Advanced monitoring of water systems using *in situ* measurement stations: data validation and fault detection

Janelcy Alferes, Sovanna Tik, John Copp and Peter A. Vanrolleghem

ABSTRACT

In situ continuous monitoring at high frequency is used to collect water quality information about water bodies. However, it is crucial that the collected data be evaluated and validated for the appropriate interpretation of the data so as to ensure that the monitoring programme is effective. Software tools for data quality assessment with a practical orientation are proposed. As water quality data often contain redundant information, multivariate methods can be used to detect correlations, pertinent information among variables and to identify multiple sensor faults. While principal component analysis can be used to reduce the dimensionality of the original variable data set, monitoring of some statistical metrics and their violation of confidence limits can be used to detect faulty or abnormal data and can help the user apply corrective action(s). The developed algorithms are illustrated with automated monitoring systems installed in an urban river and at the inlet of a wastewater treatment plant.

Key words | data quality assessment, fault detection, on-line monitoring, water quality

Janelcy Alferes (corresponding author)
Sovanna Tik
Peter A. Vanrolleghem
model*EAU*,
Université Laval,
Département de génie civil et de génie des eaux,
Québec,
QC G1V 0A6,
Canada
E-mail: *janelcy.alferes@gci.ulaval.ca*

John Copp
Primodal Inc.,
Hamilton,
ON L8S 3A4,
Canada

INTRODUCTION

A new generation of *in situ* automatic water quality monitoring stations is proposed adhering to the mon*EAU* vision (Rieger & Vanrolleghem 2008). With flexibility and standardization as the main drivers of recent developments, important advances have been made regarding several monitoring tasks and measurement applications (water bodies, wastewater treatment plants (WWTPs), etc.) (Winkler *et al.* 2002). However, besides the huge amount of real-time data collected in these types of implementations, the most important steps forward have been made in the field of advanced data quality evaluation. As measurements are carried out under challenging conditions (clogging, fouling, electrical interferences, flooding, etc.) raw data are frequently affected by faults like drift, bias, precision degradation or even complete failure, all of which cause the accuracy and reliability of the data to decrease (Yoo *et al.* 2008). Those conditions may lead to erroneous conclusions and to the improper use of the data (Bertrand-Krajewski *et al.* 2003). For data analysis and further applications, the collected data will be valuable only if the data are properly validated. Given the size of the data sets, automated data validation is the only viable option.

In the last few years in different fields a number of methods have been developed for fault detection and isolation (FDI) (Venkatasubramanian *et al.* 2003; He *et al.* 2005). Traditional model-based approaches make use of the generation of residuals (i.e. the difference between a measured value and its prediction by a model) and their evaluation for FDI. However, it is often difficult to identify and validate an accurate model that describes all physical and chemical phenomena occurring in the process. As an alternative, data-driven methods consider the relationships between the process variables without the explicit expression of a process model (Qin 2009). Despite these developments, methods for data validation and fault recognition used today in water systems usually follow inefficient procedures based on time series charts with a lack of systematic analysis (Branisavljevic *et al.* 2010; Mourad & Bertrand-Krajewski 2002).

In the framework of practical monitoring applications an important challenge is to develop automated data evaluation tools that can detect and correct erroneous data and assist in processing the data. This paper deals with these different issues. Data quality assessment tools that have a practical orientation and are based on multivariate analyses

doi: 10.2166/wst.2013.302

are proposed for faulty data detection. The tools have been successfully tested on water quality time series obtained from *in situ* automatic monitoring stations installed in two different water systems. According to the mon*EAU* vision, the final objective is to achieve advanced monitoring with efficient and automatic data collection, evaluation, correction and alarm triggering to create a long-term database with validated and valuable 'good' water quality data that can be used, for example, for decision support and control of water systems.

MATERIALS AND METHODS

In situ monitoring stations

Primodal Systems Inc.'s RSM30 monitoring stations were used to automatically collect *in situ* real-time water quality data. In the first application (Figure 1), a monitoring station was installed at the inlet to the primary clarifier of the municipal WWTP of Québec City, Canada. The measurement station included multiple pH, conductivity, temperature and turbidity sensors to determine if redundant signals would improve the short and long-term accuracy of the data and the detection of abnormal situations. The data for this study were collected in the spring of 2012. In the second application (Figure 2), a monitoring station was installed in a small urban river (Notre Dame) in Québec, Canada. The measurement station included several on-line sensors for collecting a large number of conventional physico-chemical parameters (temperature, dissolved oxygen, conductivity, turbidity, etc.), a UV spectrometer (nitrates, total organic carbon (TOC), dissolved organic carbon (DOC), turbidity) and ion selective electrodes (potassium, ammonia). In this case, data from the summer of 2012 are used.

