Automatic outlier detection in automated water quality measurement stations

Mémoire

Atefeh Saberi

Maîtrise en génie électrique
Maître ès sciences (M.Sc.)

Québec, Canada

Résumé

Des stations de mesure de la qualité de l'eau sont utilisées pour mesurer la qualité de l'eau à haute fréquence. Pour une gestion efficace de ces mesures, la qualité des données doit être vérifiée. Dans une méthode univariée précédemment développée, des points aberrants et des fautes étaient détectés dans les données mesurées par ces stations en employant des modèles à lissage exponentiel pour prédire les données au moment suivant avec l'intervalle de confiance. Dans la présente étude, ne considérant que le cas univarié, la détection de points aberrants est améliorée par l'identification d'un modèle autorégressif à moyenne mobile sur une fenêtre mobile de données pour prédire la donnée au moment suivant. Les données de turbidité mesurées à l'entrée d'une station d'épuration municipale au Danemark sont utilisées comme étude de cas pour comparer la performance de l'utilisation des deux modèles. Les résultats montrent que le nouveau modèle permet de prédire la donnée au moment suivant avec plus de précision. De plus, l'inclusion du nouveau modèle dans la méthode univariée présente une performance satisfaisante pour la détection de points aberrants et des fautes dans les données de l'étude de cas.

Abstract

Water quality monitoring stations are used to measure water quality at high frequency. For effective data management, the quality of the data must be evaluated. In a previously developed univariate method both outliers and faults were detected in the data measured by these stations by using exponential smoothing models that give one-step ahead forecasts and their confidence intervals. In the present study, the outlier detection step of the univariate method is improved by identifying an auto-regressive moving average model for a moving window of data and forecasting one-step ahead. The turbidity data measured at the inlet of a municipal treatment plant in Denmark is used as case study to compare the performance of the use of the two models. The results show that the forecasts made by the new model are more accurate. Also, inclusion of the new forecasting model in the univariate method shows satisfactory performance for detecting outliers and faults in the case study data.

Table of contents

List of tables

List of figures

Acronyms

AEM : Abnormal Event Management

ARMA : Auto-Regressive Moving-Average

ARIMA : Auto-Regressive Integrated Moving Average

KF : Kalman filter

MAD : Mean Absolute Deviation

MMFA : Multi Model Filtering Algorithm

NTU : Nephelometric Turbidity Unit

PC : Principal Component

PCA : Principal Component Analysis

RMSE : Root Mean Square Error

VARMA: Vector Auto-Regressive Moving Average

WS : Window Size

WWTP : Waste Water Treatment Plant

Acknowledgment

I would like to express my deepest gratitude, admiration and respect to my supervisors André Desbiens and Peter A. Vanrolleghem for their unwavering dedication, helpful advices and ongoing supports during my M.Sc. and this research work. I have learned a lot from them and without their helps this work would have never been completed.

I would like to extend my gratitude to Janelcy Alferes Castano for her technical supports and advices. I always felt comfortable to discuss my problems with her and she never hesitated to help me in spite of her occupations.

I also wish to express my thanks to my colleagues and friends in model*EAU* with whom I had the precious educational and social experiences.

Many thanks to my friends with whom I have shared many memorable moments during this challenging experience.

I cannot express my loving thanks to my dearest friend and husband, Mohammad who has never stopped believing in me. He has always reminded me of my capabilities and capacities when I was filled with stress, despair and gloom. We have passed many happy and sad moments together in our challenging life in Canada and during all these times he was kindly by my side. Without his supports and understanding, I wouldn't have finished this work.

Last, but not least, I will thank my parents. In spite of being thousand miles away from me, they have always been the source of my energy. All my exhaustions and frustrations have been vanishing away as I heard their kind voices. They have always admired me because of just being me! Thanks for being my parents, I love you so much!

1. Introduction

Fault detection and diagnosis which can be regarded as a part of the topic Data Quality Evaluation, has been an important problem in different industries. In the petrochemical industry, for instance, the Abnormal Event Management (AEM) problem was rated as the number one problem to be solved (Venkatasubramanian et al., 2003a). Data quality evaluation is a subject associated to all domains where sensors monitor the state of a system. Thus a variety of mechanisms and techniques have been developed to evaluate the quality of data obtained from application fields as different as nuclear power engineering, chemical engineering, air pollution control, medicine, water supply and distribution networks (Conejo et al., 2007).

Particularly in water systems effective management of water networks and their applications, such as waste water treatment systems and water pollution control, requires reliable information about water quality parameters. Control and modeling of water systems and subsequently making decisions about their performance and judgment about their variations can be carried out with more confidence if the acquired information is reliable. Especially the early diagnosis of process and sensor faults in the systems is of great importance since it facilitates a more precise understanding of the system and more trustable decisions can be made to control, model and supervise the system. Therefore, the reliability of data shall be evaluated attentively and in case of detecting faults, the quality of data shall be improved.

Implementation of automated water quality monitoring stations together with in situ continuous measurement devices measuring with high frequency has been a great step forward in obtaining a better description of water systems. The fast sampling rate enabled by these stations, helps getting along with fast dynamics of water systems and also improving the reliability of the data describing water quality parameters. In other words, the main purpose of implementing automated online water quality measuring stations is to monitor water bodies efficiently and to have a more advanced description of the system's fast dynamics.

However, the high sampling intensity enabled by online water quality measurement stations leads to collection of huge data sets consisting of a large number of physical-chemical parameters (Alferes et al., 2013b). However, because of the challenges in measuring conditions and sensors, especially during rain events, data collected by means of automated water quality measurement stations are prone to different sources of errors and faults (Rieger & Vanrolleghem, 2008; Hill & Minsker, 2010). In addition, due to the intrinsic properties of water systems, the data sets which describe water quality parameters are often co-linear or auto correlated and not normally distributed (Alferes et al., 2012). Thus, the main objective of implementing the automated water quality monitoring stations will not be efficiently attainable unless the quality of acquired data is verified.

Quality of data registered with automated water quality measurement stations can be largely affected by the conditions of the installed hardware in the field. Some procedures can be followed to improve the quality of acquired data. Regular manual cleaning of the measurement sensors, application of innovative measuring devices with self-cleaning system (Mourad & Bertrand-Krajewski, 2002), increasing the number of measurement sensors and selecting the measurement locations which meet certain criteria (Campisano et al., 2013) are examples of procedures that are followed in the field. However, despite taking all mentioned procedures, the reliability of sensors still remains insufficient and different types of faults in the data are inevitable.

To detect faults in practical systems, outliers in the data should be detected and removed before, since they can negatively affect the performance of the fault detection algorithms. Outliers are known as sample values that behave significantly different from the data points which are believed to depict normal behavior. Outliers might be generated due to different reasons. Sensor noise, temporary sensor failures and human-related errors are examples of sources that contaminate the data with outliers (Liu et al., 2004; Ting et al., 2007). Existence of outliers change considerably different features of the data like mean and variance and also might have negative impact over our interpretation from the system. Therefore, they must be detected, omitted or replaced by more reliable values (Alferes et al., 2012). In automated water quality measurement stations with high frequency which

lead to very large data sets, visual detection of outliers is not feasible thus the need for automated outlier detection methods is motivated (Pearson, 2002).

Peter Vanrolleghem and his research team (model*EAU*, Université Laval, Québec, Canada) have developed automated water quality monitoring stations, called mon*EAU*s, with collaboration of public organizations and private companies from North America and Europe (Rieger & Vanrolleghem, 2008). These stations, now available from Primodal Inc. RSM-30, can automatically register different water quality parameters with fast sampling rates in various measurement locations such as rivers, waste water treatment systems and sewers. They comprise sensors for conventional and physical-chemical parameters (temperature, dissolved oxygen, …) as well as innovative sensors like a UV spectrometer and an ion selective device together with water level measurement sensors (Rieger & Vanrolleghem, 2008). Univariate and multivariate methods to detect and replace outlier and to diagnose probable faults in real time have already been developed for these monitoring stations and practically tested successfully by Alferes et al. (2012; 2013b).

The univariate method proposed by Alferes et al. (2012) consists of three successive steps: outlier detection, data smoothing and fault detection. The first and the most challenging step is the outlier detection. This applied method is based on fitting an exponential smoothing model to the data and defining the smoothing model parameters according to historical data. The fitted model is then projected into the future to calculate a one step-ahead forecast value together with its prediction error interval. The prediction error interval is calculated by estimation of the standard deviation of the forecast error. The forecast value plus/minus a multiple of the standard deviation of the forecast error will be considered as the prediction error interval. If the real observed value falls outside this interval, it is considered an outlier and it is replaced by the forecast value.

This research project focuses on the outlier detection step of the previously proposed univariate method. The objective is to propose an alternative method to identify a model for the water quality parameters time series according to which the forecast data will be calculated. The idea is to find a model which results in a better fit to the system and a better forecast of the future behavior of the system in comparison with the exponential smoothing model, in particular with respect to the Root Mean Square Error (RMSE) criterion. The

outliers will then be identified automatically and replaced by the forecast value according to the same approach taken in the previously proposed univariate method. Performance of the new model in detecting outliers will be evaluated in comparison with the exponential smoothing model with respect to a particular set of criteria. In order to achieve this objective, the structure of the work presented in this thesis is as follows:

In the *Introduction*, the advantage of using automated water quality monitoring stations was discussed. The chapter contained an overview of the general properties of the data series obtained from automated water quality monitoring stations and the necessity of data quality evaluation in such systems.

In chapter 2, *Literature Review*, first a general introduction of fault detection methods found in literature will be presented ranging from laborious manual techniques to more sophisticated automatic methods. Then the water quality monitoring stations will be introduced and the measurement challenges which lead to the collection of datasets with specific properties will be discussed. Different data quality evaluation approaches found in literature will be presented and feasibility of their application to the case study system will be studied. The exponential smoothing models and their corresponding forecasting equations will be presented for the approach proposed by Alferes et al. (2012), to cope with challenging conditions of monitoring water quality parameters.

Chapter 3, *Materials and Methods,* begins with the presentation of the case study water quality monitoring stations whose measured water quality parameters will be used to evaluate the performance of the alternative model. The chapter continues with introducing the three steps of the univariate data quality evaluation method proposed by Alferes et al. (2012) and ends with a short discussion about the performance of the method.

Chapter 4, *Results*, principally includes a mathematical presentation of the alternative model for the system and the method followed to detect outliers in time series data. It also consists of a subchapter which revisits the exponential smoothing model used by Alferes et al. (2012) and discusses the advantage of using an alternative model to forecast the future behavior of the system. The chapter ends with a comparison of the new model and the

exponential smoothing model, regarding outlier detection and data quality evaluation performance.

Chapter 5 draws the conclusions of this work and the suggestions for possible future research plans.

2. Literature Review

A large number of data quality evaluation methods are proposed in literature ranging from simple visual pre-validation tools to more sophisticated analytical methods, artificial intelligence tools and statistical approaches. Some data quality evaluation methods can be widely applied to several areas while application of some methods is limited due to the assumptions made about their characteristics which make them incompatible to specific procedures. For example when applying a classical Kalman filter (KF), a linear model, quadratic performance criterion and Gaussian probability distribution for the observation noise are assumed (Gandhi & Mili, 2010). However, the performance of KF may degrade significantly in the presence of outliers or, in general, when the state or measurement noise is non-Gaussian (Chan et al., 2005; Ting et al., 2007). Automated water quality monitoring systems have specific characteristics that distinguish them from other systems and limit application of many of data quality evaluation approaches.

In this chapter, first a general introduction of the available approaches in literature will be presented. Subsequently, challenges of water quality monitoring will be discussed and application of the different methods will be practically validated considering the specific conditions of the case study system. An approach to model water quality parameters time series proposed by Alferes et al. (2012) will be discussed and mathematical details will be presented.

2.1. Introduction to data quality evaluation methods

In this section, different types of approaches will be presented and the way they regard the data quality evaluation problem will be discussed.

A very primary classification of the methods comprises off-line and on-line methods or according to another categorization, manual and automatic methods. Off-line data quality evaluation methods such as application of control charts when a reference measurement is available (Thomann et al., 2002) are too tedious and time consuming when huge volumes

of data are registered in on-line automated systems. Application of manual methods, on the other hand, is not feasible in such systems since it requires a full-time inspection operation which is not cost effective (Hill & Minsker, 2010). Accordingly, in automated on-line systems, a serious need for automatic on-line data quality evaluation tools is identified.

Automatic on-line data quality evaluation methods can be regarded as univariate or multivariate approaches according to Alferes et al. (2012). In univariate methods the information from single variables are extracted. Multivariate methods detect correlations in the high dimensional measurement space according to which useful information about the measurements is extracted. For a huge amount of correlated data an appropriate method to draw useful information from the data is needed (Aguadoa & Rosen, 2008). For instance in water quality measurement stations, multivariate methods can be used to infer significant information from highly correlated variables. Since univariate methods cannot handle correlation among variables, they can be only applied to evaluate the quality of single water quality parameters. A more reliable decision can be made about the performance of the measurement system if the evaluation result of any single variable done by an univariate method is regarded with respect to other variables.

Another fundamental classification of methods in literature concerns parametric (statistical) methods and non-parametric (model-free) methods. Statistical parametric methods either assume a known underlying distribution of the observations or, at least, they are based on statistical estimates of unknown distribution parameters (Ben-Gal, 2005). These methods are often unsuitable for high-dimensional data sets and for arbitrary data sets without prior knowledge of the underlying data distribution (Ben-Gal, 2005). Within the class of non-parametric approaches, a special categorization can be done for the data-mining methods, also called distance-based methods. These approaches are usually based on local distance measures and are capable of handling large databases (Ben-Gal, 2005). Another category of non-parametric methods consists of the clustering techniques. In these techniques, each object is assigned to the cluster of its nearest neighbor within a certain distance (Abu-el-zeet et al., 2002).

According to Venkatasubramanian et al. (2003a; 2003b; 2003c), fault detection strategies are largely influenced by the type of a priori knowledge which is available about the

system. The a priori knowledge can be developed from a fundamental understanding of the physics of the process expressed in terms of functional relationships between the inputs and outputs of the system which is referred to as, model-based knowledge. On the other hand, the a priori knowledge can be based on the past experience with the process and the availability of a large amount of historical data while it does not assume any form of model information.

Consequently, according to Venkatasubramanian et al. (2003c), from a modeling perspective, fault detection methods can be classified as model-based and process history-based approaches. The model-based methods can be subsequently categorized as quantitative or qualitative. In quantitative model-based methods, the a priori knowledge about the process is expressed in terms of mathematical functional relationship between inputs and outputs of the system while in qualitative model-based method this understanding is represented in terms of qualitative functions centered around different units in the process.

There are different ways in which the process history knowledge can contribute as a priori knowledge to the fault detection system. This is known as feature extraction from process history data. Again, one can classify the feature extraction process as either quantitative or qualitative. Quantitative feature extraction can be performed by either statistical or non-statistical methods.

It should be noted that Venkatasubramanian et al. (2003a) give a classification of quantitative model, qualitative model and process history based methods in terms of the intrinsic way these methods approach the problem of fault detection. For instance, KFs (classified under the quantitative model-based methods), which are based on state-space models, broadly use quantitative approaches for generating fault detection results. Similarly neural network methods which basically approach the problem of fault detection from a pattern recognition point of view and hence are classified as process history based methods intrinsically use state-space models (Venkatasubramanian et al., 2003a). If such an overlap of method classifications is accepted, this classification of methods is comprehensive.

By presenting a general introduction of data quality evaluation methods we can have a better perspective on the existing approaches and the feasibility of their application to the case study system. In next section, first the general information about the water quality monitoring stations will be presented. Then the challenging conditions of monitoring of water quality parameters which lead to specific properties of data series will be discussed. Subsequently, different fault detection methods will be presented and the efficiency of their application to the system will be discussed.

