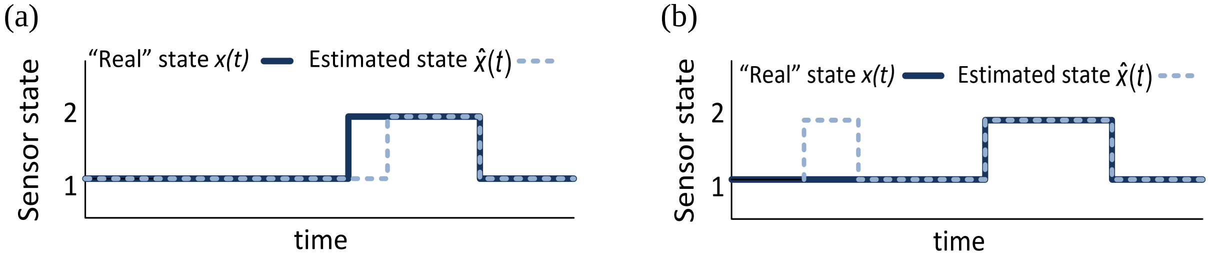

Figure 2. Penalization examples within the global index. (a) False acceptance, (b) False alarm.

As soon as the penalty functions are obtained, the penalty points are obtained by multiplying the difference $d(t)$ between the true state $x(t)$ and the estimated state $\hat{x}(t)$ with the penalty functions $P_{FAC}(t)$ and $P_{FAL}(t)$ to obtain the accumulated penalties A_{FAC} and A_{FAL}. The total penalization A is then obtained by adding such accumulated penalties as $A = A_{\mathrm{FAC}} + A_{\mathrm{FAL}}$.

With the objective of evaluating the reliability of each validation method, the maximum penalties for $\mathrm{P_{FAC}}$ and $\mathrm{P_{FAL}}$ ($\mathrm{A_{FAC,max}}$ and $\mathrm{A_{FAL,max}}$ respectively) are calculated by assuming $d(t)$ equal to 1 for all time instants. Then a measurement of the reliability R is calculated as:

$$R = \left(1 - \frac{A_{\mathrm{FAC}} + A_{\mathrm{FAL}}}{A_{\mathrm{FAC,max}} + A_{\mathrm{FAL,max}}}\right) \times 100 \qquad (3)$$

According to equation (3) a perfect detection case will receive 100% reliability and a totally wrong detection 0% reliability. Reliability in detecting false alarms and false acceptance individually can also be calculated. Additionally, individual indices can be calculated to evaluate the number of detections of true faults, the number of not detected faults and the number of false detections. The global and individual monitoring indexes are then calculated for each evaluated method by using the surrogate true sensor state and the estimated state provided by the evaluated data validation method.

Data validation methods

Method 1. Based on univariate time series analysis for outlier handling and fault detection (Alferes et al., 2013). Autoregressive models are used to forecast future expected time series data. Outliers are then identified by comparing the measured values with the forecast value with their dynamic prediction error interval. Once the outliers have been removed a smoothed time series is created and for fault detection purposes several statistical data features are calculated together with their acceptability limits. Figure 3 shows the results of applying the method to a conductivity time series collected at the inlet of the Eindhoven WWTP (the Netherlands). Blue and red lines represent the prediction interval within which normal data should fall and the green line represents the accepted smoothed data. Different kinds of faults are detected like a high percentage of outliers (period I), an excessive slope and residual standard deviation values (RSD) in period II and values with an abnormal constant value (period III). Once all data features have been assessed for each data point, data is validated according its degree of reliability (last subplot “Label”): 0 - valid (green colour, all data quality tests passed), 1 – doubtful (orange colour, some tests failed), and 2 – not valid (red colour).