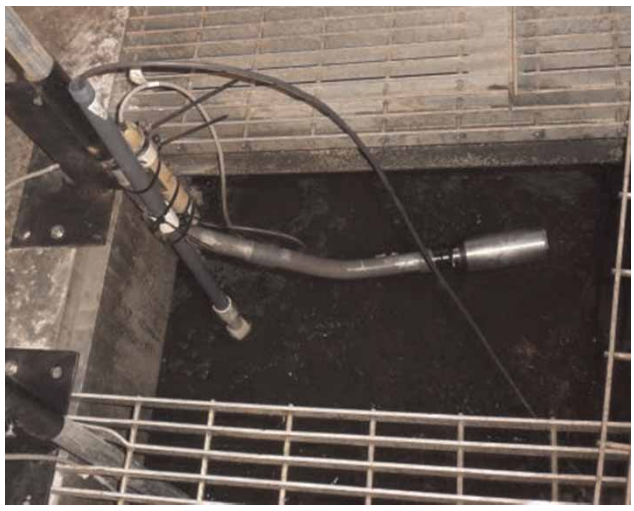

Figure 1 | Installation of sensors at the WWTP.

Figure 2 | Installation of sensors at the river.

To properly describe the dynamics of both water systems all sensors were set to record data at short intervals (5–60 seconds). This implementation generated information-rich but also complex and huge data sets. To increase the likelihood of good quality data from the on-line measurements, the application of a maintenance protocol including cleaning and systematic calibration tasks was essential (Poirier 2013).

Faulty data assessment

Ensuring the data quality from on-line measurements is one of the most important issues concerning effective monitoring today. In hostile environments like wastewater systems, sensors are subjected to failures that compromise the precision and the reliability of the measurements (Rosén *et al.* 2008; Yoo *et al.* 2008), which may result in incorrect perceptions of the monitored system and/or in erroneous control actions. Typical faults in online sensors are shown in Figure 3. The detection and diagnosis of these kinds of sensor faults are crucial if the water system is to be successfully monitored. Even if most researchers and practitioners agree with this statement, the reality is that little attention has been given to the study of sensors in a realistic manner (Rosén *et al.* 2008).

The tools for faulty data assessment proposed in this paper are based on multivariate methods. The multivariate

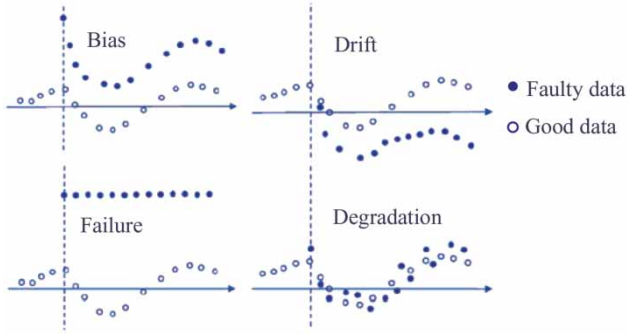

Figure 3 | Common sensor faults (Yoo *et al*. 2008).

process monitoring methods based on principal component analysis (PCA) and partial least squares (PLS) models have been shown before to be practical approaches for fault detection and diagnosis (Villez *et al*. 2008). These methods exploit the redundant information present in highly correlated variables (typical for real water quality data) to reduce the dimensionality. By exploring the original data set, PCA is used to find a new set of uncorrelated variables, called principal components (PCs), which explain most of the data variability in a more visual coordinate system with fewer dimensions. Given an autoscaled $[m \times n]$ data matrix X for n process variables and m samples, performing PCA allows decomposing X as follows:

$$X = \hat{X} + E = TP^T + T_e P_e^T = \sum_{i=1}^{a} t_i p_i^T + \sum_{i=a+1}^{n} t_i p_i^T \quad (1)$$

where $\hat{X}$ is the model matrix which describes the system variations and E is the residual or error matrix which captures the noise or unmodelled variations. The matrix P $[n \times a]$ is the loading matrix and its column vectors (p_i) are called loadings or PCs of X. The matrix T $[m \times a]$ is the score matrix and its column vectors (t_i), called scores, represent the values of the original data in the new coordinate system. Finally, a represents the number of PCs to be retained in the model. The matrix P can be obtained by performing a singular value decomposition (SVD) on the covariance matrix Cx of X that can be written as $Cx = R\Lambda R^T$, Λ being the diagonal matrix of the eigenvalues of Cx sorted in decreasing order $(\lambda_1 > \lambda_2 > \ldots > \lambda_n)$ and R the eigenvectors of Cx. As the λ_i values are a measure of the variance of X along each PC p_i, the reduced dimension matrix P is obtained by choosing the a eigenvectors of Cx associated with the a largest eigenvalues capturing the largest fraction of the data variance. The P_e matrix is generated with the remaining n–a eigenvectors. The goodness of the model depends on the right choice of a and should consider both the dimensionality reduction and the loss of data information. In this case, the method based on the eigenvalue scree plot (Jolliffe 2002) was used. Once the PCA model is obtained new data X_{new} can be projected onto the existing model while preserving the matrix P. New scores are calculated as $T = X_{new}P$.