2.2. Water quality monitoring challenges and alternatives

2.2.1. The mon*EAU* water quality monitoring stations

According the mon*EAU* vision for water quality monitoring presented by Rieger and Vanrolleghem (2008), the concept of the resulting mon*EAU* water quality monitoring stations is demonstrated in Figure 1. This figure shows the different components of the monitoring network. It includes the measurement sensors, the monitoring stations and a central server. The mon*EAU* monitoring stations consist of the input/output modules, the industrial computer and the data transmission modules. The sensors measure the water quality parameters and transmit the data to the monitoring stations. The data can be registered in the stations for different monitoring purposes or data quality evaluation or can be transmitted to the central servers for further monitoring and control purposes. The communication protocol between the sensors and the base stations can be quite diverse, but preference is given to bus protocols like Profibus. To connect the monitoring stations to the central server, different telemetry modules are available such as telephone line, xDSL, dedicated radio link and satellite. The monitoring stations are designed to be flexible to be used at different locations such as rivers, WWTPs and sewers with a wide range of types of sensors and sampling methods and with a large set of standard communication protocols for sensor connections. The number of measured variables used so far with these stations can be between 3 and 10 and the sampling time, T_s, can be 5 seconds or 1 minute. Figure 2 shows a typical monitoring station installed in the field, in this case at the inlet of the

primary clarifier of the Lynette wastewater treatment plant in Denmark. Figure 3 shows the equipments panel of the mon*EAU* station.

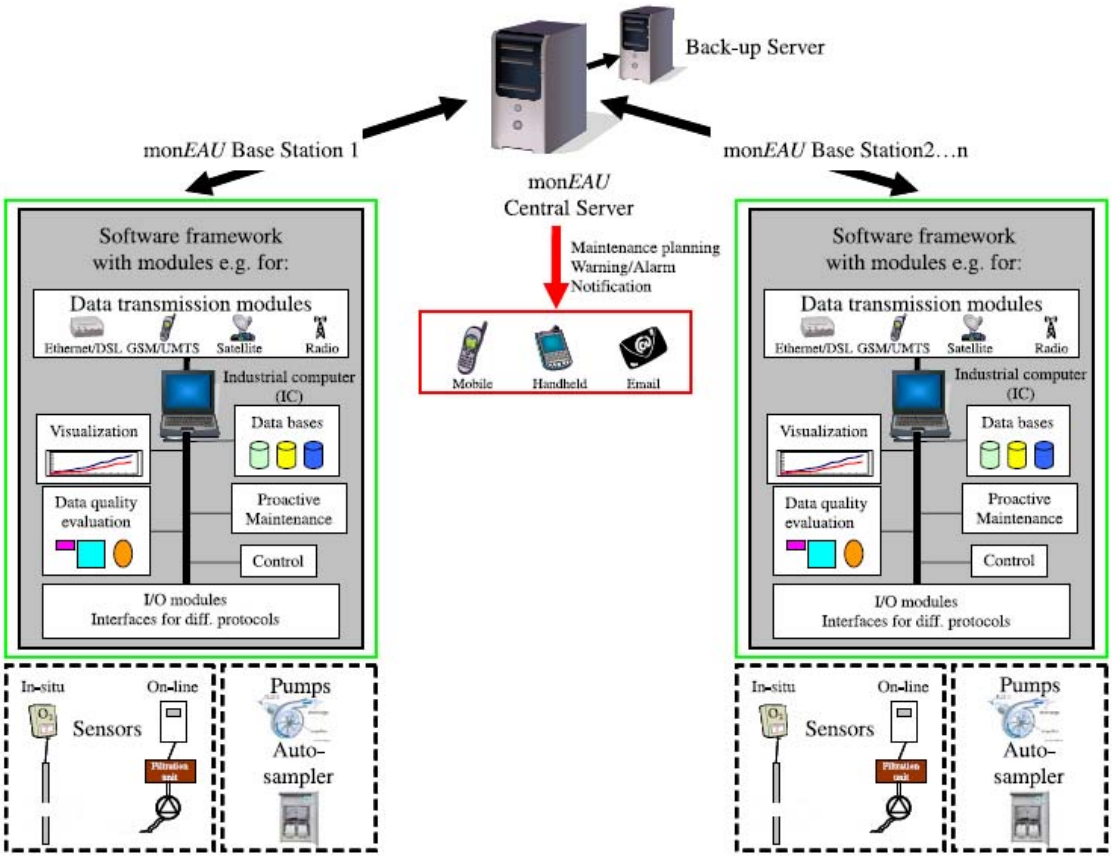

Figure 1 – The mon*EAU* network concept

Figure 2 – The mon*EAU* monitoring station installed at the inlet of the primary clarifier at the Lynette wastewater treatment plant in Denmark

Figure 3 – The equipment panel of the mon*EAU* monitoring station

2.2.2. Characteristics of the water quality monitoring system

As discussed in chapter 1 the high sampling intensity enabled by online water quality monitoring stations leads to collection of huge data sets consisting of a large number of physical-chemical parameters (Alferes et al., 2013b). However, since the measurements are carried out in very difficult and challenging conditions, especially during rain events, data collected from such systems is frequently affected by different types of faults like shift, drift, inaccuracy and sometimes complete failure of the measurement system (Alferes et al., 2013a). In addition, due to the intrinsic properties of water systems, the data sets which describe water quality parameters are often auto-correlated, not normally distributed and noisy (Alferes et al., 2012). Consequently, finding a fault detection approach which is compatible with the operating conditions of this stochastic system with fast dynamics is challenging.

2.2.3. Methods in literature and their applicability in the system

There is an abundance of methods in literature covering model-based and process history-based methods (Venkatasubramanian et al., 2003a) ranging from simple pre-validation tests (Bertrand-Krajewski et al., 2000; Mourad & Bertrand-Krajewski, 2002) to more sophisticated statistical univariate or multivariate tests, model-based and data mining methods (Branisavljevic et al., 2011). Due to the specific characteristics of the case study system, application of a large number of data quality evaluation methods in practice might not be feasible. Model-based methods for instance, need an explicit model of the system to generate residuals (the difference between the actual and expected behavior according to the model) and evaluate them statistically for fault detection purposes (Venkatasubramanian et al., 2003a). However, finding an exact model that explains all physical and chemical variations that might occur especially in this system is a difficult task (Alferes et al., 2013a). On the other hand, process history-based methods can be employed to extract the relationship between input and output of the system according to the history of data without having an accurate model of the system (Venkatasubramanian et al., 2003c; Alferes et al., 2013a).

Automated water quality measurement stations need automatic methods for their quality assurance. As mentioned before, the off-line manual methods are not appropriate for this system. Application of semi-automated data quality evaluation tools in water quality monitoring systems also result in a relatively large percentage of data loss (5-40%) (Bijnen & Korving, 2008; Thomann, 2008). This can be caused by either the inefficiency of the practical methods or by the intrinsic invalidity of data in such systems. On the other hand, on-line application of data quality evaluation approaches is limited to the methods which are not computationally complicated.

Statistical methods have been broadly addressed in literatures as powerful techniques in fault detection in different fields (Willsky, 1976; Basseville, 1988; Conejo et al., 2007; Patcha & Park, 2007); however to apply them safely, accurate information about the system characteristics needs to be available (Conejo et al., 2007). Due to the intrinsic properties of the system, application of pure statistical methods (such as statistical classifiers) might not be feasible since not all behaviors of the system can be expressed as exact statistical distributions (Patcha & Park, 2007).

Patcha and Park (2007) discussed the application of machine learning techniques in the domain of fault detection. They are capable of changing their execution strategies and improve their performance based on previous information. Despite the variety of machine learning approaches available in this domain, these techniques have drawbacks that make them inappropriate for real-time fault detection (Hill & Minsker, 2010). The major problem with many of the machine learning techniques is that they are computationally complex for large volumes of data as collected by automated water quality measurement stations (Patcha & Park, 2007). However, as proposed by Patcha, the idea of the Sliding Window approach in the frame of machine learning techniques can be inspiring to find a way to get adapted to the varying dynamics in the case study system. The idea of using a moving-window approach has also been proposed by Liu et al. (2004) to capture the dynamic variations in on-line process data for outlier detection purposes.

As previously discussed in 2.1, according to the process history-based model categorization (Venkatasubramanian et al., 2003c), there are feature extraction methods that present the historical process data as a priori knowledge to the fault detection system. This is also

known as data-mining in the literature. It is the process of extracting patterns from the data to allow later diagnosis according to the deviation of future behavior with the historical pattern (Patcha & Park, 2007; Hill & Minsker, 2010). For stochastic systems, like the water quality monitoring stations, whose future state is not completely determined by the past and present status of the system, statistical feature extraction methods can be useful too.

The previously discussed idea of generating residuals is also known as analytical redundancy in literature (Basseville, 1988; Bloch et al., 1995; Hill & Minsker, 2010). In this approach the output of a model which is fitted to the data stream is considered as a redundant sensor whose measurement can be compared with that of the actual sensor. Classification of a data point as anomalous or non anomalous is performed according to the difference between model prediction and sensor measurement considering a threshold value (Pastres et al., 2003; Garcia et al., 2010; Hill & Minsker, 2010; Alferes et al., 2012). Since faults in the system cause changes in state variables or model parameters, analytical redundancy for fault detection in dynamic systems, mostly concerns monitoring of estimated states or model parameters of the system (Venkatasubramanian et al., 2003a). According to the solution given by Willksy (1976), Basseville (1988) discussed the application of KFs to give optimal state estimates according to the model of the system. KF is a well known model-based tool that uses underlying dynamic models to estimate the current state of process variables given the noisy measurement data (Bai et al., 2006). The filter can also estimate the corresponding output data by using the estimates of the states of the dynamic system (Ting et al., 2007).

However, as mentioned at the beginning of chapter 2, some assumptions are made in the application of the classical KFs that may degrade their efficiency (Chan et al., 2005; Ting et al., 2007; Gandhi & Mili, 2010). To address this problem, different solutions have been presented in literature comprising consideration of non-Gaussian distributions for random variables (West, 1981; Smith & West, 1983) or addressing the sensitivity of the squared error criterion to noises (Huber, 1964). To get along with the nonlinearity of systems, Extended KFs can be used to estimate the states of the system by a nonlinear state space model of the system (Basseville, 1988).

Application of KFs to water quality monitoring stations can be challenging because of the specific properties of these systems. The mentioned robustification approaches can be numerically complicated and thus might not be applicable to the case study water system. In addition, due to the existence of unknown disturbances in this system, the stochastic behavior of the monitored water quality parameters can hardly be precisely known and thus, one single model cannot describe the behavior of the system. A possible solution can be adopting the system model by using the Extended KFs. Another possible approach is the application of a bank of KFs, working independently, designed for all possible modes of behavior in the system (Basseville, 1988; Maldonado et al., 2010).

Maldonado et al. (2010) presented the Multi Model Filtering Algorithm (MMFA). In their method, at the calibration phase that uses the history of data, a set of linear models is identified for different behavioral modes of the system. For each of the identified models a KF is designed according to which the states of the system are estimated. To improve state estimation by applying Bayes' rule the conditional probability of each model to represent the actual observed system behavior based on input and output measurements is determined. The final estimated state of the system is calculated as a weighted sum of model probabilities and their associated states. In water quality monitoring systems, it is hard to find the optimal number of possible models of the system; therefore the size of bank of KFs may increase (Venkatasubramanian et al., 2003a). This is the reason that the application of this approach in this system may not be as straightforward as it seems.

Auto-Regressive data-driven methods are a class of data quality evaluation approaches that are widely used in literature. They are known to be successful in empirical modeling of time series data. Hipel and McLeod (1978) demonstrate that data-driven Auto-Regressive Moving-Average (ARMA) models can be successful in modeling hydrological time series data. According to them, the stochastic model is fitted to the system empirically and it does not exactly represent the physical model of the system. Similarly, Berthouex and Box (1996) state that an Auto-Regressive Integrated Moving Average (ARIMA) model can be fitted empirically to the stochastic time series obtained from waste water treatment plant to auto-forecast (uses just its own past to forecast future) future values of the series. Although the obtained model is identified empirically and does not exactly represent all physical-

chemical phenomena, it can be interpreted as a faithful description of physical-chemical realities. Garcia et al. (2010) also use a Vector Auto-Regressive Moving Average (VARMA) model (which is the extension of ARIMA model to the multivariate case) to predict and detect failures in railway networks. They identify parameters of the model using the Maximum Likelihood approach.

2.2.4. The alternative approach

Considering the intrinsic characteristics of the automated water quality measurement systems, it can be concluded that the alternative method should be automatic to scale well with the large volume of data, on-line applicable in real-time and fast enough to get along with the rate of data collection. In addition, it should not be computationally complicated and should not need any form of model as a priori knowledge. Since finding a physics-based model is hard while a large volume of data history is available, a data-driven time series model can be used (Hill & Minsker, 2010). Consequently, it is proposed to use the ARMA model as an alternative to the expected characteristics. The model can be identified according to the history of the data and the expected behavior in the future can be forecasted based on the identified model. To adapt to the varying dynamics, the moving window idea can be employed and to detect outliers, according to the analytical redundancy concept, deviation of the expected behavior from the real behavior can be identified.

2.2.5. Methods proposed by Alferes et al. (2012)

Regarding the utilization of data-driven models, Alferes et al. (2012) discussed two different approaches to assess data quality in automated water quality monitoring stations according to the history of data without having a theoretical model of the system. One is the univariate method which employs time series data of a single variable to check the acceptability of the measurement noise (Campisano et al., 2013) by means of autoregressive models. The other one is the multivariate method which infers significant

information from highly correlated variables based on Principal Component Analysis (PCA) (Yoo et al., 2008).

The univariate method comprises three consecutive steps (Figure 4). The first and the most challenging step is the outlier detection. According to this approach, a 3rd order exponential smoothing model is fitted to the time series data and its parameters are estimated. The model is then projected into the future to generate the one-step ahead forecast data. The standard deviation of the forecast error (the difference between the forecast and observed values) is predicted at each time step by means of the 1st order exponential smoothing model. The prediction interval is defined by adding or subtracting a multiple of the standard deviation of forecast error to the forecast value. If the observed value at each time step falls out of the prediction interval, it will be regarded as an outlier. Once an outlier is detected, it is replaced by the forecast value. The output of this step is called the "Accepted data" which is smoothed in the next step of the method, using a kernel smoother to remove noise. Potential sensor faults are then identified in the third step by applying acceptability limits to the data features extracted from the filtered data calculated in the previous step.

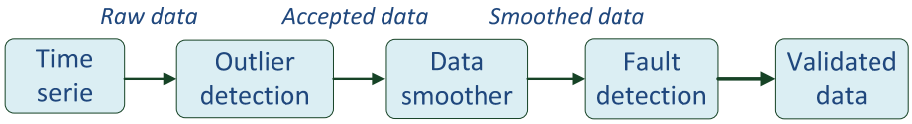

Figure 4 – Three steps of the univariate method proposed by Alferes et al. (2012)

Since water quality data are highly auto and cross-correlated and univariate methods cannot intrinsically handle correlation among variables, Alferes et al. (2012) propose a multivariate method to infer significant information from highly correlated variables.

The multivariate method is based on Principal Component Analysis (Yoo et al., 2008). In the PCA method, a new set of uncorrelated and orthogonal variables, called Principal Components (PCs), are extracted. Each PC is a linear combination of original variables which describe the largest process variability in a space with fewer dimensions than the original one. In order to detect probable faults in the multivariate method, it is proposed to calculate two statistics, T^2 and Q, which describe the fit of the model to the system.

Comparison of the calculated statistics with their correspondent confidence intervals, leads to detection of probable faults in the original system.

The univariate and multivariate methods developed by Alferes et al. (2012) have been successfully applied to detect and replace doubtful data and to diagnose probable faults in real time data series collected by the automated monitoring stations in model*EAU*. In this research project, the univariate method is regarded with special attention. The outlier detection step will be studied in more details and its efficiency will be commented. The objective is to provide improvements to the univariate method regarding the model fitted to the system and the way the one-step ahead forecasts are generated.

2.3. The objective of the project

As mentioned before, the objective of this research project is to improve the univariate data quality evaluation method of Alferes et al. (2012) that detects outliers and probable faults in water quality parameters measured by the automated water quality monitoring stations. The evaluation method consists of three consecutive steps and the focus of this work will be on the first step which is where outliers are detected. To detect outliers in the univariate method, a 3rd order exponential smoothing model is fitted to the time series data and is used to perform a one-step ahead forecast. In this work the exponential smoothing models will be critically evaluated and the manner in which they fit to the data series and forecast one-step ahead will be discussed. Subsequently, another forecasting model will be suggested to replace the 3rd order exponential smoothing model to give one-step ahead forecasts. As for the exponential model this alternative model must be accompanied with an approach to adapt the model to the varying dynamics in the data. The proposed forecasting approach will be applied to a case study water quality parameter and the accuracy of the one-step ahead forecasts will be compared to the ones obtained by the 3rd order exponential smoothing model. Both forecasting models will finally be compared through their performance within the univariate data quality evaluation method. Their efficiency will be compared in terms of detecting probable faults in the case study time series according the specified criteria.