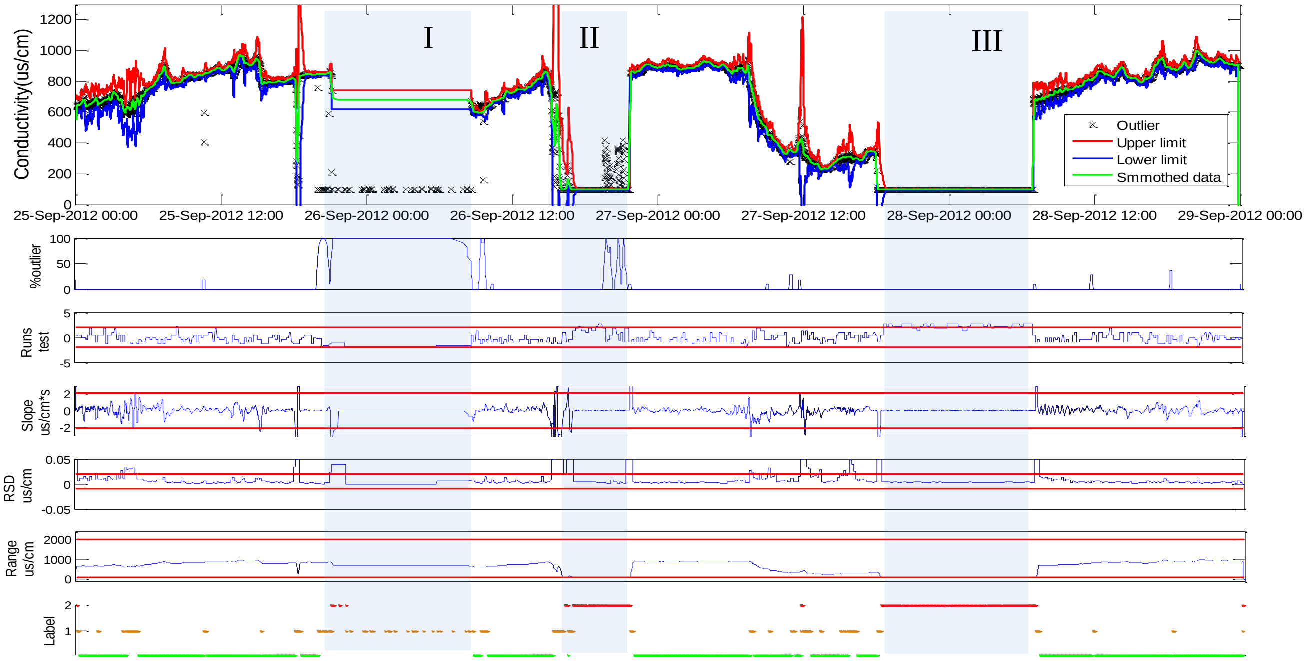

Figure 3. Application of a time series analysis method for data validation purposes

Method 2. As part of Primodal Systems’ RSM30 PrecisionNow software, a real-time step-wise analysis process (Copp et al., 2010) evaluates the existence of seven different fault types including:

(1) measurements outside acceptable limits, (2) difference between sequential points outside an acceptable tolerance, (3) constant value or regular oscillations, (4) constant difference between points, (5) standard deviation of the raw data below an acceptable tolerance, (6) residuals outside an acceptable tolerance and (7) raw data outside the dry weather tolerance (if flow data is available). Because of the real-time approach, the challenging part of the method is that it must detect not only faulty data in real-time, but it also incorporates criteria to detect that new data that is no longer faulty in real-time so that high quality data recording can resume after a period of time when faulty data has been rejected. Figure 4 shows an example of the application of the method (last subplot: 0 – valid data, 1 – invalid data) for a conductivity time series collected at the inlet of the Eindhoven WWTP. In this case invalid data resulted mostly from constant values, abnormal difference between sequential points and data outside acceptable standard deviation range.

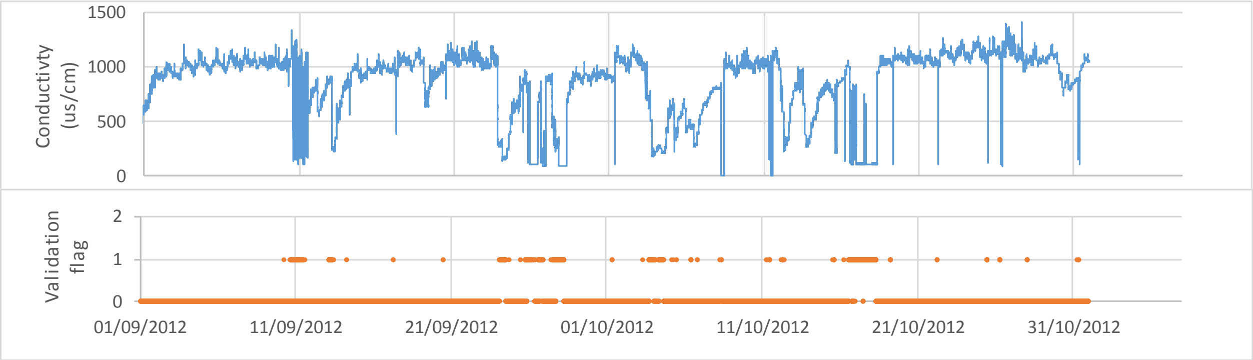

Figure 4. Fault flags (method 2) for a conductivity time series

Method 3. As part of the FEWS software (van Heeringen, 2015), the method evaluates different fault types providing quality flags to each single value accounting among others for missing data, values outside physical limits, values exceeding jumping limits and liveliness of the signal. Values can also be manually rejected. Figure 5 shows an example of the application of the method for time series of conductivity, temperature and dissolved oxygen. Validation results are shown with a color code at the bottom of each time series and different labels in the graph. In this case three periods where the sensor tubes were clogged were detected with measurements exceeding fixed jumping limits (Labeled ROR in the subplots)