Using the transformed data, sensor faults can be detected by measuring variations from the normal conditions both in the model and in the residual space. The 'normal' conditions are defined by the choice of the data set with which the PCA model is built. For fault detection, two statistical metrics are calculated and their violations of confidence limits are monitored. In contrast to univariate tests, the monitoring of these statistics takes into account the correlation in the data. A measure of the variation within the PCA model is obtained at time k by the T^2 statistic which is defined as the sum of normalized squared scores:

$$T^2(k) = x^T(k) P \Lambda_a^{-1} P^T x(k) \quad (2)$$

where x is the sample vector and Λ_a^{-1} is the diagonal matrix containing the a eigenvalues associated with the a eigenvectors or PCs retained in the model. Statistical confidence limits T_α^2 for T^2 are obtained by using the α-percentile Fisher distribution $F_{a,m-a,\alpha}$ with $(a,\ m-a)$ degrees of freedom and a level of significance α (usually between 0.01 and 0.05) (Yoo *et al*. 2008). A measure of the variation outside the PCA model space (residual space) is obtained at time k by the Q statistic which is defined as the sum of squared residuals:

$$Q(k) = x^T(k)\left(I - PP^T\right)x(k) \quad (3)$$

The Q statistic not only detects events that are not taken into account in the model space but also indicates the lack of model fit for each sample. An upper control limit Q_α for Q can be obtained assuming that x follows a normal distribution (Montgomery 2009). The process is therefore considered normal if $T^2 < T_\alpha^2$ and $Q < Q_\alpha$. An increase in T^2 can be interpreted as an abnormal increase in the main normal source of variance of the model, whereas an increase in Q can be seen as the introduction of an additional source of variance that breaks the normal correlation between the variables (Perera *et al*. 2006). A geometric interpretation of Q and T^2 is shown in Figure 4. The T^2 statistic defines an ellipse on the model plane defined by the PCs within which the operating points normally

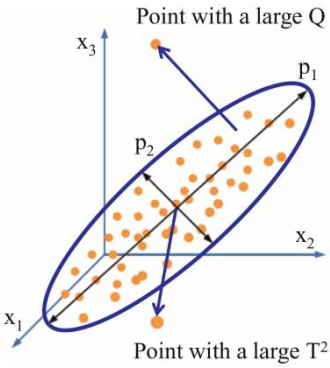

Figure 4 | Geometrical interpretation of T^2 and Q statistics (Montgomery 2009).

project. While the Q statistic measures the orthogonal distance from the sample to the model plane, the T^2 statistic is a measure of the distance from the sample to the intersection of the PCs.

The proposed sensor validation procedure is shown in Figure 5. The first step includes the development of the PCA model using a set of training data. In order to obtain a representative and valid PCA model, data are pre-treated to remove outliers and perform auto-scaling (mean centring and variance scaling). Outlier detection is carried out by using univariate autoregressive models which compare measured values with forecast values. Details of this procedure can be found in Alferes *et al.* (2012). Pre-treated data are then used to build the PCA model and to determine the confidence limits for the T^2 and Q statistics.

The second step involves auto-scaling of the new data and the projection of these new data onto the reference PCA space. If one or several variables are found to deviate from the normal model region (expected variability) the T^2 and/or Q statistics will increase above their normal values. Faults or abnormalities in the data are thus detected by comparing the T^2 and Q values against their thresholds. After a fault is detected, an alarm is generated and further analysis is carried out to identify the fault. This identification will then lead to the application of the necessary corrective actions in the field to eliminate or reduce the abnormal condition.

RESULTS AND DISCUSSION

To illustrate the potential of the proposed procedure the figures below show some of the results obtained from the time series of the first application with redundant sensors. While the difference between two redundant sensor signals can already be used for outlier identification, multivariate methods allow more analysis including the identification of multiple sensor faults and the detection of abnormal trends.