2.4. Exponential smoothing models in forecasting time series

The performance of the outlier detection step in the univariate method is principally based on forecasting time series data according to the historical behavior by means of exponential smoothing models. Therefore, in this section more details are presented about these models and the way they generate forecasts.

As mentioned by Montgomery et al. (1990), exponential smoothing is a popular method for smoothing discrete time series in order to forecast immediate future. Exponential smoothing models use special weighted moving averages and a seasonal factor that is multiplied by the weighted moving average in order to forecast the immediate future. These weighted moving averages are referred to as smoothing statistics. The exponential smoothing models are an extension of the moving average model. Generally, an exponential smoothing method uses three smoothed statistics that are weighted, so that the more recent the data, the more weight is given to the data in producing the forecast. These three averages are referred to as single, double, and triple smoothing statistics and are moving averages that are weighted in an exponential declining way.

According to Montgomery et al. (1990), forecasting systems use three separate forecasting equations: one model called a constant model, the second called a linear model and a third called a quadratic model. The constant forecast uses only the single smoothed statistic and performs well when the time series has little trend. The linear forecast model uses the single and double smoothed statistics and is good when there is a linear trend in the time series. The quadratic forecast model uses all three statistics namely, single, double, and triple smoothed statistics.

In the next section, different orders of exponential smoothing models will be discussed. It includes details about the models, estimation of their unknown parameters and the way they produce forecasts for the future according to Montgomery et al. (1990). Since the objective is to improve the quality of the model in generating more accurate forecasts, a mathematical revision of the models is essential. An alternative model can then be proposed once all the details are understood about the current model.

2.4.1. Simple exponential smoothing for a constant process

If one could assume that the average of a time series does not change over time or if it does, it changes very slowly, the fitted discrete-time model might take the form:

$$x_k = b + \varepsilon_k \tag{1}$$

where k represents a discrete time sample, x_k is the data value at time k, b is the mean value of the data set and the model's unknown parameter and ε_k denotes a random error term corresponding to that part of the data that cannot be fitted by the model. The random term in the model is considered to have an expected value of zero and the constant variance.

At the end of time step k, a data history $x_1, x_2, \ldots, x_k$ is available according to which we wish to estimate the unknown model parameter b. To do so, the simple (first) exponential smoothing model can be used.

In this forecasting system, the model parameters are re-estimated at each time k. In order to take into account the data observed in the most recent period, it can be assumed that at the end of period k we have available the estimate of b made at previous time $k-1$, i.e. $\hat{b}_{k-1}$ and the actual value of current time k, x_k to calculate an updated estimate $\hat{b}_k$. A reasonable way to get the updated estimate is to modify the old estimate with a fraction of the forecast error. The forecast error at time k, e_k results from the difference between the current observed value and the old estimate.

$$e_k = x_k - \hat{b}_{k-1} \tag{2}$$

The new estimate is computed according to the following equation:

$$\hat{b}_k = \hat{b}_{k-1} + \alpha[x_k - \hat{b}_{k-1}] \tag{3}$$

α, namely the smoothing constant, is the fraction according to which we desire to contribute the forecast error to the calculation of the new estimate.

For simplification of the further development of the methods, we define the first exponentially smoothed statistic as $s_k \equiv \hat{b}_k$. So the above equation will be:

$$s_k = s_{k-1} + \alpha[x_k - s_{k-1}] \tag{4}$$

In other words:

$$s_k = \alpha x_k + (1-\alpha) s_{k-1} \tag{5}$$

The presented procedure is called Simple (First) Exponential Smoothing.

Generally speaking, the smoothed statistic constitutes a weighted average of past observations where the weights sum to unity and decrease geometrically with the age of the observations. These statements can be proven if s_{k-1} in the right-hand side of equation (5) is substituted recursively by its equivalent from equation (5). This operation leads to the following equation:

$$s_k = \alpha \sum_{m=0}^{k-1} (1-\alpha)^m x_{k-m} + (1-\alpha)^k s_0 \tag{6}$$

where s_0 is the initial estimate of b and is used to initialize the algorithm.

The weights add to 1 since:

$$\alpha \sum_{m=0}^{k-1} (1-\alpha)^m = \alpha \frac{1-(1-\alpha)^k}{1-(1-\alpha)} = 1-(1-\alpha)^k \tag{7}$$

The smoothing constant α presented in equation (3) is the forgetting factor in this method which controls the rate of decay and determines the behavior of the forecast system with respect to the changes in b. Small values of α give more weight to the historical data promoting a slow response. With large values of α, more weight is assigned to the current observation while leading to a faster response. Generally, the α value can range between 0 and 1, taking typical values between 0.01 and 0.3.

As already mentioned and proven in equation (6), the weights associated to previous observations decay with time and since these weights decrease exponentially with time, the name exponential smoothing is assigned to this procedure.

For a large enough value of k, i.e., the term $(1-\alpha)^k s_0$ is very close to zero, the exponential smoothing operator described in equation (6) leads to an unbiased estimate of the real value of the process average b, since:

$$E(s_k) = E[\alpha \sum_{m=0}^{\infty} (1-\alpha)^m x_{k-m}] = \alpha \sum_{m=0}^{\infty} (1-\alpha)^m E(x_{k-m}) = b\alpha \sum_{m=0}^{\infty} (1-\alpha)^m = b \tag{8}$$

So, it is logical to take s_k as the estimator of the unknown parameter of the model b at time k:

$$\hat{b}_k = s_k \tag{9}$$

Finally, because a constant model is considered, the forecast value of the data for any future time steps $k+j$ would be:

$$\hat{x}_{k+j} = \hat{b}_k \tag{10}$$

To sum up, a first order exponential smoothing operator fits a constant model to the time series data (equation (1)) whose only unknown parameter is re-estimated at each time k as the new data point emerges (equations (4) and (9)). Since a constant model is considered, for any time in future the estimated model at current time is projected into future and according to that the forecast of the data is produced (equation (10)).

2.4.2. Double exponential smoothing for a linear trend process

A second order exponential smoothing model is the extension of the first order exponential smoothing to the cases in which the mean of the process changes linearly with time and there is a trend in the data, according to the following discrete-time model:

$$x_k = a + bk + \varepsilon_k \tag{11}$$

where a the intercept and b the slope of the linear model are the unknown parameters which are supposed to be estimated. The definitions of x_k and ε_k remain the same as the ones for the constant model in the previous section.

If the simple exponential smoothing operator described in equations (6) is applied to the linear process defined in equation (11) and the expected value is calculated, we obtain:

$$\begin{aligned} E(s_k) &= \alpha \sum_{m=0}^{k-1} \beta^m E(x_{k-m}) + \beta^k s_0 \\ &= \alpha \sum_{m=0}^{k-1} \beta^m [a + b(k-m)] + \beta^k s_0 \end{aligned} \tag{12}$$

where β is equivalent to $(1 - \alpha)$ and is considered for simplification of the demonstrations.

For sufficiently large value of k, β tends to zero, then equation (12) can be expressed as:

$$E(s_k) = (a + bk)\alpha \sum_{m=0}^{\infty} \beta^m - b\alpha \sum_{m=0}^{\infty} m\beta^m = a + bk - \frac{\beta}{\alpha} b \tag{13}$$

and since,

$$E(x_k) = a + bk \tag{14}$$

for equation (13) we have:

$$E(s_k) = E(x_k) - \frac{\beta}{\alpha} b \tag{15}$$

This means that the expected value of the first order exponentially smoothed statistic, when applied to a linear model, lags behind the process by a value equal to $\frac{\beta}{\alpha} b$.

If the exponential smoothing operator is applied once more to equation (5), we obtain:

$$s_k^{[2]} = \alpha s_k + (1-\alpha) s_{k-1}^{[2]} \tag{16}$$

where $s_k^{[2]}$ denotes double (second-order) exponential smoothing.

Similarly for the output of equation (15) we can show that:

$$E(s_k^{[2]}) = E(s_k) - \frac{\beta}{\alpha} b \tag{17}$$

And consequently,

$$b = \frac{\alpha}{\beta}[E(s_k) - E(s_k^{[2]})] \tag{18}$$

So it is reasonable to estimate *b* at the end of time period *k* as:

$$\hat{b}_k = \frac{\alpha}{\beta}[s_k - s_k^{[2]}] \tag{19}$$

By substituting equation (17) in equation (15), the expected value of the data at the end of the time period *k* can be obtained as:

$$\begin{aligned} E(x_k) &= E(s_k) + \frac{\beta}{\alpha} \cdot \frac{\alpha}{\beta}[E(s_k) - E(s_k^{[2]})] \\ &= 2E(s_k) - E(s_k^{[2]}) \end{aligned} \tag{20}$$

Therefore, it seems reasonable to say:

$$\hat{x}_k = 2s_k - s_k^{[2]} \tag{21}$$

On the other hand, estimation of a can be performed according to two different approaches which both lead to the same estimation of the intercept. One is to estimate the intercept at the original time origin by employing equations (19) and (21), as:

$$\hat{a}_k = \hat{x}_k - k\hat{b}_k = 2s_k - s_k^{[2]} - k\frac{\alpha}{\beta}\left(s_k - s_k^{[2]}\right) \tag{22}$$

According to this approach, the calculation of the forecast value made at time k for any time step j in the future will be:

$$\hat{x}_{k+j} = \hat{a}_k + \left(k + j\right)\hat{b}_k \tag{23}$$

Another approach is to think the origin of time as shifted to the end of time period k and then estimate the current-origin intercept according to the following equation:

$$\begin{aligned} \hat{a}_k &= \hat{x}_k \\ &= 2s_k - s_k^{[2]} \end{aligned} \tag{24}$$

In this case, the forecasting equation for any time step in future is as follows:

$$\hat{x}_{k+j} = \hat{a}_k + j\hat{b}_k \tag{25}$$

Initial values of s_0 and $s_0^{[2]}$ are obtained from initial estimates of the two coefficients a and b which may be developed through simple linear regression analysis of historical data.

2.4.3. Triple exponential smoothing for a process with quadratic model

A third order (triple) exponential smoothing model takes into account trends and seasonal changes. The corresponding model takes the quadratic form:

$$x_k = a + bk + \frac{1}{2}ck^2 + \varepsilon_k \tag{26}$$

According to the explanations presented in the previous sections, assuming that we have estimates of the model parameters based on the original origin of time, the forecasting equation at the end of period k for τ step ahead is:

$$\hat{x}_{k+j} = \hat{a}_k + (k+j)\hat{b}_k + \frac{1}{2}(k+j)^2 \hat{c}_k \tag{27}$$

However, if we define the origin of time at the end of the period k, the coefficients will take different values and the forecasting equation will be:

$$\hat{x}_{k+j} = \hat{a}_k + \hat{b}_k j + \frac{1}{2}\hat{c}_k j^2 \tag{28}$$

For the latter case, the coefficients of the model, $\hat{a}_k$, $\hat{b}_k$ and $\hat{c}_k$ are computed using the first, second and third exponentially smoothed statistics, $s_k, s_k^{[2]}$ and $s_k^{[3]}$ calculated at the end of time period k. The smoothed statistics are calculated by:

$$\begin{aligned} s_k &= \alpha x_k + (1-\alpha) s_{k-1} \\ s_k^{[2]} &= \alpha s_k + (1-\alpha) s_{k-1}^{[2]} \\ s_k^{[3]} &= \alpha s_k^{[2]} + (1-\alpha) s_{k-1}^{[3]} \end{aligned} \tag{29}$$

Once the statistics have been calculated, the coefficients of the model are obtained as follow:

$$\begin{aligned} \hat{a}_k &= 3s_k - 3s_k^{[2]} + s_k^{[3]} \\ \hat{b}_k &= \frac{\alpha}{2(\alpha-1)^2}\left[(6-5\alpha)s_k - 2(5-4\alpha)s_k^{[2]} + (4-3\alpha)s_k^{[3]}\right] \\ \hat{c}_k &= \left(\frac{\alpha}{\alpha-1}\right)^2 \left(s_k - 2s_k^{[2]} + s_k^{[3]}\right) \end{aligned} \tag{30}$$

Therefore, in the third order exponential smoothing model, a quadratic model (equation (26)) will be considered for the system. The unknown parameters of the model, a, b and c, can be estimated at each time step through equations (30), while the smoothing statistics have been already calculated through equations (29). The forecast for any time step in future can be generated by projecting the quadratic model into the future through equation (27).

2.4.4. Simple exponential smoothing in calculating the Standard Deviation of forecast error

The exponential smoothing models are used in the univariate method to first, produce the one step-ahead forecasts and then to calculate the standard deviation of the forecast error to detect and replace outliers in water quality parameter time series. At each time step, the forecast is calculated together with its prediction error interval which gives the amount by which the real observed data can deviate from the forecast data.

The prediction error interval is indentified by analyzing the one-step-ahead forecast error at each time step, $e_k(1)$ calculated as:

$$e_k(1) = x_k - \hat{x}_k \tag{31}$$

where $\hat{x}_k$ is the forecast for time k made at previous time step, $k-1$ and x_k is the real observed value at time k.

To provide an estimation of the local variance and to quantify the extent by which the actual value differs from the forecast according to Montgomery (2009), a method is applied to estimate the variance of forecast error, σ_e^2, through the estimation of the Mean Absolute Deviation (MAD), Δ, by means of a simple exponential smoothing model. According to this method, assuming that the forecast error is normally distributed at time k, the estimate of σ_e^2 is obtained as:

$$\hat{\sigma}_{e,k} = 1.25\hat{\Delta}_k \tag{32}$$

where $\hat{\sigma}_{e,k}$ is the estimate of the standard deviation of forecast error for time k and $\hat{\Delta}_k$, calculated according to the 1st order exponential smoothing method as:

$$\hat{\Delta}_k = \eta\left|e_k(1)\right| + (1-\eta)\hat{\Delta}_{k-1} \tag{33}$$

where η is the forgetting factor.

Finally, the prediction error interval, $xlim$, is defined based on a probability statement about the forecast error by adding or subtracting to the forecast data a multiple of the standard deviation of the forecast error:

$$xlim_k = \hat{x}_k \pm L\hat{\sigma}_{e,k} \tag{34}$$

where L is a proportional constant. Smaller values of L make the limits more restrictive while larger values lead to less restrictive limits.

In this chapter, the exponential smoothing models used in forecasting the data as well as used for estimation of the prediction error interval, were explained. The next chapter is dedicated to the explanation of the water quality monitoring system and its corresponding measured variables. Details of the three steps of the previously explained univariate data quality evaluation method will then be presented. An alternative method cannot be proposed unless a complete knowledge about the previous method is acquired.

3. MATERIALS AND METHODS

As discussed in section 2.3, the objective of this project is to propose a new model to fit to the fast dynamics of the time series that describe water quality parameters in the mon*EAU* automated water quality monitoring stations. General information about these stations is presented in section 2.2.1. The monitoring stations have been so far installed in different water sources with diverse purposes. The case study in this project is a station installed at the municipal treatment plant in Denmark. Time series related to one of the water quality parameters in this station will be selected and the new method will be calibrated and its performance will be evaluated according to that.

In this chapter, first the case study monitoring station will be introduced. The water quality parameter registered by this station, which will be used for tuning and evaluation of the new model, will be introduced. The chapter continues with an explanation of the different steps of the previously discussed univariate method. With a focus on the outlier detection step, the chapter ends with a discussion about the potential of substituting the forecasting model of the outlier detection system.

3.1. In situ monitoring station – Case study

As mentioned before, the automated water quality monitoring stations, mon*EAU*s, have been installed in different water systems with different purposes. The case study in this project used two automated monitoring stations (RSM30, Primodal System) which were installed one at the inlet and the other one at the outlet of a primary clarifier of the 700,000 PE municipal treatment plant in Lynetten (Copenhagen, Denmark) (Alferes et al., 2013b). Figure 2, illustrates one of the mentioned monitoring stations. The objective of installing these stations was to study the inflow dynamics as well as the performance of the primary clarifier. To achieve this objective, these stations comprise the following measurement instruments:

1. Instruments for measuring conventional physical-chemical parameters, such as temperature, pH, turbidity and conductivity.

