Proceedings New Developments in IT & Water Conference.
Amsterdam, The Netherlands, November 4-6 2012

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

Janelcy Alferes[1], Sovanna Tik[1], John Copp[2] and Peter A. Vanrolleghem[1]

[1]model*EAU*, Université Laval, Département de génie civil et de génie des eaux, Québec, QC G1V 0A6, Canada
(Email: *janelcy.alferes@gci.ulaval.ca*)
[2]Primodal Inc., Hamilton, ON L8S 3A4, Canada

Abstract
In situ continuous monitoring at high frequency is used to collect water quality information of water bodies. However, it is crucial that the collected data be evaluated and validated for an effective monitoring programme and the proper interpretation of the large data set. Software tools for data quality assessment with a practical orientation are proposed. Since water quality data often contain redundant information, multivariate methods can be used to detect correlation and significant 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 help the user to apply the corrective action. The developed algorithms are illustrated with automated monitoring systems installed in an urban river and at the inlet of a wastewater treatment plant.

Keywords
Data quality assessment, Fault detection, On-line monitoring, Water quality

INTRODUCTION

A new generation of in situ automatic water quality monitoring stations is proposed adhering to the mon*EAU* vision (Rieger and Vanrolleghem, 2008). With flexibility and standardisation as the main drivers of recent development, important advances have been made regarding several monitoring tasks and measurement applications (water bodies, wastewater treatment plants, 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 is 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., 2006). 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 thus be valuable only if the data is properly validated. Given the size of the data sets, only automated data validation is an option.

In the last few years some methods have been developed for fault detection and isolation (FDI) in different fields (Venkatasubramanian et al., 2003, He et al., 2005). Traditional model-based approaches make use of the generation of residuals (difference between measured value and its prediction by a model) and their evaluation for decision making. 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 (Mourad and Bertrand-Krajewski, 2002; Branisavljevic et al., 2010).

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 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 e.g. for decision support and model based control of water systems.

MATERIALS AND METHODS

In situ monitoring stations

Primodal System 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 wastewater treatment plant (WWTP) in 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…), a UV spectrometer (nitrates, TOC, DOC, turbidity) and ion selective electrodes (potassium, ammonia). In this case, data from the summer of 2012 are used.

Figure 1. Sensors installation at the WWTP

Figure 2. Sensors installation 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). That implementation allows generating information-rich data but also complex and huge data sets. To increase the likelihood of good quality from the on-line measurements, the application of a maintenance protocol including cleaning and systematic calibration tasks was essential (Poirier, 2012).

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 at al., 2008; Yoo et al., 2008), which may result in a false perception of the monitored system and/or in erroneous control actions. Typical hard and soft faults in online sensors are shown in Figure 3. The detection and diagnosis of these kinds of sensor faults is 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 process monitoring methods based on principal components 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 their dimensionality. By exploring the original data set, PCA is used to find a new set of uncorrelated variables, called principal components (PC) which explain most of the data variability in a more visual coordinate system with fewer dimensions. Given an autoscaled [*m x n*] data matrix X for *n* process variables and *m* samples, performing PCA allows decomposing X as follows:

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

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 x a*] is the loading matrix and its column vectors (p_i) are called loadings or principal components of X. The matrix T [*m x 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 principal components 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 are a measure of the variance of X along each principal component 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 $T = X_{new}P$.

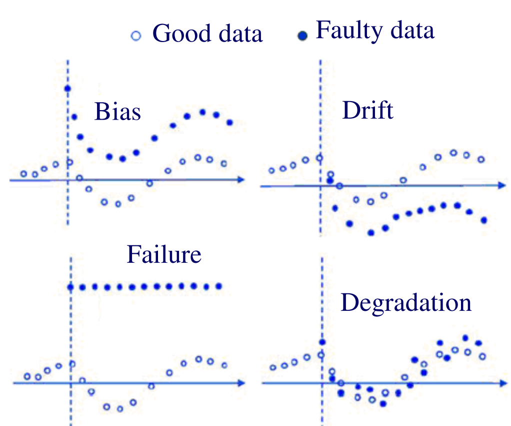

Figure 3. Common sensor faults (Reference: Yoo et al., 2008)

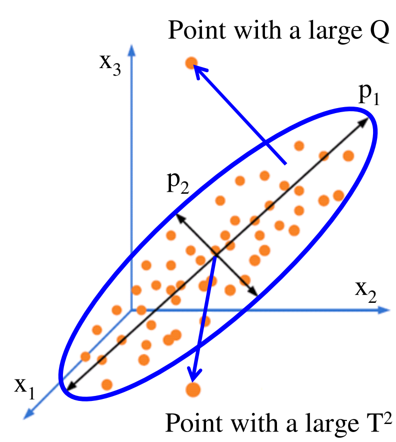

Figure 4. Geometrical interpretation of T^2 and Q statistics (Reference: Montgomery, 2009)