WWTP case

Time series from eight on-line variables at the inlet of the WWTP (Figure 6) were considered including: Conductivity sensor 1 (Cond1), Temperature at Conductivity sensor 1 (CondTemp1), Conductivity sensor 2 (Cond2), Temperature at Conductivity sensor 2 (CondTemp2), pH sensor 1 (pH1), pH sensor 2 (pH2), Turbidity sensor 1 (Turb1) and Turbidity sensor 2 (Turb2). All sensors recorded data at 5-second intervals. A representative training data set over a 3-day period was used to build the PCA model. Prior to the PCA modelling, training data were properly auto-scaled (mean centring and variance scaling) and outliers were removed. Performing the PCA showed that the first three PCs explain more than 90% of the total variance of the process. Therefore, three PCs were retained in the PCA model for further analysis. After calculation of the Q statistic and its threshold,

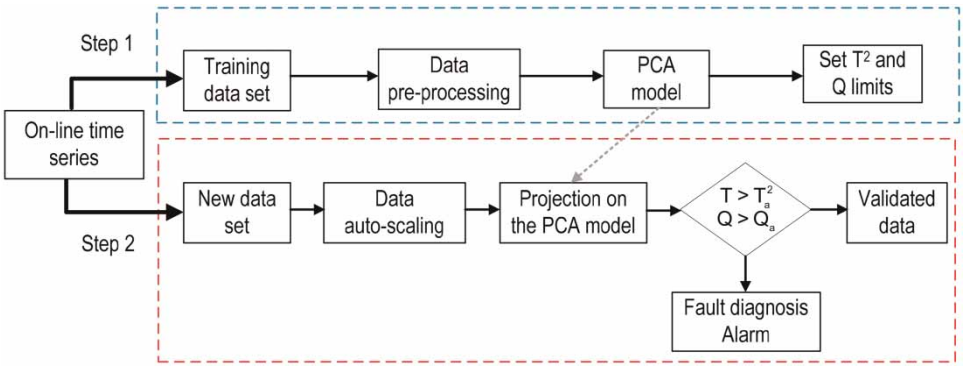

Figure 5 | Proposed sensor validation procedure.

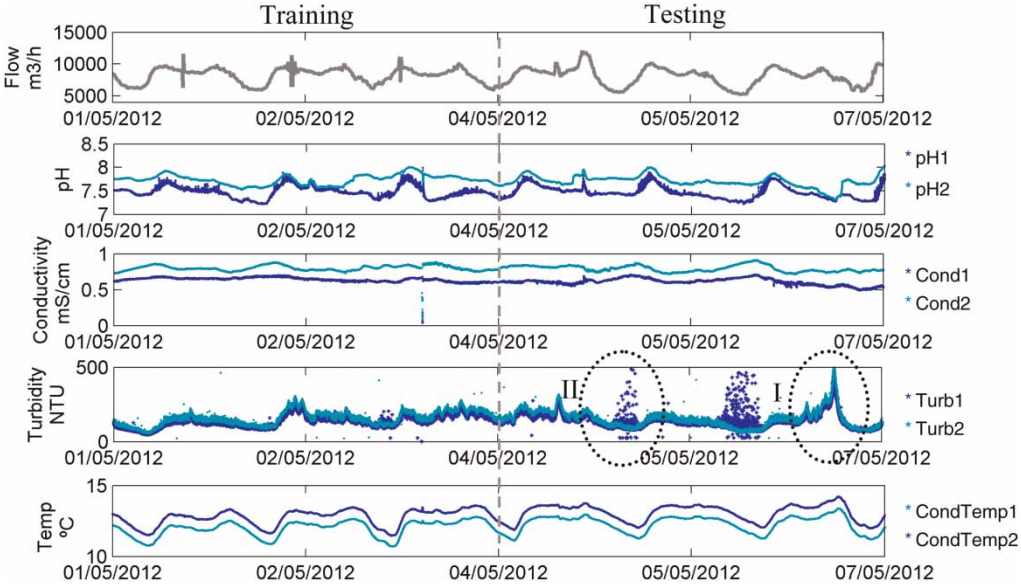

Figure 6 | On-line measurements of flow, turbidity, conductivity, pH and temperature at the inlet of the primary clarifier of the WWTP of Quebec City.

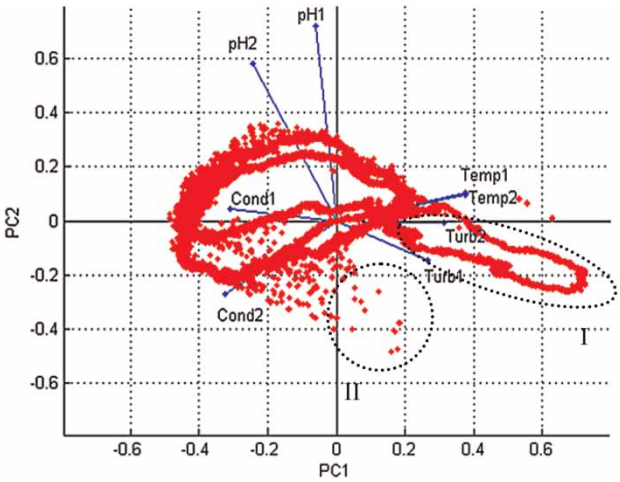

Figure 7 | PCA representation, WWTP application (1/5/2012–7/5/2012).