2. A UV spectrometer for measuring total suspended solids (TSS), total chemical oxygen demand (CODt) and filtered chemical oxygen demand (CODf).

3. Ion selective electrodes for measuring ammonia, potassium and chloride.

Depending on the type of instrument, the sampling time of the data collected by the different measurement devices can be 5 or 60 seconds, which is very fast and generates a huge volume of information-rich data sets.

Among all water quality parameters registered in this system, turbidity was selected as the case study variable. Turbidity, Nephelometric Turbidity Unit (NTU), is known as the cloudiness of the fluid that is caused by suspended particles which are not visible by naked eye. In the monitoring stations this variable is measured by a Solitax (Hach, USA) sensor each 5 seconds. According to the experiments, turbidity data are known to be more sensitive to the variations in the status of the system in comparison with many other measured water quality parameters. The turbidity sensors are easily affected by clogging and fouling that may cause many errors in the data. That is why it was selected as the case study variable and according to that time series the new method will be tuned and its performance will be evaluated. We believe that any model that is successful in representing the real dynamics of this variable can work successfully for other variables. Figure 5 illustrates the case study turbidity sensor installed at the inlet of the primary clarifier of the Lynette wastewater treatment plant. This can be considered one of the most challenging locations for such sensor. In Figure 6 , the same sensor is shown after two weeks of utilization in sewage. This figure shows an example of the fouling problems that affect measurement in this system.

Figure 5 – The case study turbidity sensor installed at the primary clarifier of the Lynette wastewater treatment plant in Denmark

Figure 6 – The case study turbidity sensor – Sensor fouling problems

Figure 7 shows 45 days of raw turbidity data collected at the inlet of the wastewater treatment plant where probably the hardest measurement conditions for sensors can be found (Alferes et al., 2013a). Since the new method will be tuned according to this variable, the properties of this variable should be studied in more details. Figure 8 shows a 86 minute zoom.

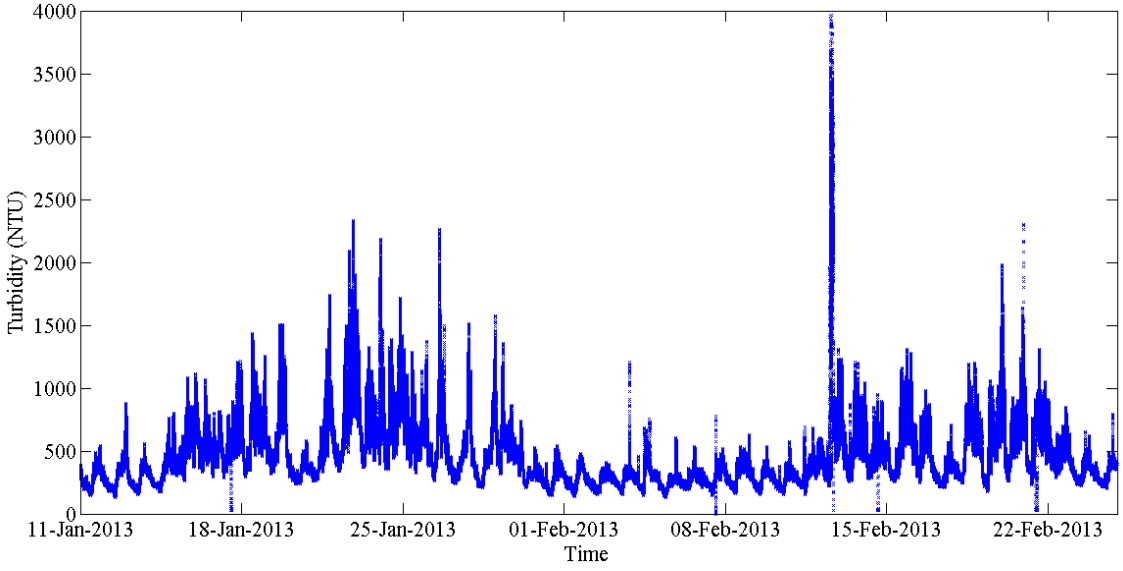

Figure 7 – Raw data – Turbidity – Ts = 5s

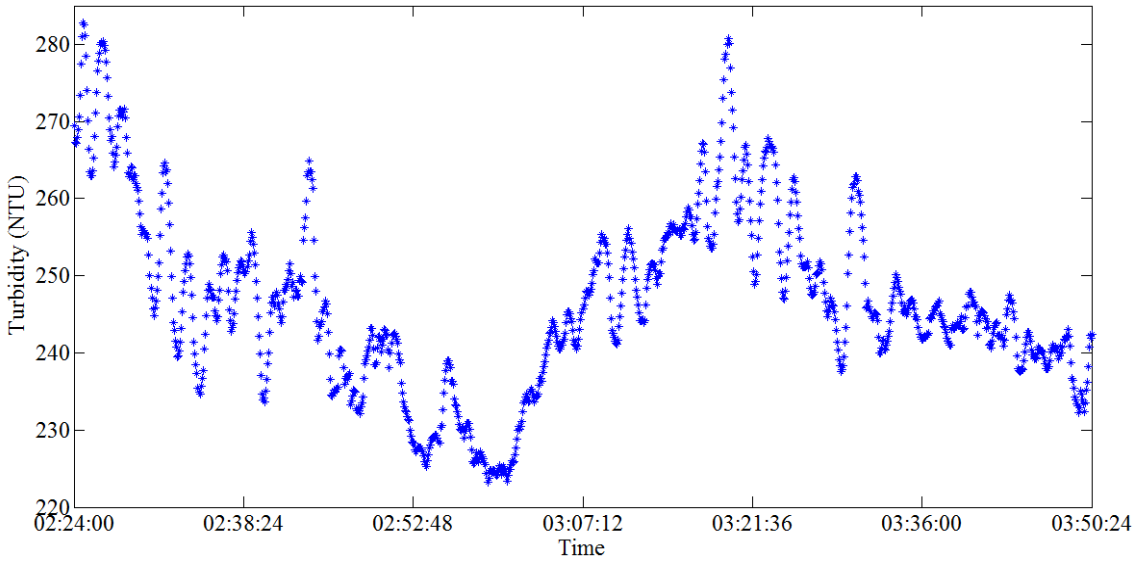

Figure 8 – Turbidity – Ts = 5s – A closer look

Taking a glance at the dynamics of the raw turbidity data in Figure 7 and Figure 8 shows that a 5 second sampling interval is relatively fast and may lead to intensive computations for data validation. A slower sampling interval, e.g. 1 minute, will not miss the important dynamics in the data while it may accelerate the mathematical calculations. In this study the raw turbidity data will thus be re-sampled at 1/12 the rate of the original data set, which leads to a 1 minute final sampling period. The result is shown in Figure 9.

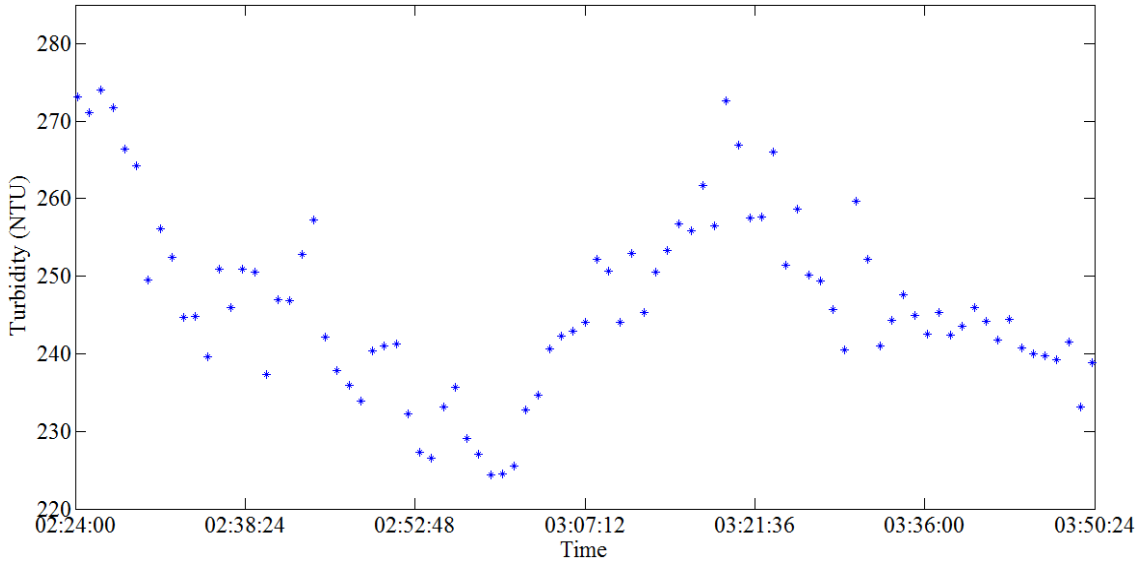

Figure 9 – Re-sampled turbidity – Ts = 1min – A closer look

Figure 10 shows a graph of rainfall data collected by the same monitoring system. Theoretically, it is expected that the rainfall intensity directly affects the level of contamination of water bodies by suspended particles and thus increases the level of turbidity. In other words, a positive correlation between precipitation and turbidity dynamics is expected. However, when the dynamics of the turbidity data in Figure 7 are regarded with consideration of the rainfall information of Figure 8, it can be observed that not all the large variations in the turbidity data correspond to rainfall events. However, as also previously discussed, due to the characteristics of water systems, such correlation does not necessarily exist in practice. Specifically for this case study system, different operation modes of the wastewater treatment system during the 24 hours of a day caused variations in turbidity data, which are not necessarily related to rainfall events. This behavior was the

main limitation in view of the automatic detection of abnormal behaviors in these types of data.

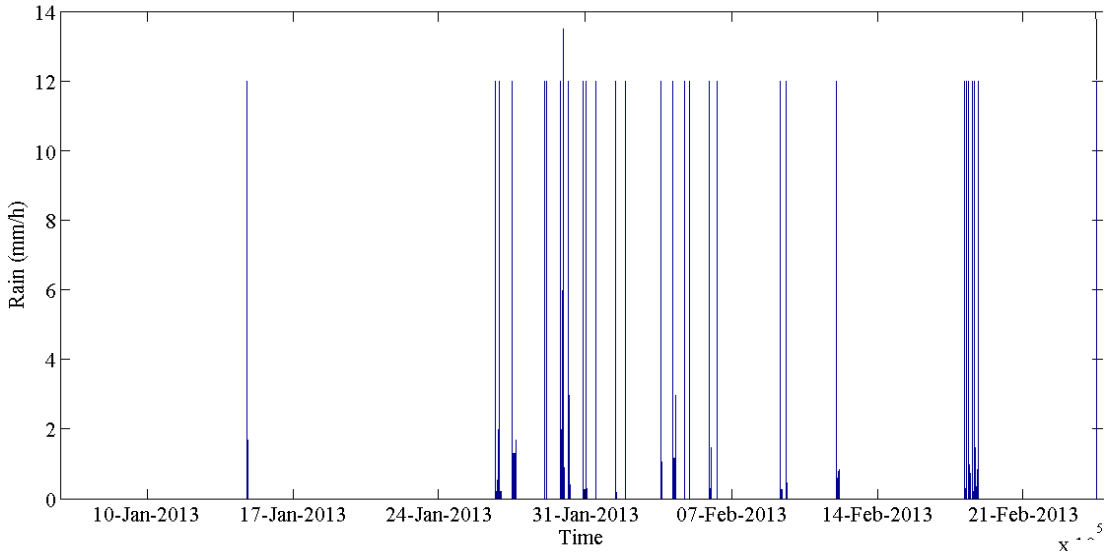

Figure 10 – Rainfall intensity data

A closer look of the turbidity data in Figure 11 shows that, harmonics of 1 hour and 24 hours can be observed. As mentioned before, there are different modes of operation in this wastewater treatment plant, such as pumping modes during the day, which may cause these hourly and daily harmonics to happen.

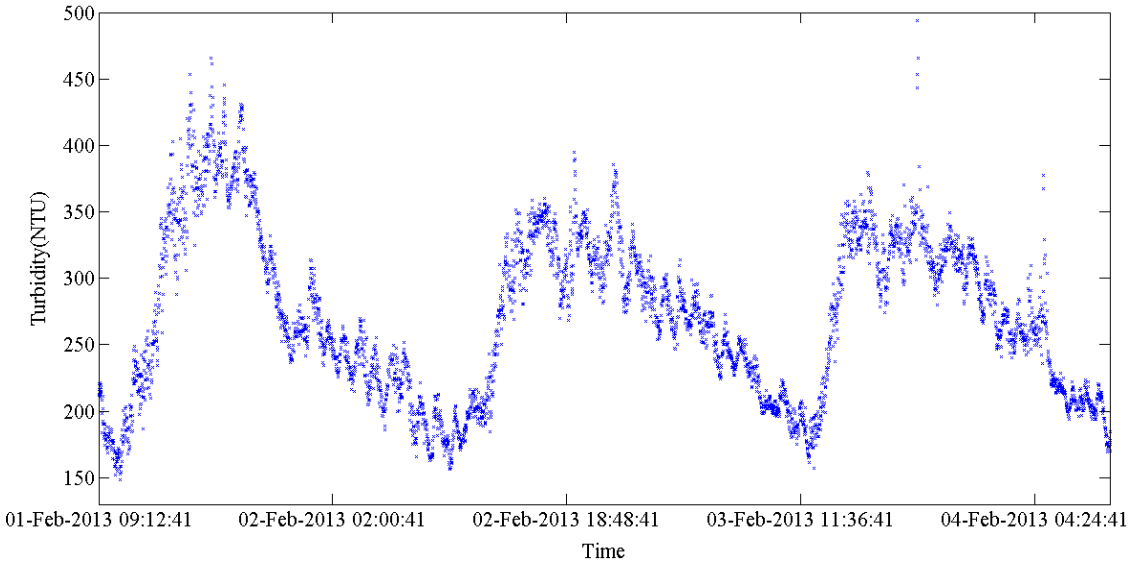

Figure 11 – Turbidity – Harmonics

This knowledge about the varying dynamics in the system could give an idea about the appropriate method that could be selected or the model that could be identified for the system.

To conclude, turbidity data collected using a sampling time of 1 minute will be employed to first select the proper structure of the new model and then to validate the performance of the method to detect and replace outliers. The results will then be compared with those obtained from the reference univariate method with the application of the exponential smoothing models to detect and replace outliers.

3.2. A univariate method for automatic data quality evaluation – Proposed by Alferes et al. (2012)

The univariate data validation method consists of three consecutive steps. The first step uses the algorithm to detect and replace outliers and to generate a proper time series which is smoothed in the next step to reduce the negative effects of noise. Potential sensor faults are then identified in the third step by calculating some data features according to the smoothed data and calculating their acceptability limits according to the raw data. Figure 12 schematizes the three steps of the univariate data quality evaluation method. In the next section, each of the mentioned steps will be explained theoretically and where necessary their mathematical equations will be provided.

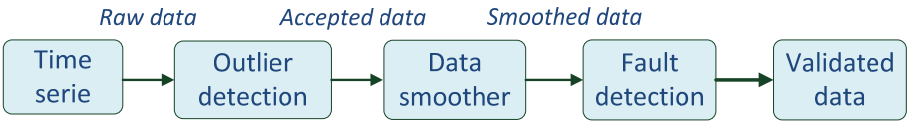

Figure 12 – Univariate method proposed by Alferes et al. (2012)

3.2.1. Outlier detection

As previously discussed, in this univariate method, an exponential smoothing model is fitted to the time series data where the unknown parameters are estimated based on the historical data. The model is then projected one step-ahead to obtain forecasts. To detect

outliers, an exponential smoothing model will be used as well to estimate the standard deviation of the forecast error, a coefficient of which will be used to decide on the acceptable prediction interval. Outliers can then be detected by comparing the observed data with the estimated prediction interval of the forecast value.

The basics of the exponential smoothing models and the way they generate forecasts are already explained in section 2.4. Exponential smoothing models are often selected to forecast time series data in immediate future due to their simplicity, their computational efficiency, the ease of adjusting their responsiveness to changes in the process and their adequate accuracy (Taylor, 2010). To be more precise, the third order exponential smoothing model which takes into account trends and seasonal changes, is chosen to forecast the time series data in the univariate method.