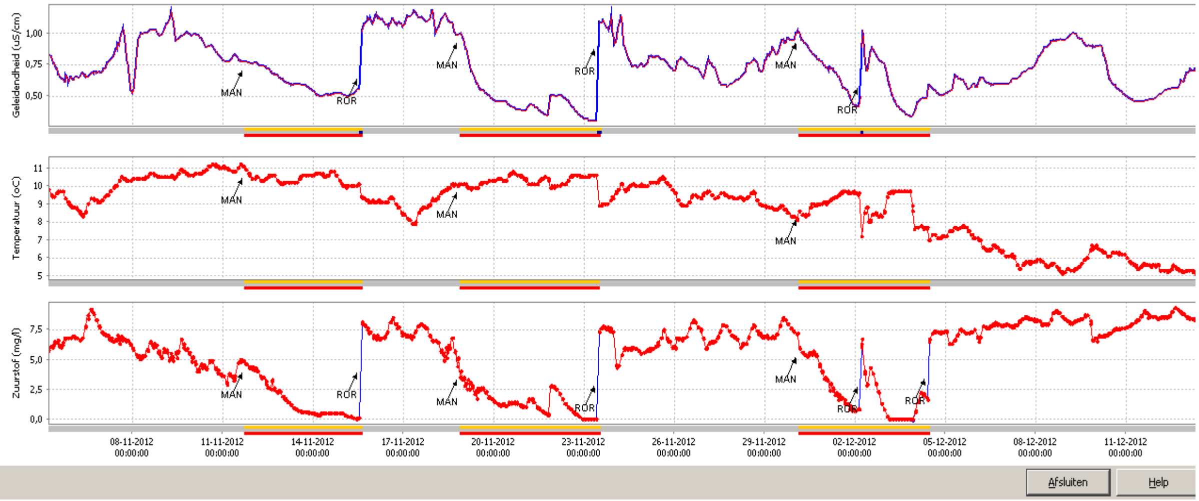

Figure 5. Quality flags (method 3) for conductivity, temperature and oxygen time series at the Dommel river

Method 4. The method (as part of internal preprocessing tools developed at CIRSEE**)** is focused on the evaluation of isolated peaks and low resolution data. A centered moving median is computed and then removed from the original signal at each time step. The absolute value of the residual is compared to the absolute value of the moving median at each corresponding time step. A coefficient is applied to the latter. If the residuals are higher than this value, they are considered as outlying i.e. having an isolated peak nature. To account for low resolution detection an indicator of the percentage of different values available among a sample of values is also calculated. This indicator is computed on a moving temporal window and then compared to a threshold. If the indicator is below the threshold, the value contained in the temporal window is considered as having low resolution. It is not necessarily to be removed, but there might be a sensor configuration problem. This also has the effect of removing the constant values after a certain amount of time steps. Figure 6 shows an example of the validation method (last subplot: 0 – valid data, 1 – invalid data) for a flow time series collected at the inlet of the Eindhoven WWTP.