less than 1% of the samples were determined to be abnormal, demonstrating that the model was able to capture the main correlations and variability among the process variables. The time series of these variables (Figure 6) shows how the two conductivity signals describe similar dynamics but in the presence of a time variable bias of about 20%. Some divergence is also shown for pH and turbidity sensors although in the latter case the bias is less significant. All temperature signals present the same behaviour with a constant 5% bias for CondTemp1. These differences were assumed to be caused by missed calibration steps and the different ages of the sensors.

Figure 7 shows the scores of the testing data set once the reference PCA model is applied. Each variable is represented in the PC-space by a vector and its length and direction indicate the contribution of the variable to the two first PCs (PC1, PC2) for each observation. Each point in the plot corresponds to a measurement. Points that cluster represent similar behaviour and points that deviate pertain to process changes. It can be seen from this analysis how the vectors for the redundant temperature sensors have the same contribution to PC1 and PC2 suggesting a strong correlation between the two sensors. As expected, vectors for redundant pH and conductivity sensors indicate a considerable divergence, accounting for the bias presented between these sensors.

When considering the variation of the data in the PC-space, an analysis of the scores allows the identification of different clusters and outlying points. For example, a cluster in area I (in the direction of the turbidity measurements) reveals changes in these variables. In fact, these samples are associated with an unusual discharge on 7th May (see Figure 6) which induced an important variation in turbidity. Some outlying points are also identified around area II in the direction of the turbidity sensors suggesting an abnormal behaviour or disturbance for these samples.

Monitoring of the T^2 and Q statistics (Figure 8) allowed for the detection and isolation of some fault situations in the process. While T^2 accounts for data variability, Q measures the goodness-of-fit of each sample to the PCA model and is directly associated with the noise level. Figure 8(a) shows how, for samples in period I, the Q statistic is maintained

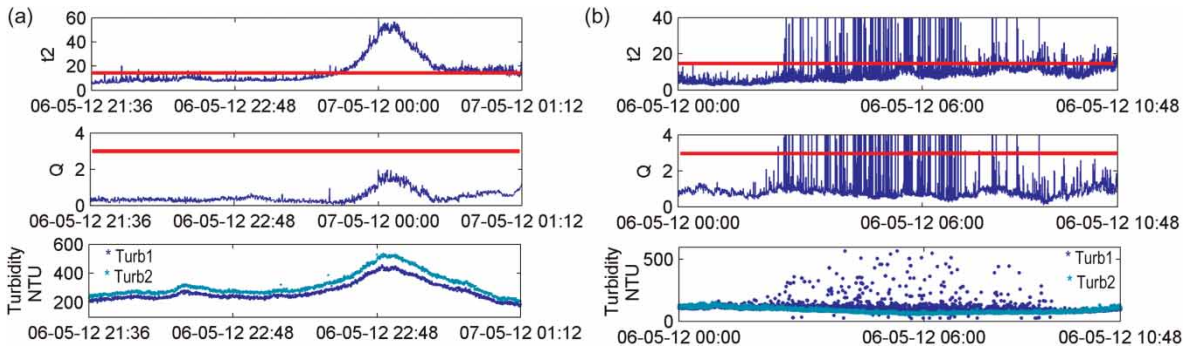

Figure 8 | T^2 and Q statistics for data corresponding to certain periods of Figure 6. (a) Period I, (b) period II. Horizontal lines are limits that allow detecting faults in the monitored data series.

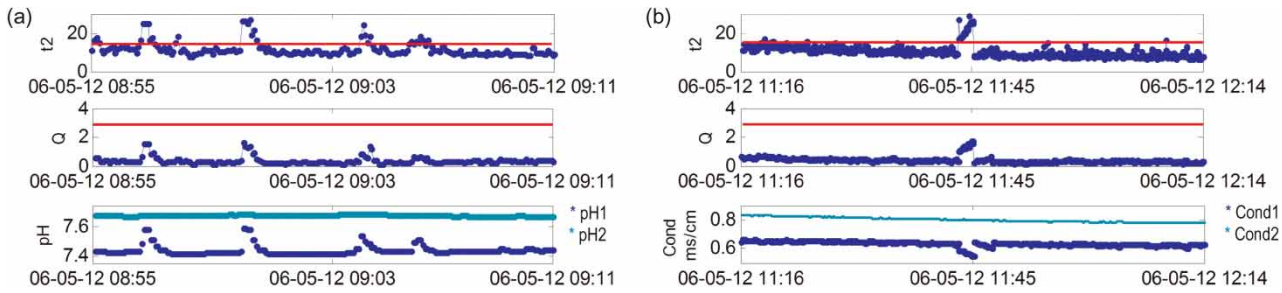

Figure 9 | T^2 and Q statistics for individual faults during certain periods of Figure 6. (a) pH, (b) conductivity. Horizontal lines are limits that allow detecting faults.