Considering the developments in section 2.4.3 for the 3rd order exponential smoothing model, equation (28) is used to give the one-step ahead forecast ($j = 1$), equations (30) are used to compute the coefficients of the model, $\hat{a}, \hat{b}$ and $\hat{c}$, using the first, second and third exponentially smoothed statistics $s_k, s_k^{[2]}$ and $s_k^{[3]}$, calculated according to equations (29).

The first step to start the algorithm is to know the initial values of the smoothing statistics, i.e. $s_1, s_1^{[2]}$ and $s_1^{[3]}$. These values are simply considered equal to x_1. Subsequent values of the smoothing statistics will be calculated at each time step using equations (29) by having α and the observed value x at each time step k.

To choose the smoothing constant α, the model should be calibrated using the so-called "Good data". A Good data set is a subset of the whole data that is selected by the expert using his/her previous knowledge of working with similar systems and refers to a set of data which does not include any important number of faults and represents modes of the normal behavior of the system. This normal behavior is also defined according to the knowledge of the expert about the system. Therefore, for all possible values that α can take from 0.01 to 1 (with an interval of 0.01) the RMSE between the observed values and the forecast values (which is calculated according to the same 3rd order exponential smoothing

method) will be calculated using the Good data time series. The α value that leads to the least RMSE will be selected as the desired value.

To sum up, the calculation of the forecast according to the 3rd order exponential smoothing method is done according to the following algorithm:

1. The initial values of $s_1, s_1^{[2]}$ and $s_1^{[3]}$ are simply taken equal to x_1

2. $\hat{a}_1, \hat{b}_1$ and $\hat{c}_1$ are calculated according to (30)

3. The one-step ahead forecast data made at time step 1, $\hat{x}_2$ is calculated according to (28)

4. For subsequent steps k, $s_k, s_k^{[2]}$ and $s_k^{[3]}$ are calculated according to equations (29) by having the observed value x_k

5. The model parameters $\hat{a}_k, \hat{b}_k$ and $\hat{c}_k$ are calculated according to (30)

6. The one-step forecast data $\hat{x}_{k+1}$ is calculated according to (28)

7. The steps 4 to 6 can be repeated until all data points have been treated

It should be noted that the algorithm mentioned above should be followed to find the best value for α according to the Good data time series before starting the main algorithm.

Once the one-step-ahead forecast value is calculated at each time step, outliers can be identified by comparing the real observed data with the prediction error interval. As discussed in section 2.4.4, at each time step the forecast data is calculated together with its confidence interval. Calculation of this confidence interval, also known as the prediction error interval, is done by analyzing the one-step-ahead forecast error calculated in equation (31).

To provide an estimation of the local variance, the Standard Deviation of Forecast Error at each time step k, $\hat{\sigma}_{e,k}$, is computed by equation (32) where the Mean Absolute Deviation,

$\hat{\Delta}_k$, is estimated according to the 1st order exponential smoothing method given in equation (33).

To select the η value in equation (33), a procedure similar to the one already explained for the selection of α in equations (29) and (30) should be followed. Accordingly, for a period of Good data, for different values of η from 0.01 to 1 (with an interval of 0.01) the RMSE between the forecast Mean Absolute Deviation, $\hat{\Delta}_k$ and the absolute value of the one-step ahead forecast error, $|e_k(1)|$ is calculated. The desired value of η is the one which corresponds to the least value of RMSE.

Finally, the prediction error interval xlim, is calculated by:

$$xlim_k = \hat{x}_k \pm L\hat{\sigma}_{e,k} \tag{35}$$

L is a proportional constant that is selected by trial and error with a Good data series during the calibration phase. Smaller values of L make the limits more restrictive while its larger values lead to less restrictive limits. This parameter may be adjusted manually according to the previous experience with similar systems so that the method does not reject the real dynamics as outliers. For different data sets the fine tuning of this parameter may be needed after the calibration phase.

To sum up, to detect outliers at each time step k, a forecast value and its prediction interval, made at time $k-1$, are available. If the observed data at time k, x_k exceeds the prediction interval, it is considered as an outlier and the data is replaced by the forecast value $\hat{x}_k$. This procedure continues until the data points are exhausted. The result of substituting detected outliers in time series data is a new data series which is called Accepted Data.

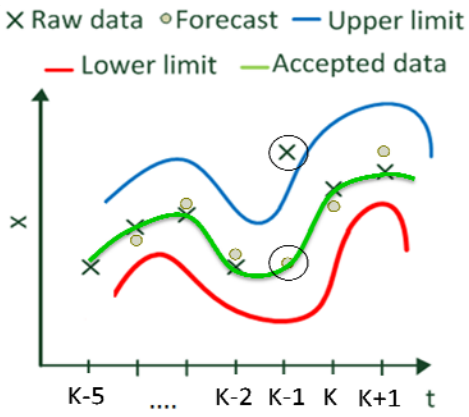

Figure 13 – Graphical representation of the outlier detection algorithm

It is important to note that the outlier detection step in the univariate method is equipped with an algorithm that consists of the backward reinitialisation of the algorithm in case the outlier detection procedure gets out of control. The algorithm is called out of control when a large number of consecutive data samples are rejected as outliers. In such situations, the algorithm reinitiates in the backward direction to recover the lost data and continues until a specific numbers of data points before the starting point of the out of control occurrence. Then the algorithm continues in the forward direction again. In this way data points that have caused the out of control situation can be skipped.

3.2.2. Data Smoothing

Subsequent to the detection and replacement of outliers (producing the accepted data) and before applying the fault detection methods, the treated data will pass through a smoother to smooth the data and reduce noise. This smoothing step is necessary since the noise present in the data will negatively affect the statistical data features that will be extracted in the fault detection block and hence will negatively affect the proper conclusion about system fault.

According to the univariate method proposed by Alferes et al. (2012), the accepted data will pass through a Kernel smoother to decrease the amount of noise by which the data is contaminated. According to the definition of Takahama and Sakai (2009), a kernel

smoother estimates a smooth function on noisy observations when no parametric model for this function is available. To estimate the smooth value Schimek (2013) proposes using the Nadarya-Watson kernel estimator to estimate the smooth function as a locally weighted average using a kernel as a weighting function. According to this estimator, at each point x_0 the estimation of the smooth value by using the neighboring points is:

$$\hat{y}_h(x_0) = \sum_{i=1}^{n} W(x_0, x_i; h).y(x_i) \tag{36}$$

where $\hat{y}_h(x_0)$ is the estimation of the smooth value of the observed point at x_0, n is the number of observed points, $W(x_0, x_i; h)$ is a weighting function, $y(x_i)$ are the observations at x_i points and h, the band width, is the number of neighboring points around x_0 that are considered to estimate the smooth value.

The weighting function is given by:

$$W(x_0, x_i; h) = \frac{K(\frac{x_0 - x_i}{h})}{\sum_{i=1}^{n} K(\frac{x_0 - x_i}{h})} \tag{37}$$

where $K(.)$ is a kernel function.

Equation (36) means that the kernel smoother estimator is just a weighted sum of the observations $y(x_i)$. In other words, when a data series is smoothed by a kernel smoother, all the points of the time series are weighted using the weights which are the results of the computation of the kernel function.

Kernel functions are normally symmetric with non negative values that decrease from a central (maximum) value to zero. In literature one finds symmetric probability density functions used as kernel functions. They can take different forms like Uniform, Triangular, Quadratic, Gaussian or Normal (Takahama & Sakai, 2009). The kernel smoother used in the univariate data quality evaluation method proposed by Alferes et al. (2012) employs a Gaussian kernel function with the following form:

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{(-\frac{x^2}{2})} \tag{38}$$

in which $x = \frac{x_0 - x_i}{h}$

h is a positive parameter which controls the smoothness or roughness of the estimates and leads to under-smoothing or over-smoothing situations. For very small values of h the variance of the estimates will be too large while for very large values of h the bias of the estimates will be too large. There are various criteria according to which h can be selected. Normally it is selected automatically in a way that it minimizes the sum of squared residuals between the raw data and the smoothed data (Schimek, 2013). In the univariate method there is the possibility of selecting h automatically to minimize the sum of squared residuals or manually according to the further applications of the data. For example, when the data is going to be used for modeling purposes, if the simplicity of the model is desired, the larger values of h may be required.

The principal theories about the kernel smoother were just presented. The accepted data passes through a kernel smoother that smoothes the data and reduces the noise. According to this smoothed data, some statistical data features will be calculated in next step which help detection of probable faults in the data. These data features will be defined in next section and their contribution in detecting potential sensor faults will be discussed.

3.2.3. Fault detection

At each time step k, Alferes et al. (2012) propose a number of statistical data features to be calculated from the smoothed data sets. For each of these data features, the acceptability limits are calculated based on the raw data time series. Probable faults in the system can be detected by comparing each of these data features with reference to their acceptability limits according to which a proper validation mark can be assigned to each data point. Table 1 summarizes the definitions of the data features calculated for fault detection purposes according to Alferes et al. (2013a)

Table 1 – Data features for fault detection

Data feature	Definition	Purpose of calculation
Percentage of replaced data	Fraction of forecast values that have replaced the outliers in the raw data	Demonstrates the situation in which the abnormal behavior of data is replaced with the forecast value and evaluates the goodness of the smoothed data and data features
Rate of change	Slope between two consecutive data points in smoothed data	Gives information about the dynamics of the data and helps detection of too sudden changes
Physical realistic range	Range where values are usually observed in a given measurement site with the corresponding sensor	Investigates if the data lies inside the expected range
Residual's standard deviation	Standard deviation of residuals (difference between accepted data and smoothed data)	Estimation of variance of the data, large standard deviation of residuals can be a sign of faulty data
Auto-correlation of residuals	Applying the Run Test (Dochain & Vanrolleghem, 2001) to the residuals (difference between accepted and smoothed data) to detect autocorrelation among them	Evaluates whether residuals are randomly distributed. If the run test fails, residuals are auto-correlated and if so, either the smoothed data is not representative of the accepted data or the noise is not randomly distributed

Acceptability limits for the data features in Table 1 are defined based on realistic values obtained in the field and can be adjusted to make the fault detection phase more or less restrictive. They are calculated by using raw data and can vary according to the measured variables, measurement sensors and measurement locations (Alferes et al., 2013a).

After calculating the data features together with their acceptability limits, data points are validated by regarding the data features with reference to their acceptability limits. For any specific period of data, if all the data features lie inside their acceptability limits, the validation mark 0 is assigned to the data which means the data is "Valid". If some of the tests fail for a specific period of data, the validation mark 1 is given to the data, which means the data is "Doubtful" and further analysis is necessary to judge about the validity of the data. Finally, if all the tests or the most important ones fail for a data period, the validation mark 2 is assigned to that period which means the data is "Not valid".

3.2.4. Discussion

Three consecutive steps of the univariate data quality evaluation method were presented and detailed to 1) detect and replace outliers, 2) remove noise and 3) detect probable faults in time series data collected by mon*EAU* automated water quality monitoring stations.

It is believed that the most drastic step in this univariate method is the first step, which is the detection and replacement of outliers. The reason behind this claim concerns in the first place the intrinsic characteristics of the system under study. The outlier detection is based on forecasting the future behavior of the system according to a model of the system. However, as also discussed in section 2.2, finding an exact model that describes all physical and chemical phenomena in water systems is difficult. As Alferes et al. (2013a) have proposed, one alternative is the identification of the system model according to historical data. This approach also has its own challenges since the system under study has very fast dynamics and is prone to significant noise and many uncertainties.

Practical results presented by Alferes et al. (2012; 2013b) demonstrate that the application of the 3rd order exponential smoothing model, which is selected as the model of the water

system in their univariate approach, gives very good performance. However, theoretically it can be shown that this model has drawbacks that made us decide about a more systematic approach to identify a new type of model for the system. For instance, as discussed in 2.4.3, according to this method it is assumed that the water quality parameter time series follow a quadratic model which is presented in equation (26), while the system does not necessarily behave like a quadratic system. That is why the next chapter is dedicated to revise the exponential smoothing models, with a different theoretical point of view. This revision of the model includes a mathematical reorganization of the equations presented in section 2.4 that will help us better understand the actual function of the models.

With this improved understanding, in the next chapter the idea of presenting an alternative model will be discussed and hence the procedure related to this new approach will be developed. At the end of the chapter, the results of applying the new model to the univariate method will be presented in comparison with the exponential smoothing model.

4. RESULTS

This chapter includes three principal sections. First the exponential smoothing models in modeling and forecasting the behavior of the water quality monitoring system will be theoretically revisited. Then, the proposed approach and the underlying theories will be presented as an alternative to the third order exponential smoothing model. The section continues by presenting the steps followed to tune the new model and identify its final structure. Finally, the performance of the univariate method with reference to the specific criteria will be evaluated for the new model and compared with the exponential smoothing model.

4.1. Exponential Smoothing Models – Revisited

The purpose of this section is to discuss the essence of exponential smoothing models through a different theoretical approach. It will help to better comprehend the intrinsic performance of the smoothing methods in forecasting time series data.

4.1.1. 1st order exponential smoothing model

If equation (5) in section 2.4.1 is written in the form of an input-output transfer function model in the discrete-time domain, we obtain:

$$s_k = \frac{\alpha}{1-(1-\alpha)z^{-1}} \cdot x_k \tag{39}$$

in which z^{-1} is the backward-shift operator and $s_k z^{-1}$ represents a one-step delay in s_k Equation (39) expresses that the first order exponential smoothing operator is actually a first order low-pass filter.

On the other hand, according to equations (9) and (10), the forecast data for any time in the future, $\hat{x}_{k+\tau}$, is in fact equal to the value of the filtered data at time k, s_k, which is held

constant for any time in future. The performance of the explained behavior is demonstrated graphically in Figure 14. It should be noted that the values in this figure are demonstrative and they do not necessarily correspond to a real system.

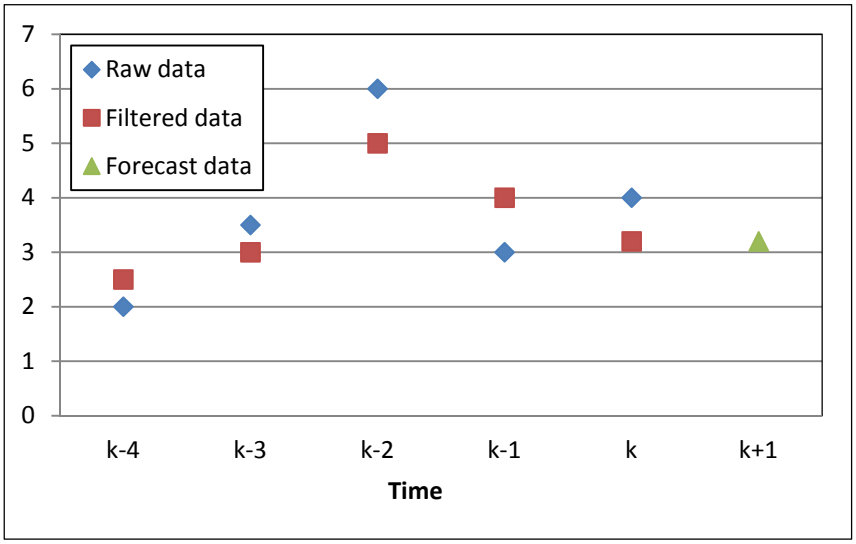

Figure 14 – 1st order exponential smoothing – Forecasting

4.1.2. 2nd order exponential smoothing model

A similar approach to the 1st order exponential smoothing model can be followed to explain the forecasting behavior of the second order exponential smoothing model. If the equations (5) and (16) in sections 2.4.1 and 2.4.2 are rewritten as an input-output transfer function model in the discrete-time domain, we get:

$$s_k^{[2]} = \left(\frac{\alpha}{1-(1-\alpha)z^{-1}}\right)^2 \cdot x_k \tag{40}$$

According to the above equation, the 2nd order exponential smoothing operator can be regarded as the combination of two identical low-pas filters in series that take x_k as the input and give $s_k^{[2]}$ as the output.

To develop a new approach regarding the forecasting behavior of the 2nd order exponential smoothing operator, the previously stated equations in section 2.4.2 shall be reconsidered. It can be demonstrated that the forecasting equation (25) can be expressed as a function of

the simple and double smoothed statistics s_k and $s_k^{[2]}$. To address this, the procedure mentioned below should be followed.