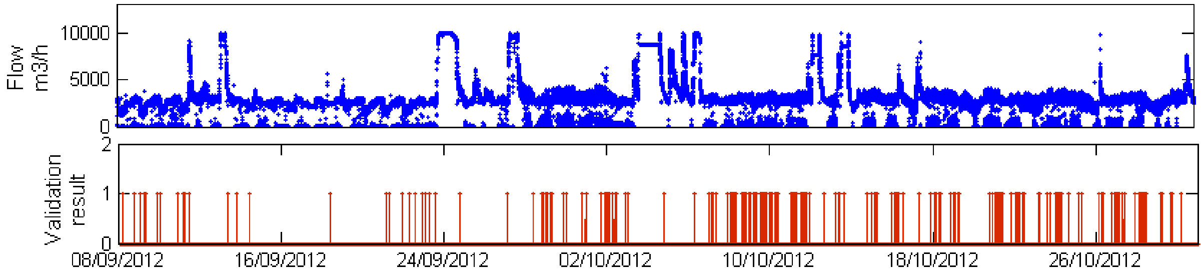

Figure 6. Validation output (method 4) for a flow time series

Method 5. As part of the Aquadvanced Suez tool the solution encompasses three different methods: (1) dedicated analysis for detecting pulse points based on the gradient threshold; (2) outlier detection based on the analysis of the absolute difference to the median absolute deviation (MAD) for a given interval, a point being considered as an outlier if its current value is greater than the median of the interval plus a certain fraction of the MAD; and (3) a dedicated method for detecting points indicating a beginning of change in the trend of the time series by analysing the absolute values of gradients. For all the methods, the threshold limits and coefficients are fixed by a learning process from historical data. Figure 7 shows an example of the validation method (subplots: 0 – valid data, 1 – invalid data) for the flow time series collected at the inlet of the Eindhoven WWTP. While the dedicated method (1) detects abnormal points due to unusual gradient values (typical during the on-off pumping dry weather strategy), the dedicated method 2 also detects outlier points associated to wet weather conditions.

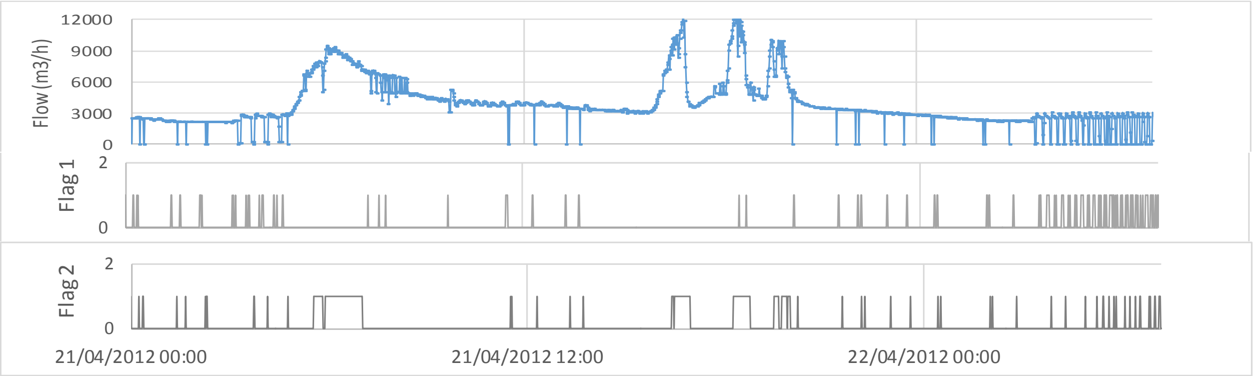

Figure 7. Validation outputs (method 5) for a flow time series

Ongoing work

The comparison methodology is being applied to long time series considered relevant for the partners in their future applications and collected at different locations (ammonia, flow and conductivity collected at the inlet of a WWTP; level measurements in the sewer system, level, velocity, dissolved oxygen and conductivity in the receiving water body). The application of the methodology and some relevant results will be further discussed in the paper presentation. It is important to clarify that methods here evaluated are focused on univariate analysis. Methods considering cross validation through several signals will provide additional information about the nature of the fault and to discern between a change in the sensors' properties or in the process variable itself.

ACKNOWLEDGEMENTS

Peter A. Vanrolleghem holds the Canada Research Chair in Water Quality Modelling.

REFERENCES

Alferes J., Tik S., Copp J. and Vanrolleghem P.A. (2013). Advanced monitoring of water systems using in situ measurement stations: Data validation and fault detection. Water Sci. Technol., 68(5), 1022-1030.

Branisavljevic N., D. Prodanovic and D. Pavlovic. (2010). Automatic, semi-automatic and manual validation of urban drainage data. Water Sci. Technol., 62(5), 1013-1021, 2010.

Copp J., Belia E., Hübner C., Thron M., Vanrolleghem P.A. and Rieger L. (2010) Towards the automation of water quality monitoring networks. In: Proceedings 6th IEEE Conference on Automation Science and Engineering (CASE 2010). Toronto, Ontario, Canada, August 21-24, 2010.

Corominas Ll., Villez K., Aguado D., Rieger L., Rosén C. and Vanrolleghem P.A. (2010) Performance evaluation of fault detection methods for wastewater treatment processes. Biotechnol. Bioeng., 108, 333-344.

Garcia D., Quebedo J. Puig V., Cuguero M.A. (2014). Sensor data validation and reconstruction in water networks: a methodology and software implementation. In Proceedings 9th International Conference on Critical Information, Infrastructure and Security (CRITIS). Limassol, Cyprus, October 13-15.

van Heeringen, K. (2015). BOS Brabant, gebruikershandleiding. Deltares. Rapport 1208557-000-ZWS-0016 (In Dutch).