inside the limit while T^2 detects variations in the data that are larger than the variations expected under normal operation. The turbidity data indeed show a variation that is indicated by the high T^2 statistic. In period II, T^2 reveals important variations in the measurements (Figure 8(b)), but Q also reveals important noise. With reference to the location of the faulty data in Figure 7, the turbidity data were scrutinized and it turned out that the Turb1 data were a probable cause of the fault detection.

Similar conclusions could be drawn from the pH and conductivity measurements (Figure 9). For instance, some faulty situations are identified by the T^2 statistic around 6th May (end of period II). Time series for the process variables revealed some abnormal behaviour in the pH and conductivity measurements by sensors pH2 (Figure 9(a)) and Cond2 (Figure 9(b)), respectively. Although in both cases the Q statistics remained inside the limit, the T^2 statistics detected some drifts that changed the normal trend of those variables. The underlying causes could not be diagnosed.

Multivariate methods can also be applied to investigate correlations between redundant sensors. Figure 10 shows for example the representation in the resulting PC-space of the turbidity testing data set when only those sensors are

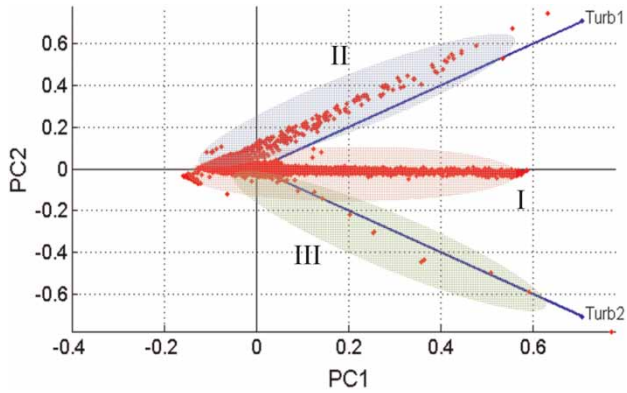

Figure 10 | PCA representation of the redundant turbidity data of Figure 11.

considered. While PC1 accounts for the dominant variability in the turbidity data set, PC2 accounted for the remaining variance between the turbidity sensors. Three clusters are identified: (1) area I, in which both sensors have the same behaviour; (2) area II, in the direction of the sensor Turb1 suggesting an abnormal behaviour for these samples; and (3) area III, in the direction of the sensor Turb2 accounting for disturbances in these samples. Traditional analysis of redundant signals through calculation of their differences

and standard deviation (StD) allow detecting outliers and some drift situations. However, multivariate methods help in the detection of abnormal trends and identifying which sensor is misbehaving. As shown in Figure 11, while monitoring the Q statistic mainly detected divergences between the two sensors (areas II and III), the T^2 statistics revealed also changes in the operating conditions not taken into account in the StD analysis shown in Figure 11 (area I).

River case

The following figures show some results obtained from the time series of the second application in which a small urban river was monitored. In this case, the data set included time series for turbidity, total organic carbon (TOCeq), pH, temperature, nitrates (NO_3) and potassium (K^+). Figure 12 shows some of the collected data. Performing PCA over a representative training data set showed that three PCs can explain more than 87% of the total variance of the process. Figure 13 shows the scores from the 12-day testing data set. Some clusters and outlying points are identified, for example in the marked areas (I, II) in the direction of turbidity and potassium measurements, respectively. Graphical representation of the T^2 and Q statistics time series in Figure 12 also allows the identification of different faulty or abnormal events.

Samples around period I are associated with an unusual event on 18th September (rain event), which mainly caused an important variation in the turbidity and TOC behaviour. In that period an abnormal condition was detected by the Q and T^2 statistics for a short time. However, T^2 remained longer outside the limits revealing still larger variations in the data than expected in normal conditions. On the other hand, around period II both statistics confirm an abnormal behaviour around 25th September. In this case, not only T^2 showed abnormal variations in the measurements but also Q identified events not taken into account in the model, clearly indicating a faulty condition that mainly affected the potassium measurements.