According to the current origin of time k, the forecasting statement for one step ahead is:

$$\hat{x}_{k+1} = \hat{a}_k + \hat{b}_k \tag{41}$$

in which the unknown parameters $\hat{a}_k$ and $\hat{b}_k$ can be estimated according to equations (19) and (24). If s_k and $s_k^{[2]}$ in equations (19) and (24) are substituted by their equivalents of equations (5) and (16), after some simplifications the new expressions of $\hat{a}_k$ and $\hat{b}_k$ can be replaced in equation (41), yielding the one-step ahead forecast:

$$\hat{x}_{k+1} = 2s_k - s_{k-1}^{[2]} \tag{42}$$

From equation (42) it can be inferred that the one step-ahead forecast data calculated at time k, simply follows a linear function of the simple smoothed statistic value at time k, s_k and the double smoothed statistic value at time $k-1$, $s_{k-1}^{[2]}$. In other words, it can be simply proven that $\hat{x}_{k+1}$ lies on the straight line that passes through s_k and $s_{k-1}^{[2]}$. The performance of the explained behavior is demonstrated graphically in Figure 15.

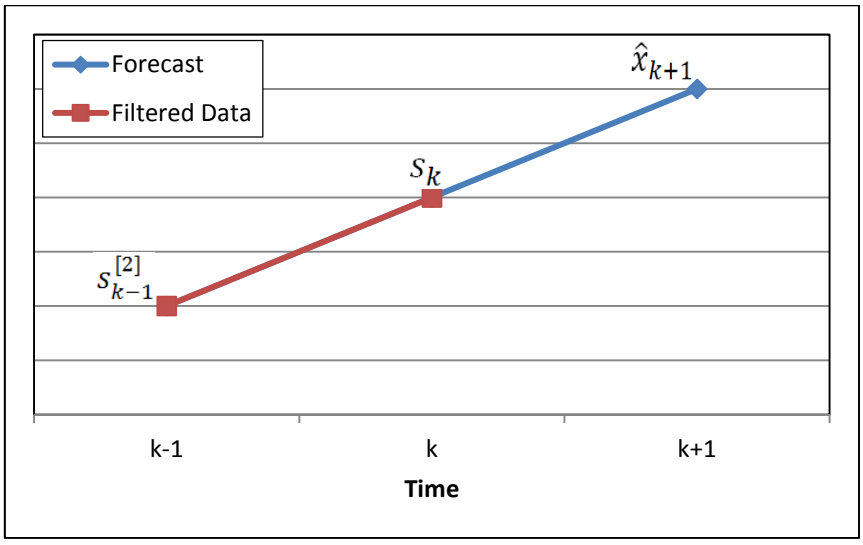

Figure 15 – 2nd order exponential smoothing – Forecasting

To prove this, the general equation of a straight line should be considered. For a straight line we have:

$$f(k) = ak + b \tag{43}$$

We declare that:

$$\begin{aligned} f(k-1) &= s_{k-1}^{[2]} = a(k-1) + b \\ f(k) &= s_k = ak + b \end{aligned} \tag{44}$$

Equations (44) can be re-organized with respect to a and b which are the unknown parameters of the straight line in equation (43). The result will be:

$$\begin{aligned} a &= s_k - s_{k-1}^{[2]} \\ b &= (1-k)s_k + ks_{k-1}^{[2]} \end{aligned} \tag{45}$$

If a and b in (43) are substituted by their equivalents from equation (45), a unique straight line results that passes through s_k and $s_{k-1}^{[2]}$ with the following equation:

$$f(k) = (s_k - s_{k-1}^{[2]})k + (1-k)s_k + ks_{k-1}^{[2]} \tag{46}$$

The value of $f(k+1)$ will be:

$$f(k+1) = 2s_k - s_{k-1}^{[2]} \tag{47}$$

While, according to equation (42), we know that:

$$\hat{x}_{k+1} = 2s_k - s_{k-1}^{[2]} \tag{48}$$

So:

$$f(k+1) = \hat{x}_{k+1} = 2s_k - s_{k-1}^{[2]} \tag{49}$$

Therefore it is proven that the one step ahead forecast data made by a 2nd order exponential smoothing operator lie on the straight line that passes through s_k and $s_{k-1}^{[2]}$.

4.1.3. 3rd order exponential smoothing model

A similar approach can be followed for a 3rd order exponential smoothing model and the results mentioned below can be developed.

The 3rd smoothing statistic, $s_k^{[3]}$, can be regarded to be the result of filtering the data string x_T three times:

$$s_k^{[3]} = \left(\frac{\alpha}{1-(1-\alpha)z^{-1}}\right)^3 \cdot x_k \tag{50}$$

Similar to the 1st and 2nd order exponential smoothing models, a new mathematical relationship can be developed for a 3rd order exponential smoothing operator to produce the forecasts.

According to equation (28) in section 2.4.3, the one-step ahead forecast given by the 3rd order exponential smoothing model is:

$$\hat{x}_{k+1} = \hat{a}_k + \hat{b}_k + \frac{1}{2}\hat{c}_k \tag{51}$$

in which the unknown coefficients of the model, $\hat{a}_k, \hat{b}_k$ and $\hat{c}_k$, are computed by equations (30).

By considering the equations in (29), the unknown coefficients of the model, $\hat{a}_k, \hat{b}_k$ and $\hat{c}_k$, can be rewritten as functions of $s_k, s_{k-1}^{[2]}$ and $s_{k-2}^{[3]}$ as:

$$\begin{aligned}
\hat{a}_k &= (\alpha^2 - 3\alpha + 3)s_k + (-2\alpha^2 + 5\alpha - 3)s_{k-1}^{[2]} + (\alpha^2 - 2\alpha + 1)s_{k-2}^{[3]} \\
\hat{b}_k &= -\frac{3}{2}(\alpha^2 - 2\alpha)s_k + (3\alpha^2 - 5\alpha)s_{k-1}^{[2]} + \frac{1}{2}(-3\alpha^2 + 4\alpha)s_{k-2}^{[3]} \\
\hat{c}_k &= \alpha^2 s_k - 2\alpha^2 s_{k-1}^{[2]} + \alpha^2 s_{k-2}^{[3]}
\end{aligned} \tag{52}$$

Substitution of equations (52) in (51) and the consecutive simplifications lead to the following relationship:

$$\hat{x}_{k+1} = 3s_k - 3s_{k-1}^{[2]} + s_{k-2}^{[3]} \tag{53}$$

By having a priori information about the second order exponential smoothing model, it can be inferred that for a third order exponential smoothing model, a quadratic relationship exists among the three smoothing statistics as demonstrated in Figure 16.

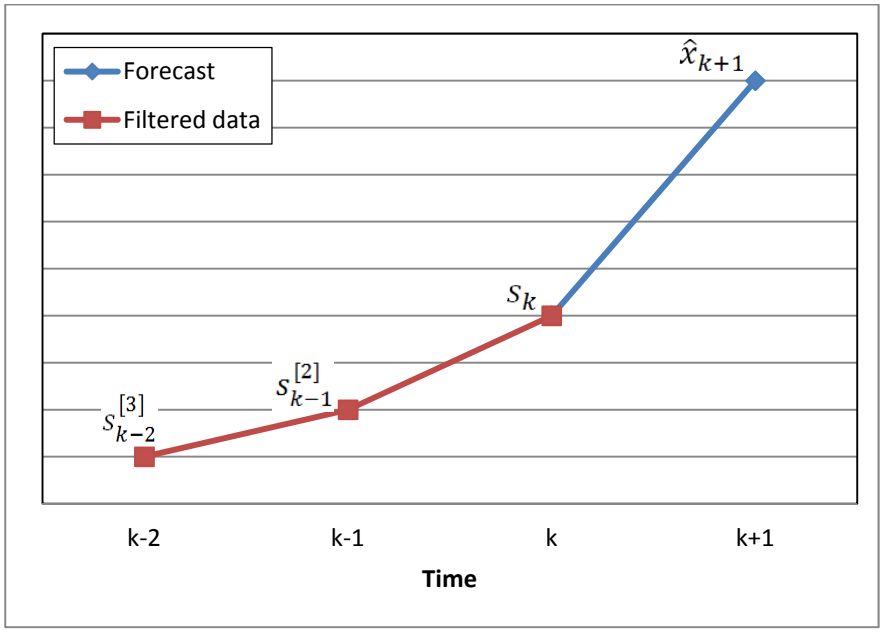

Figure 16 – 3rd order exponential smoothing – Forecasting

To prove this, the general equation of a quadratic function should be considered:

$$f(k) = a + bk + \frac{1}{2}ck^2 \tag{54}$$

We declare that:

$$\begin{aligned} f(k-2) &= s_{k-2}^{[3]} \\ f(k-1) &= s_{k-1}^{[2]} \\ f(k) &= s_k \\ f(k+1) &= \hat{x}_{k+1} \end{aligned} \tag{55}$$

In the quadratic function of equation (54), estimation of the three unknown parameters, a, b and c, will be possible if at least three points of the function are available; Thus, the quadratic model will be then uniquely identifiable. Based on (55), the three points will be:

$$\begin{aligned} f(k-2) &= s_{k-2}^{[3]} = a + b(k-2) + \frac{1}{2}c(k-2)^2 \\ f(k-1) &= s_{k-1}^{[2]} = a + b(k-1) + \frac{1}{2}c(k-1)^2 \\ f(k) &= s_k = a + bk + \frac{1}{2}ck^2 \end{aligned} \tag{56}$$

Equations (56) can be re-organized and the unknown parameters a, b and c can be stated as functions of $k, s_k, s_{k-1}^{[2]}$ and $s_{k-2}^{[3]}$.

On the other hand, we know that:

$$f(k+1) = \hat{x}_{k+1} = a + b(k+1) + \frac{1}{2}c(k+1)^2 \tag{57}$$

If the calculated parameters a, b and c, are substituted in equation (57) it leads to:

$$f(k+1) = 3s_k - 3s_{k-1}^{[2]} + s_{k-2}^{[3]} \tag{58}$$

Equation (58) is equal to the equation (53) for $\hat{x}_{k+1}$.

The presented development proves that $\hat{x}_{k+1}$ lies on the quadratic relation that exists among $s_k, s_{k-1}^{[2]}$ and $s_{k-2}^{[3]}$.

4.1.4. Discussion

According the above discussions, it can be concluded that using the exponential smoothing approach consists of two parts, filtering and prediction. The data is filtered by the low-pass filters (once, twice and three times) and then the predictions are made according to the filtered data.

The purpose of filtering the data with an ideal low-pass filter is to remove noise while the informative part of the data does not change. The cut-off frequency in the filters must be selected according to the frequency contents that are meant to be eliminated. However, in the exponential smoothing approach, selection of the order of the filter and the cut-off frequency is done empirically, without detailed analysis of the system. Specifically, for the 3rd order exponential smoothing method, filtering the data repeatedly with the same time constant cannot be logically explained from a cut-off frequency point of view.

On the other hand, prediction of the behavior of the system should be done according the available information about that system or according the model of that system. The limitation of the exponential smoothing approaches is in the way they address the models they fit to the system and the way these models produce forecasts. For example, the first order exponential smoothing model fits a constant model to the data. The last estimated value by this model is kept constant and projected into the future to produce forecasts. However, as will be demonstrated later, this is in fact an integrator model that keeps the last value constant as forecast for the future. On the other hand, the 3rd order exponential smoothing model fits a quadratic model to the history of the data and the same quadratic model continues into the future to produce forecasts. However, it seems that this model was selected empirically for the system. Water quality dynamics can be very fast, auto-correlated and noisy but the choice of the quadratic model, considering the way it generates the forecasts, is not systematic.

With respect to the previous discussions, it is attempted in this work to replace the model of the system and the forecasting approach. As mentioned before, this is the main objective of this research project. In section 2.2, the challenging conditions of the data quality evaluation in water quality monitoring stations were discussed and the desired characteristics of the alternative approach were enumerated. Accordingly, the idea is to find an alternative approach that respects the following characteristics:

- It should be automatic and online applicable so that it can be applied to the automated water quality monitoring systems.

- It should be fast to get along with the fast dynamics of the system under study.

- It should be selected systematically according to the characteristics of the system.

- It should not be computationally complicated and costly.

In the next section the alternative approach, that respects the desired characteristics, will be proposed and the corresponding theories will be presented.

4.2. Forecasting Time Series – An Alternative Method

As discussed in section 4.1.4, the 3rd order exponential smoothing model has some limitations. The results with the univariate method show that the 3rd order exponential smoothing model can be successful in detecting the outliers in a time series (Alferes et al., 2012); (Alferes et al., 2013a); (Alferes et al., 2013b). However, we propose using a more systematic approach for the system.

In this chapter, the mathematical background of the proposed model and the forecasting expression are developed. A procedure to find a more representative structure of the model will be presented in further steps.

4.2.1. Theoretical background

4.2.1.1. Auto-Regressive Moving-Average (ARMA) model

As discussed before, due to the specific characteristics of the water quality monitoring stations, dynamic of the data collected by these systems can be affected by numerous factors. For instance, variations in the turbidity parameter can be affected by precipitation, the discharges from the water sources, defects in measurement devices, measurement and system noise and a few other factors that are not precisely known. As proposed by Berthouex and Box (1996), a reasonable approach to describe such systems can be the consideration of a stochastic behavior for the system and trying to fit an auto-regressive moving average model to it. In these models, the term auto-regressive relates the current

value of the time series to its own previous values and the term moving-average relates the current value of the series to the current or past white noise error terms (Berthouex & Box, 1996).

Application of the auto-regressive class data-driven models has been successful in empirical modeling of time series data in different papers (Hipel & Mcleod, 1978; Hau & Tong, 1989; Berthouex & Box, 1996; Garcia et al., 2010). According Berthouex & Box (1996), an Auto-Regressive Integrated Moving Average (ARIMA) model can be fitted empirically to the time series obtained from a wastewater treatment plant. They state that the stochastic models that describe a time series can be employed to give optimal forecasts of future values of that time series. Therefore, the ARIMA class of models is selected to be tested for this system. The mathematical theories related to these models are developed according to Ljung (1987) in the next section.

Mathematical model and forecasting equation

The discrete-time input-output model of a linear ARMA model takes the form:

$$y(k) = \frac{C(z^{-1})}{A(z^{-1})} e(k) = H(z^{-1}) e(k) \tag{59}$$

where $e(k)$ is the white noise signal and $C(z^{-1})$ and $A(z^{-1})$ are the polynomials from orders n_c and n_a respectively (denoted as ARMA$[n_a, n_c]$) and defined by:

$$\begin{aligned} A(z^{-1}) &= 1 + a_1 z^{-1} + a_2 z^{-2} + \ldots + a_{na} z^{-na} \\ C(z^{-1}) &= 1 + c_1 z^{-1} + c_2 z^{-2} + \ldots + c_{nc} z^{-nc} \end{aligned} \tag{60}$$

where z^{-1} is the backward-shift operator.

In the data-driven proposed ARMA model, the unknown parameters of the model, $[a_1\ a_2\ \ldots\ a_{na}]$ and $[c_1\ c_2\ \ldots\ c_{nc}]$, will be identified according to the history of the data and based on the minimization of the variance of the prediction error.

The prediction error is defined as:

$$\varepsilon(k+j) = y(k+j) - \hat{y}(k+j \mid k) \tag{61}$$

where the term $\hat{y}(k+j|k)$ is the optimal prediction made for time $k+j$ according to the available data until time k, calculated as:

$$\hat{y}(k+j \mid k) = \frac{F_j}{C(z^{-1})} y(k) \tag{62}$$

in which, F_j is calculated according to the following Diophantine equation:

$$\frac{C(z^{-1})}{A(z^{-1})} = Quotient + Remainder = E_j + q^{-j} \frac{F_j}{A(z^{-1})} \tag{63}$$

for which E_j has degree $j-1$ and j is the prediction interval.

For the special case that only a one-step ahead forecast data needs to be calculated, j will be equal to 1. To calculate F_j in this case, the long division of $\frac{C(z^{-1})}{A(z^{-1})}$ will be carried out once to have E_j of order zero.