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. To that end 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 normalize squared scores:

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

where x is the sample vector and Λ_a^{-1} is the diagonal matrix containing the *a* eigenvalues associated with the *a* eigenvectors or principal components 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 90-95%) (Yoo et al., 2006). 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:

$$Q(k) = x^T(k)\left(I - PP^T\right)x(k) \qquad (4)$$

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 principal components within which the operating points normally 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 principal components.

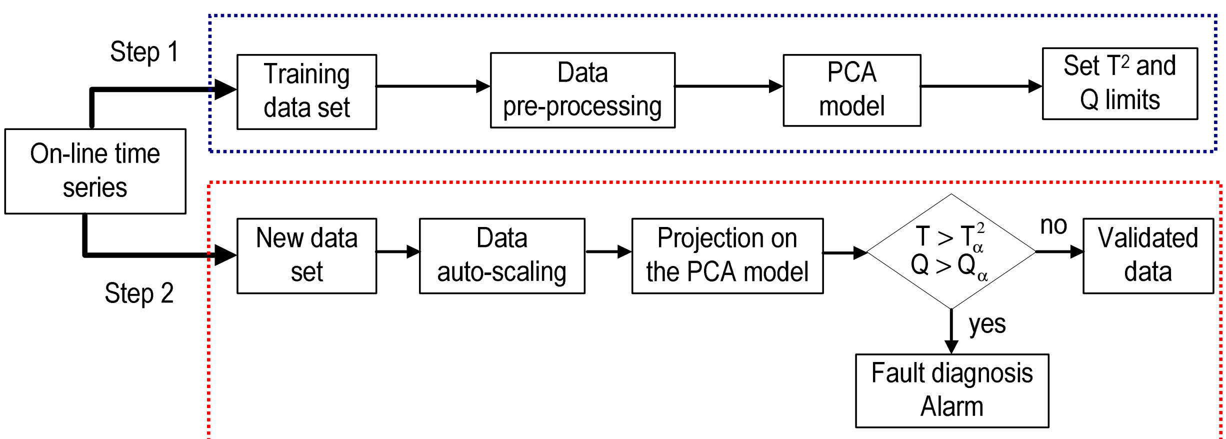

Figure 5. Proposed sensor validation procedure

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 is 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 is 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 this new data onto the reference PCA space. If one or several variables are found to deviate from the normal model area (expected variability) the T^2 and/or Q statistics will increase over 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.

Time series from 10 on-line variables at the inlet of the WWTP were considered including: Conductivity sensor 1 (Cond1), Temperature at conductivity sensor 1 (CondTemp1), pH sensor 1 (pH1), Temperature at pH sensor1 (pHTemp1), pH sensor 2 (pH2), Temperature at pH sensor 2 (pHTemp2), Turbidity sensor 1 (Turb1) and Turbidity sensor 2 (Turb2). All sensors recorded data at 5 seconds intervals. A representative training data set over a 3-day period was used to build the PCA model. Prior to the PCA modeling training data were properly auto-scaled (mean centering and variance scaling) and outliers were removed. Performing the PCA showed that the first three principal components 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, 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.