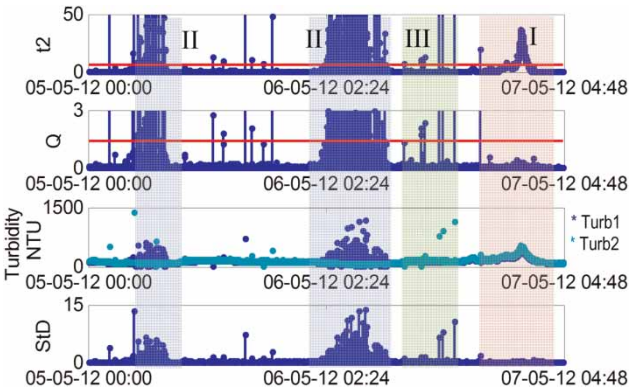

Figure 11 | Time series for T^2, Q, turbidity and StD for redundant turbidity sensors.

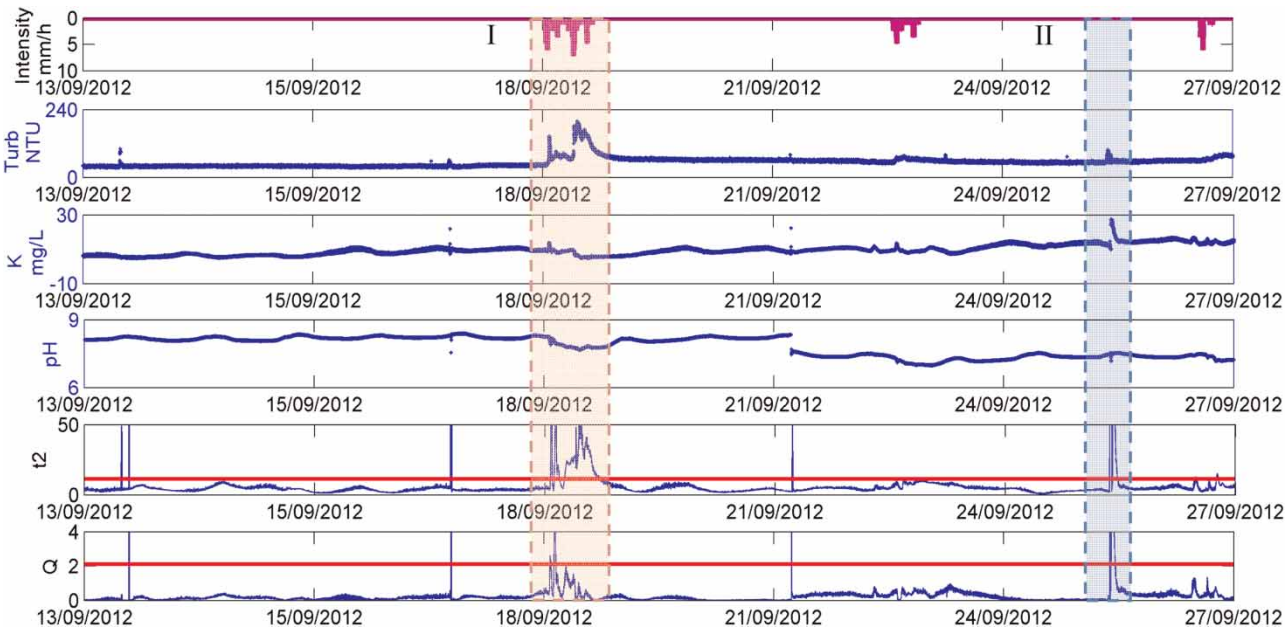

Figure 12 | Data and statistics time series for the second application in an urban river.

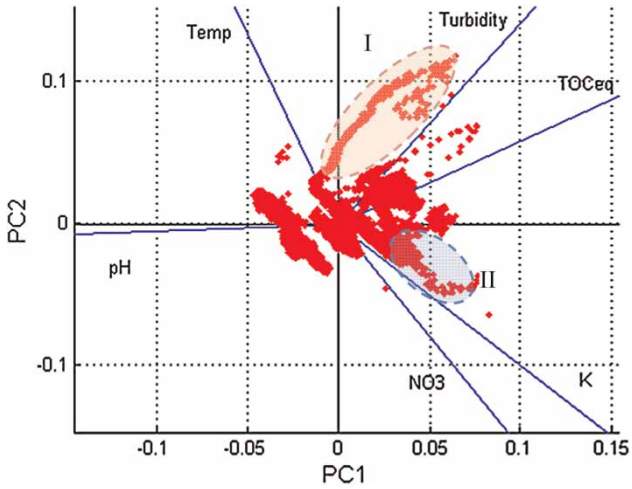

Figure 13 | PCA representation, urban river case.