A challenge in this approach is the selection of the order of the model. There are different criteria to identify the order of the model, such as Akaike's information criterion (Akaike, 1974), the final prediction error and the minimum description length suggested by Rissanen (1978). Indeed, according to the principle of parsimony, the model should possess as few parameters as possible while it adequately represents the data (Berthouex & Box, 1996). In this research project, by considering the principal of parsimony, different orders of the model will be tried and their fit to the system will be verified using the RMSE criterion.

4.2.1.2. Forcing integrator to the ARMA model – ARIMA model

In order to represent a non-stationary time series, which is the case of many environmental applications, a useful solution is to add an integrator to the ARMA model presented in (59) (Berthouex & Box, 1996) to have an ARIMA model.

Mathematical model and forecasting equation

According to Ljung (1987), the ARIMA model transfer function takes the following form:

$$y(k) = \frac{C(z^{-1})}{A(z^{-1})(1-z^{-1})} e(k) = \frac{C(z^{-1})}{A(z^{-1})\Delta} e(k) \tag{64}$$

where the term $\Delta = (1 - z^{-1})$ represents the differencing operator, which forces the model to possess a pole on the unitary circle in z-plane.

To calculate the j-step ahead forecasts, a similar procedure to that presented for the ARMA model is followed. The only difference is that the denominator possesses an integrator in addition to $A(z^{-1})$.

4.2.1.3. Pure integrator model

If we tend to model a system by a pure integrator, in the ARIMA model we have $A(z^{-1}) = C(z^{-1}) = 1$, so the transfer function in the discrete-time will be:

$$y(k) = \frac{1}{1-z^{-1}} e(k) = \frac{1}{\Delta} e(k) \tag{65}$$

A pure integrator model produces forecasts according to the equation below:

$$\hat{y}(k+j \mid k) = y(k) \tag{66}$$

4.2.1.4. Prediction theory for an ARMA model

To demonstrate how an ARMA model performs the forecasting of the unseen future in a stochastic system, the following simple system can be considered:

$$y(k) = \frac{1}{1 - az^{-1}} e(k) \tag{67}$$

where $e(k)$ is a white noise signal.

According to equation (62), the forecasting expression for j step-ahead will be:

$$\hat{y}(k + j \mid k) = a^j y(k) \tag{68}$$

The forecasting behavior of the model is simulated for three different values of $a = 0, 0.5$ and 1, $k = 1, 2, \ldots, 90$ and $j = 10$.

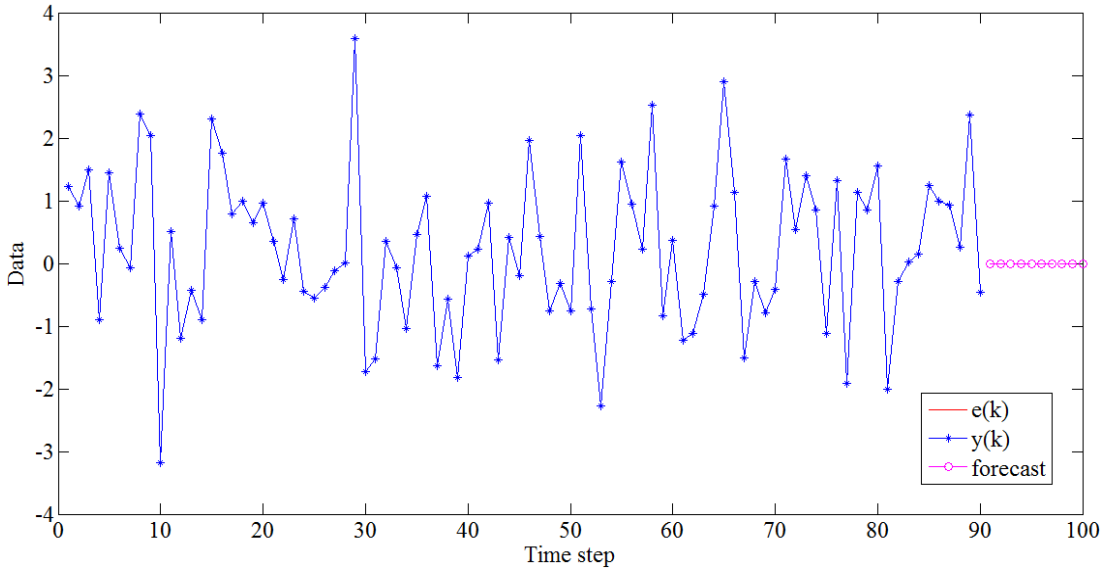

Figure 17 – Forecasting behavior for a = 0

As demonstrated in Figure 17, for $a = 0$, $y(k) = e(k)$, the forecasts are equal to zero.

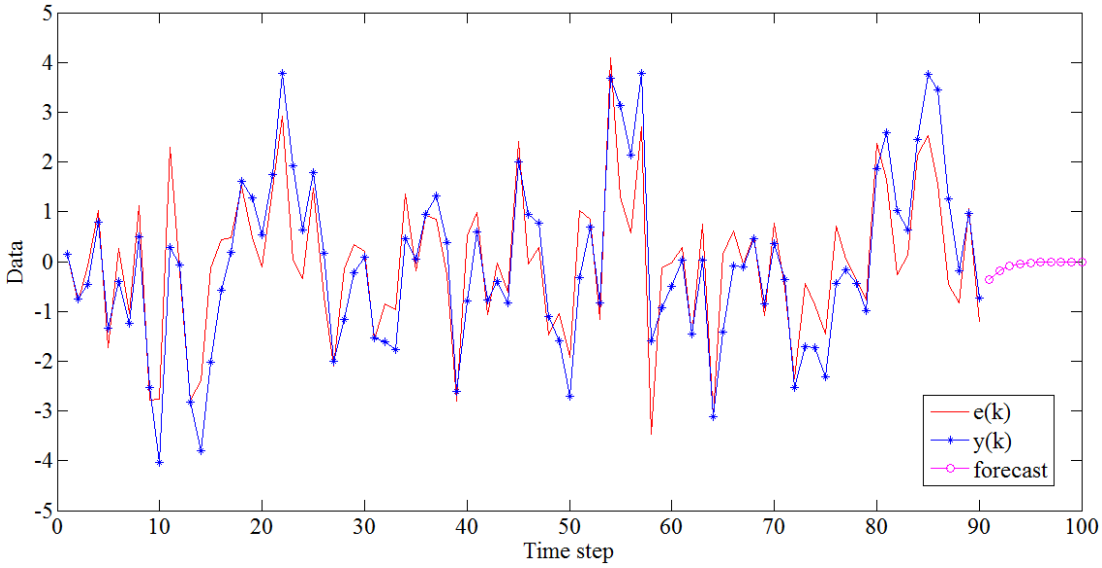

Figure 18 – Forecasting behavior for a = 0.5

Figure 18 shows the forecasts for $a = 0.5$. As expected from equation (68), for all the values of $0 < a < 1$, the forecasts tend to zero exponentially and the rate of this tendency to zero depends on the magnitude of a. As a gets closer to zero, the forecasts tend faster to zero.

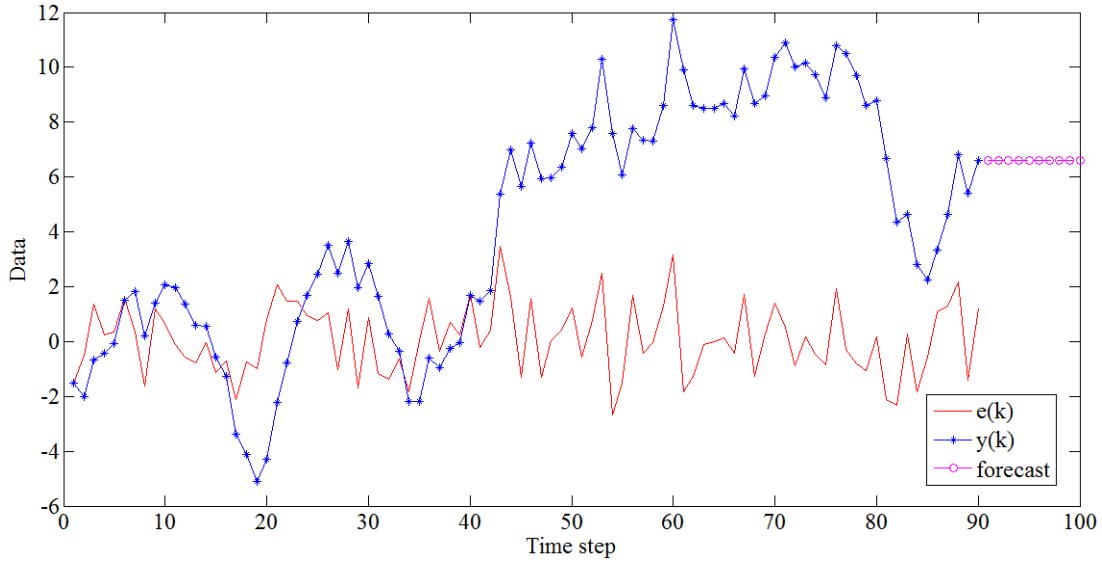

Figure 19 – Forecasting behavior for a = 1

For a pure integrator system, as demonstrated in Figure 19 and also discussed in section 4.2.1.3, the forecasts for the whole prediction interval will be equal to the latest value of the data.

4.2.1.5. Moving window data approach

In order to follow the varying dynamics in the data, according to the idea presented by Liu et al. (2004) for outlier detection, we propose using the moving window approach. Accordingly, a limited number of consecutive data samples, called a window of data, will be considered and it will move along the whole length of the data to catch the dynamics.

The moving window approach used with an ARMA-class model is principally based on the identification of the model for the window of data, producing forecast data according to the identified model and moving the window one step ahead while discarding the oldest data point from the window. Figure 20 demonstrates the performance of the moving window approach for a Window Size (WS) of 4, starting at time $k-4$.

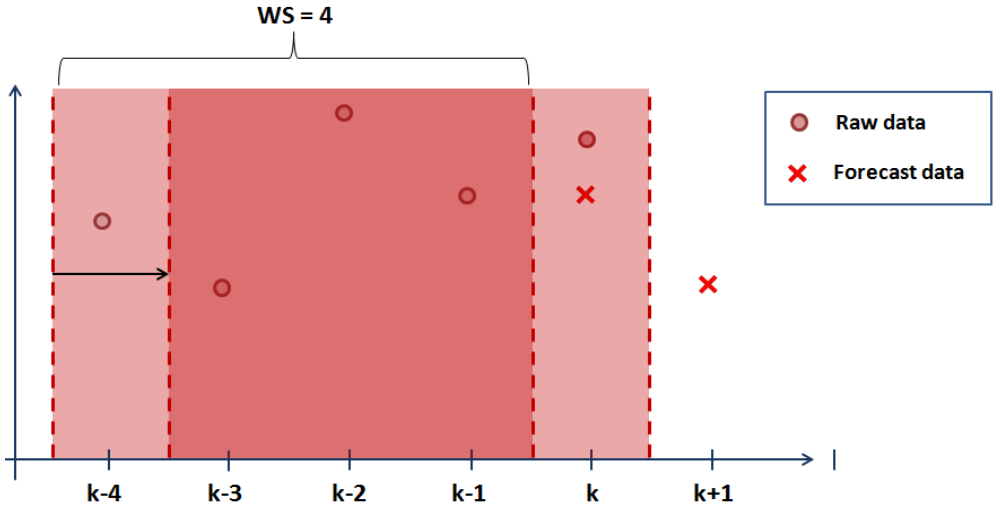

Figure 20 – Moving Window approach

In general, if N is the size of the moving window, i.e. the number of data samples included in the moving window data for each iteration of the algorithm, the proposed algorithm to fit

an ARMA-class model to a moving window data and calculate a j step-ahead forecast data is proposed below:

1- N number of data points are considered.

2- The parameters of the ARMA-class model are identified according to this window of data.

3- Based on the identified model, the j step-ahead forecast data is calculated.

4- The data window moves one data point further. Thus, a new data point is considered in the window and the oldest data point is ejected from the window.

5- Steps 2 to 4 are repeated consecutively until the last data point is reached in the window.

It should be noted that the moving window approach is actually equivalent to working with the on-line data in real-time. To be more clear, at the initialization of the algorithm one should wait until at least N data points are available. According to this period of data, the above mentioned algorithm can be followed to produce the j step-ahead forecast. The procedure continues until the end of the data collection period.

Window size selection

Regarding the application of the moving window approach, one challenge is to select the window size. Specifically for systems with special dynamics, such as those of water quality monitoring stations, selection of a proper window size is of great importance, in the sense that, for an extremely large window size, the important dynamics will be averaged and missed. On the other hand, by selection of a small window size, the algorithm will be more sensitive to noise and the model will adapt to the noise instead of the real dynamics.

In section 3.1, the case study monitoring stations and the water quality parameter were introduced. Our knowledge about the dynamics of the data and the existing harmonics give us an idea about the proper window size. However, to select a value that gives the best fit to the dynamics of the system, different window sizes should be tested and their fits to the system should be verified.

4.2.2. Model structure and window size selection

In this section, the procedure which is followed to identify the proper model, the structure and the window size will be presented.

In section 4.2.1, the models ARMA, ARIMA and the pure integrator were presented as proposed alternatives for the system. ARMA and ARIMA models can possess different numbers of poles and zeros and thus the order of the model is another effective parameter in the fit of the model to the system. As discussed before, the window size of the data is another factor that affects the quality of the approach in forecasting the data one step-ahead. Therefore, the performance of different structures in forecasting the one-step ahead data should be evaluated. To compare the accuracy of the forecasts, the RMSE criterion is calculated for the different cases.

The RMSE is a criterion which is frequently used to evaluate the quality of a model or estimator to forecast future values. According to Berthouex & Box (1996), RMSE is an estimate of the variance of the one step-ahead forecast. A model with a small RMSE leads to better forecasts than a model with large RMSE. According to the definition, RMSE is the root of the average square of residuals of the forecasts and can be formulated as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} e_i^2} \tag{69}$$

for which N is the total number of data samples and e_i is the residual, the difference between forecasts and observed values.

The ARIMA model structure and the window size will be selected according to a Good period of turbidity data, demonstrated in Figure 21, which includes a rain event on 2013.02.19. The definition of the Good data set was already given in section 3.2.1.

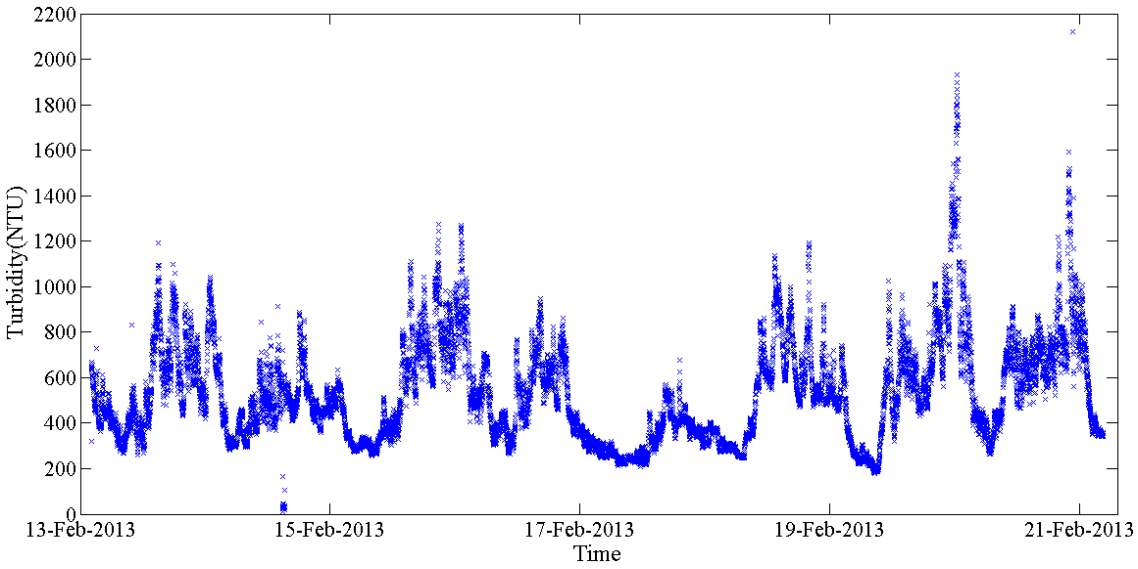

Figure 21 – Good data for model selection

According to the previous explanations, all combinations of the different models are tested for the Good data set as follows:

1- ARMA([1,1],[1,2],[2,1],[2,2]), ARIMA([1,1],[1,2],[2,1],[2,2]) and Pure integrator

2- Window sizes: 1000, 100, 30, 10

The window sizes are selected according to the existing harmonics in the data discussed in 3.1. It is important to have a window size that is smaller than the fastest harmonic in the time series. With 10 and 30 data points in the window, knowing that the sampling time is 1 minute, we can follow the 1 hour harmonic in the data. With 1000 and 100 window sizes, the 24 hours harmonic in the data can be followed.