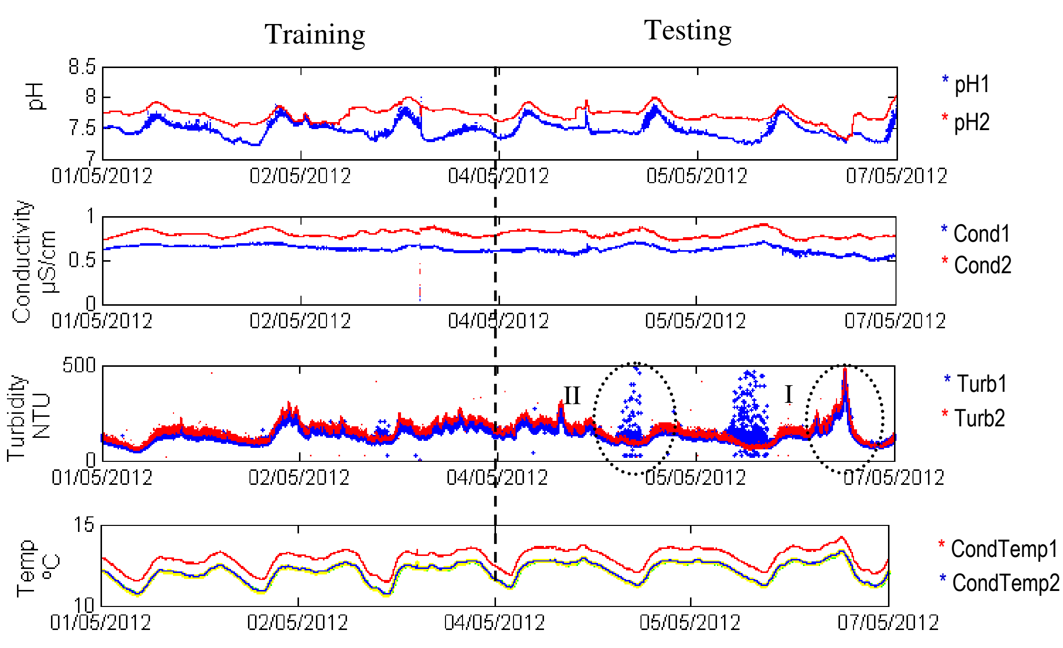

Figure 6. On-line measurements of turbidity, conductivity, pH and temperature at the inlet of the primary clarifier of the WWTP of Quebec City.

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. Missed calibration steps and the different ages of the sensors are suggested as explanation for the divergence. 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 principal components (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 a cluster in area I (in the direction of the turbidity measurements) that reveals changes in these variables. In fact, these samples are associated with a rain event on May 7th (see Figure 6) which induced an important variation in turbidity. Some outlying points are also identified around area II suggesting an abnormal behaviour or disturbance for these samples.

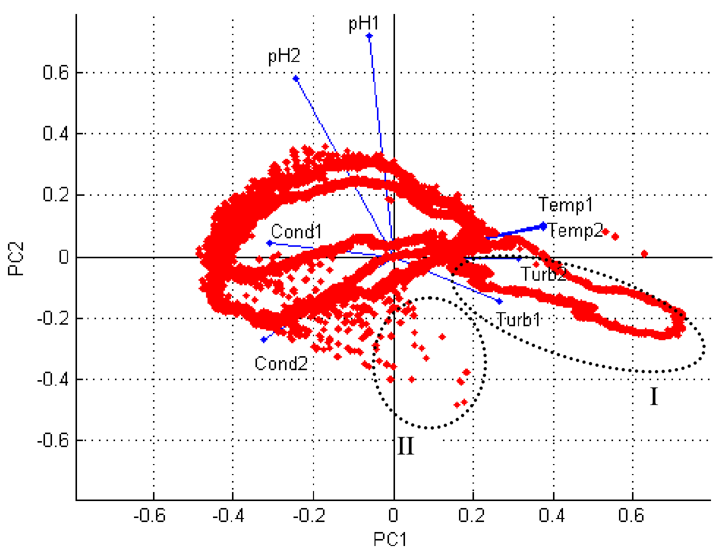

Figure 7. PCA representation, WWTP application

Monitoring of the T^2 and Q statistics (Figure 8) allowed for the detection 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 8a shows how, for samples in period I, the Q statistic is maintained inside the limit while T^2 detects variations in the data that are larger than the variations expected under normal operation. Some outliers can be detected as well. In period II, abnormal behaviour or disturbances are suggested for the turbidity data from sensor Turb1. While T^2 reveals important variations in the measurements (Figure 8b), Q also reveals important noise in the turbidity measurements for sensor Turb1 in this period as evidenced in Figure 6. Similar conclusions can be drawn from the pH and conductivity measurements.

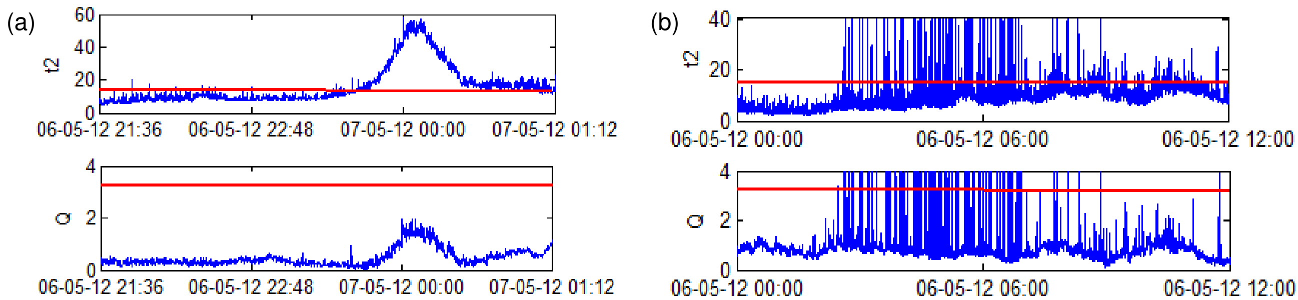

Figure 8. T^2 and Q statistics for data corresponding to certain periods of Figure 6. (a) Period I, (b) Period II. Red lines are limits that allow detecting faults in the monitored data series