CONCLUSIONS

As water quality measurements might be carried out in difficult environments, dealing with faulty sensors represents a challenge for the reliable real-time continuous monitoring of water systems. To address that challenge multivariate methods based on PCA have been tested on data sets obtained from *in situ* automatic monitoring stations storing several physical and chemical variables. After training the PCA model with normal operating data, faults or abnormal conditions can be detected by monitoring some statistical metrics and their violation of confidence limits. It was shown in two case studies that this procedure enables the detection of different kinds of faults in individual sensors. These can be used to trigger process and/or maintenance alarms. Once faults are detected and correctly diagnosed corrective actions can be applied to the measurement system. The availability and practical application of these methods to multiple and redundant water quality sensors represents a further step towards effective data quality assessment and better monitoring of water systems.

ACKNOWLEDGEMENTS

Peter Vanrolleghem holds the Canada Research Chair in Water Quality Modelling. The CFI Canada Research Chairs Infrastructure Fund project (202441) provided the monitoring stations. The authors want to thank the City of Quebec for its technical support. Peter Vanrolleghem was Otto Monsted Guest Professor at the Technical University of Denmark in 2012–2013.

REFERENCES

Alferes, J., Poirier, P. & Vanrolleghem, P. A. 2012 Efficient data quality evaluation in automated water quality measurement stations. In: *Proceedings International Congress on Environmental Modelling and Software (iEMSs2012)*, Leipzig, Germany, July 1–5, 2012.

Bertrand-Krajewski, J. L., Bardin, J. P., Mourand, M. & Beranger, Y. 2003 Accounting for sensor calibration, data validation, measurement and sampling uncertainties in monitoring urban drainage systems. *Water Science and Technology* **47** (2), 95–102.

Branisavljevic, N., Prodanovic, D. & Pavlovic, D. 2010 Automatic, semi-automatic and manual validation of urban drainage data. *Water Science and Technology* **62** (5), 1013–1021.

He, Q. P., Qin, S. J. & Wang, J. 2005 A new fault diagnosis method using fault directions in Fisher discriminant analysis. *AIChE Journal* **51** (2), 555–571.

Jolliffe, I. T. 2002 *Principal Component Analysis*, 2nd edn. Springer, Berlin, Germany.

Montgomery, D. C. 2009 *Introduction to Statistical Quality Control*, 6th edn. John Wiley & Sons, New York.

Mourad, M. & Bertrand-Krajewski, J. L. 2002 A method for automatic validation of long time series of data in urban hydrology. *Water Science and Technology* **45** (4–5), 263–270.

Perera, A., Papamichail, N., Barsan, N., Weimar, U. & Marco, S. 2006 On-line novelty detection by recursive dynamic principal component analysis and gas sensor arrays under drift conditions. *IEEE Sensors Journal* **6** (3), 770–783.

Poirier, P. 2013 Outlis automatiques d'evaluation de la qualité des données. MSc Thesis, Université Laval, Québec, QC, Canada.

Qin, S. 2009 Data-driven fault detection and diagnosis for complex industrial processes. In: *Proceedings IFAC Symposium SAFEPROCESS2009*, Barcelona, Spain, June 30–July 3, 2009.

Rieger, L. & Vanrolleghem, P. A. 2008 monEAU: a platform for water quality monitoring networks. *Water Science and Technology* **57** (7), 1079–1086.

Rosén, C., Rieger, L., Jeppsson, U. & Vanrolleghem, P. A. 2008 Adding realism to simulated sensors and actuators. *Water Science and Technology* **57** (3), 337–334.

Venkatasubramanian, V., Rengaswamy, R., Kavuri, S. N. & Yin, K. 2003 A review of process fault detection and diagnosis. Part III. Process history based methods. *Computer Chemical Engineering* **27** (3), 327–346.

Villez, K., Ruiz, M., Sin, G., Colomer, J., Rosén, C. & Vanrolleghem, P. A. 2008 Combining multiway principal component analysis and clustering for efficient data mining of historical data sets of SBR processes. *Water Science and Technology* **57** (10), 1659–1666.

Winkler, S., Kreuzinger, N., Pressl, A., Fleischmann, N., Gruber, N. & Ecker, M. 2002 Innovative technology for integrated water quality measurement. In: *Proceedings International Conference on Automation in Water Quality Monitoring (AutMoNet2002),* Vienna, Austria, May 21–22, 202 pp.

Yoo, C. K., Villez, K., Van Hulle, S. W. & Vanrolleghem, P. A. 2008 Enhanced process monitoring for wastewater treatment systems. *Enviromet rics* **19** (6), 602–617.

First received 30 November 2012; accepted in revised form 2 April 2013