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^+). Performing PCA over a representative training data set showed that three principal components can explain more than 85% of the total variance of the process. Looking at the scores from the 2-day testing data set in Figure 9 some outlying points can be identified in the marked areas (I, II) in the direction of potassium and turbidity measurements respectively. Graphical representation of the T^2 and Q statistics and time series for these variables in Figure 10 confirm an abnormal behaviour or disturbance for these samples around July 12th. In both periods not only T^2 revealed abnormal variations in the measurements but also Q identified events not taken into account in the model clearly indicating a faulty condition.

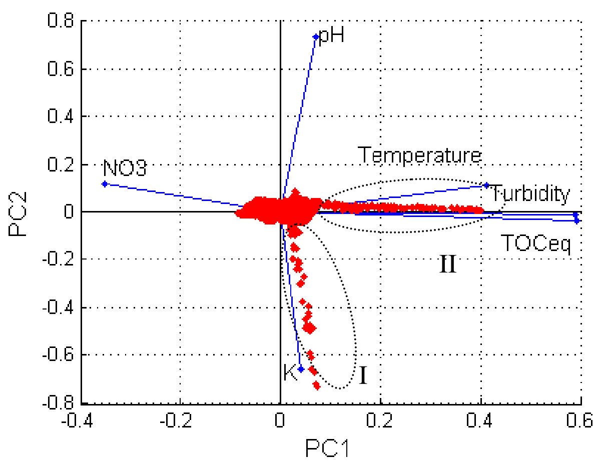

Figure 9. PCA representation, urban river

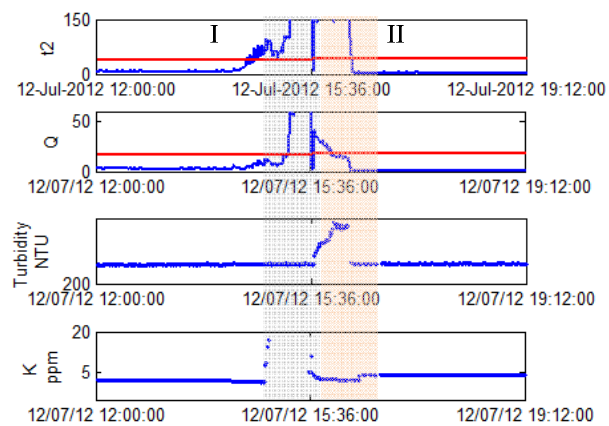

Figure 10. Time series for T^2, Q Turbidity and K^+ for periods I and II

CONCLUSIONS

As water quality measurements might be carried out in a difficult environment, 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 proposed and 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. Using this procedure enables the detection of different kinds of faults which 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.

ACKNOWLEDGMENTS

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 is currently Otto Monsted Guest Professor at the Technical University of Denmark.

REFERENCES

Alferes J., Poirier P. and 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. and 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. and 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. and Wang J. (2005). A new fault diagnosis method using fault directions in Fisher discriminant analysis. AIChE Journal, 51(2), 555-571.

Montgomery D.C. (2009). Introduction to Statistical Quality Control, 6th edition, John Wiley&Sons, New York.

Mourad M. and Bertrand-Krajewski J.L. (2012). 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. and 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. (2012). Outlis automatiques d'evaluation de la qualité des données. Mémoire, Université Laval, Quebec, 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. and P.A. Vanrolleghem. (2008). monEAU: A platform for water quality monitoring networks. Water Science and Technology, 57(7), 1079-1086.

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

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

Villez K., Ruiz M., Sin G., Colomer J., Rosén C. and 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. and 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.

Yoo C.K., Villez K., Van Hulle S. and Vanrolleghem P.A. (2006). Sensor validation and reconciliation for a partial nitrification process, Water Science and Technology, 53(4-5), 513-521.

Yoo C.K., Villez K. and Van Hulle S.W. (2008). Enhanced process monitoring for wastewater treatment systems. Envirometrics, 19(6), 602-617.