It should be noted that the tests have been limited to the orders of the model structures mentioned above since more numbers of poles and zeros may lead to the overfit of the model to the data.

Table 2 shows the results of the accuracy of the one step-ahead forecasts made by the different structures.

Table 2 – Model and window size selection, ranked according to the RMSE order

Rank	Model $[n_a, n_c]$	Window Size	RMSE
1	ARIMA [2,2]	10	26.773
2	ARIMA [2,1]	10	29.384
3	ARIMA [1,1]	10	31.453
4	ARIMA [2,2]	30	32.748
5	ARIMA [1,2]	10	32.787
6	ARMA [2,2]	10	33.141
7	ARIMA [1,2]	30	33.9
8	ARIMA [2,1]	30	33.967
9	ARIMA [1,1]	30	35.556
10	ARIMA [2,2]	100	36.139
11	ARIMA [1,2]	100	37.061
12	ARIMA [2,1]	100	37.304
13	ARMA [2,1]	30	37.377
14	ARMA [2,1]	10	37.715
15	ARMA [2,2]	100	37.762
16	ARIMA [1,1]	100	38.193
17	ARIMA [2,2]	1000	38.922
18	ARMA [1,2]	10	39.013
19	ARMA [2,2]	1000	39.172
20	ARIMA [1,2]	1000	39.174
21	ARIMA [2,1]	1000	39.267
22	ARMA [1,2]	1000	39.348
23	ARMA [1,2]	100	39.471
24	ARMA [1,1]	30	39.515
25	ARMA [1,1]	100	39.537
26	ARMA [2,1]	100	39.549
27	ARMA [2,2]	30	39.597
28	ARIMA [1,1]	1000	39.643
29	ARMA [1,1]	1000	39.919
30	ARMA [2,1]	1000	40.376
31	Pure integrator	-	41.898
32	ARMA [1,2]	30	42.005
33	ARMA [1,1]	10	43.038

A quick look at Table 2 shows that the best fits are obtained by an ARIMA model with 10 data points in the moving window and among them, the ARIMA with two poles and two zeros gives the least RMSE. However, as mentioned in section 4.2.1.5, for this small window of the data the model may adapt to the noise instead of to the real dynamics. On the other hand, with two poles and two zeros the model may overfit to the data.

According to the given explanations, in order to be more confident about our choice, we also select the ARIMA with one pole and one zero and 30 data points in the moving window and will test its performance in the univariate method to detect outliers and faults. From now on the terms ARIMA10[2,2] and ARIMA30[1,1] will be used to represent the selected model structures. Figure 22 shows the real data and forecasts made by the ARIMA30[1,1].

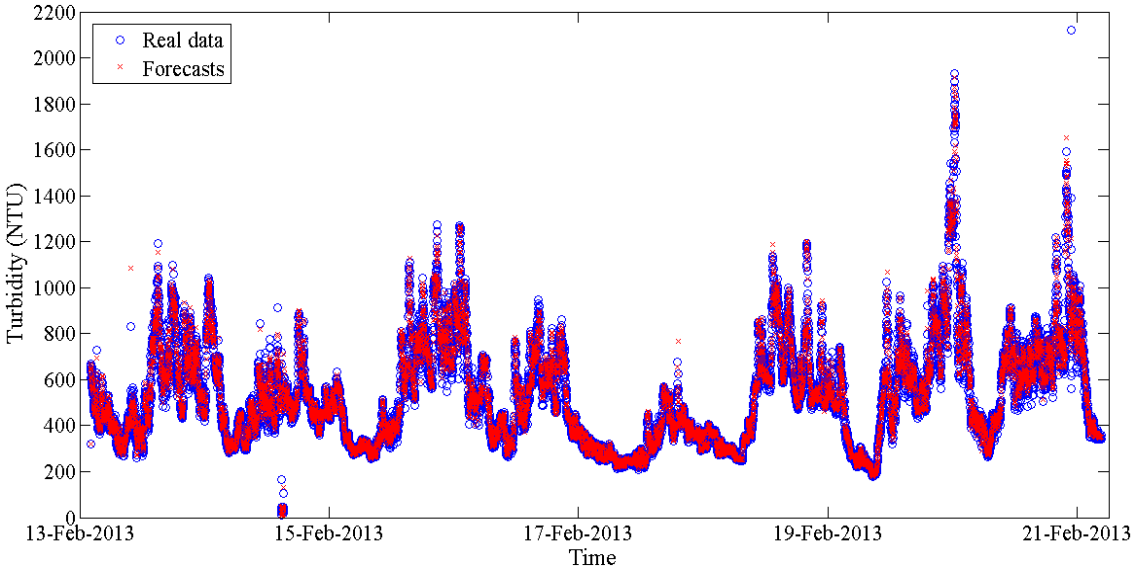

Figure 22 – Forecasts with ARIMA30[1,1]

The identified parameters of the ARIMA10[2,2] and ARIMA30[1,1] and their standard deviations (in parenthesis) are presented for 14th of February 2013 at 22:10':30'' in equations (70) and (71).

ARIMA10[2,2]:

$$A(z^{-1}) = 1 + 0.5317(\pm 0.2205)z^{-1} + 0.6316(\pm 0.1757)z^{-2}$$
$$C(z^{-1}) = 1 + 1.6774(\pm 0.1719)z^{-1} + 0.7616(\pm 0.1410)z^{-2} \quad (70)$$

ARIMA30[1,1]:

$$A(z^{-1}) = 1 + 0.7359(\pm 0.2272)z^{-1}$$
$$C(z^{-1}) = 1 + 0.9611(\pm 0.1228)z^{-1} \quad (71)$$

4.2.3. Calibration of the exponential smoothing model

In the previous section, the procedure followed to select the ARIMA model structure and the moving window size for the Good data set was presented. However, the objective is to compare the performance of this model with the exponential smoothing model in giving the one-step ahead forecasts. To do so, different orders of the exponential smoothing model should first be calibrated for the same Good data set shown in Figure 21. The calibration step for this model simply consists of calculating the optimal forgetting factor for different orders of the exponential smoothing model according to the procedure explained in section 3.2.1. The best α values for the 1st, 2nd and 3rd order exponential smoothing models were respectively 0.63, 0.28 and 0.19. Figure 23, Figure 24 and Figure 25 show the RMSE evolution as function of α, for the 1st, 2nd and 3rd order exponential smoothing models, respectively.

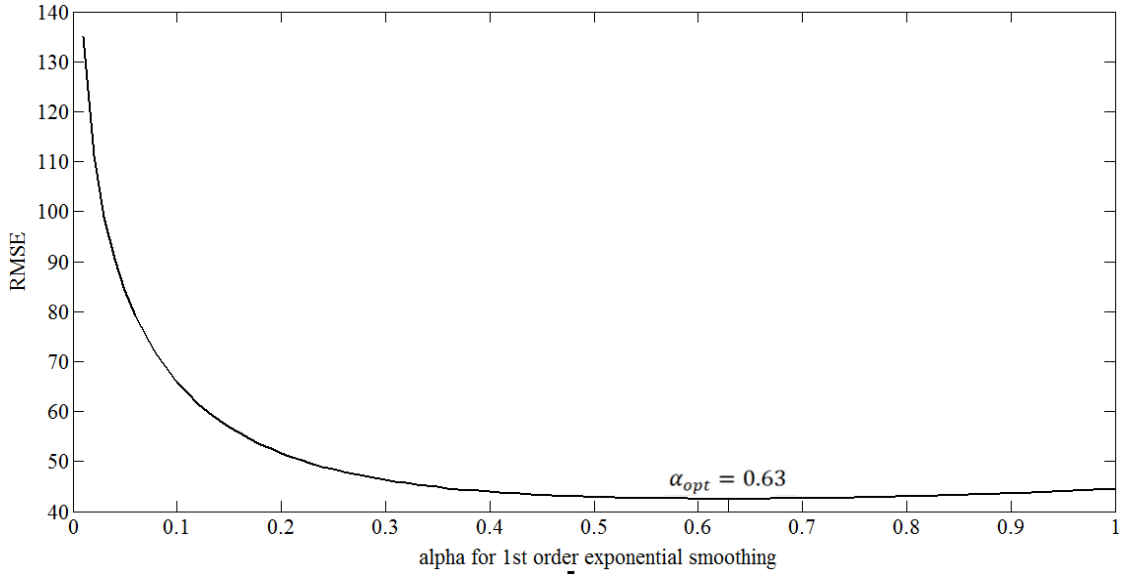

Figure 23 – RMSE minimization with the 1st order exponential smoothing model

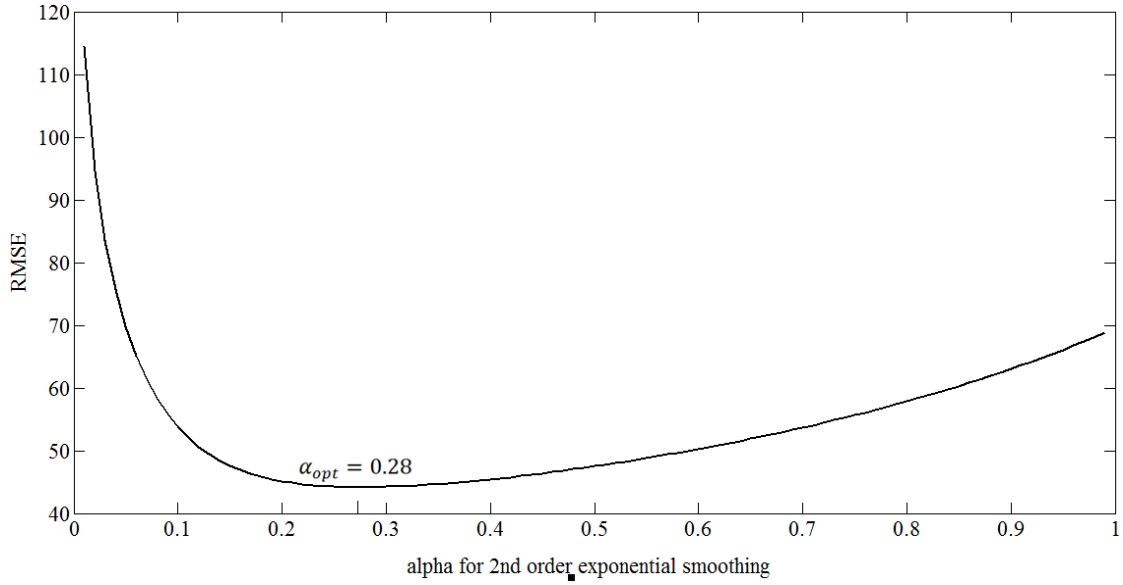

Figure 24 – RMSE minimization with the 2nd order exponential smoothing model

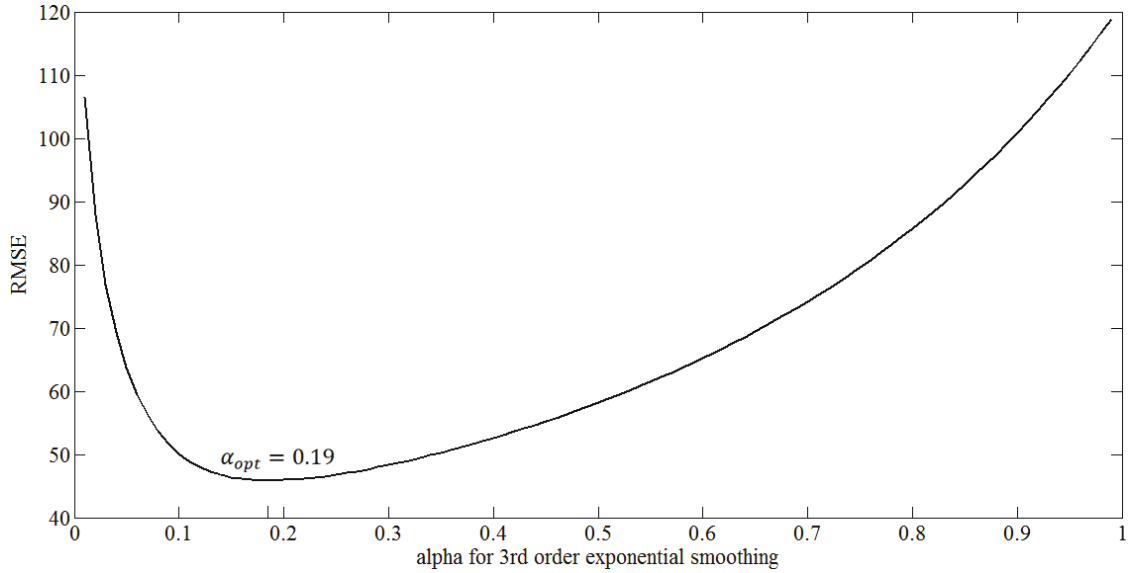

Figure 25 – RMSE minimization with the 3rd order exponential smoothing model

Once the optimal forgetting factors are obtained, the one-step ahead forecasts can be produced according to the equations in section 2.4. The accuracy of the forecasts are calculated according to the RMSE and the results are compared with the ones made by the ARIMA10[2,2] and ARIMA30[1,1] in Table 3.

Table 3 – Exponential smoothing models compared to ARIMA10[2,2] and ARIMA30[1,1] – Accuracy of the 1 step-ahead forecasts

Rank	Model	RMSE
1	ARIMA10[2,2]	26.773
2	ARIMA30[1,1]	35.556
3	1st order exponential smoothing model	42.521
4	2nd order exponential smoothing model	44.181
5	3rd order exponential smoothing model	45.896

The results in Table 2 and Table 3 show that, in general, the quality of the one-step ahead forecasts made by the moving window ARIMA model has improved according to the RMSE criterion. Specifically, if we compare the 3rd order exponential smoothing and the ARIMA10[2,2], we can notice a 43% of improvement in the accuracy of the forecasts. This can be due to either the capability of the ARIMA model to forecast the one-step ahead value or the advantage of adapting to the varying dynamics by using the moving window approach.

However, since the ARIMA model with moving window takes a very small window of data at each time of the iteration of the algorithm, at the calibration phase, it is computationally relatively heavy and time consuming in comparison with the exponential smoothing method.

4.3. Application of the univariate method to the case study system

In the previous section, the ARIMA model structure and window size were selected according to the Good data set by using the RMSE criterion. The next step is to apply the identified moving window ARIMA model to the univariate method and to study its performance in detecting outliers and faults in comparison with the univariate method using the third order exponential smoothing model.

The procedure consists of two steps. One is the calibration step and the other one is the validation step. In the calibration step, parameters of different blocks in the univariate method (explained in section 3.2) are tuned for the Good data set in Figure 21. In the validation step, the calibrated parameters will be used and the performance of the two models, i.e. the moving window ARIMA model and the 3rd order exponential smoothing model, in the univariate method will be validated for the validation data sets.

4.3.1. Calibration of the univariate method

Figure 26 shows the different blocks of the univariate method. The parameters that need to be calibrated in each step are also indicated in parentheses. In the outlier detection step, the procedure consists of 1) Forecasting the one-step ahead data and 2) Calculating the prediction error interval. To forecast the one-step ahead data, two different models are tested. 1) The moving window ARIMA model (ARIMA10[2,2] and ARIMA30[1,1]) and 2) The 3rd order exponential smoothing model. The 3rd order exponential model was calibrated for the Good data set in section 4.2.3 and $\alpha_{opt} = 0.19$ was obtained.

To calculate the prediction error interval, according to which outliers are detected, the 1st order exponential smoothing model is used to estimate the standard deviation of the forecast errors. This part is also calibrated for the same Good data set in Figure 21 and consists of calculating the optimal value for η in equation (33) and tuning the proper value for the coefficient L in equation (34).

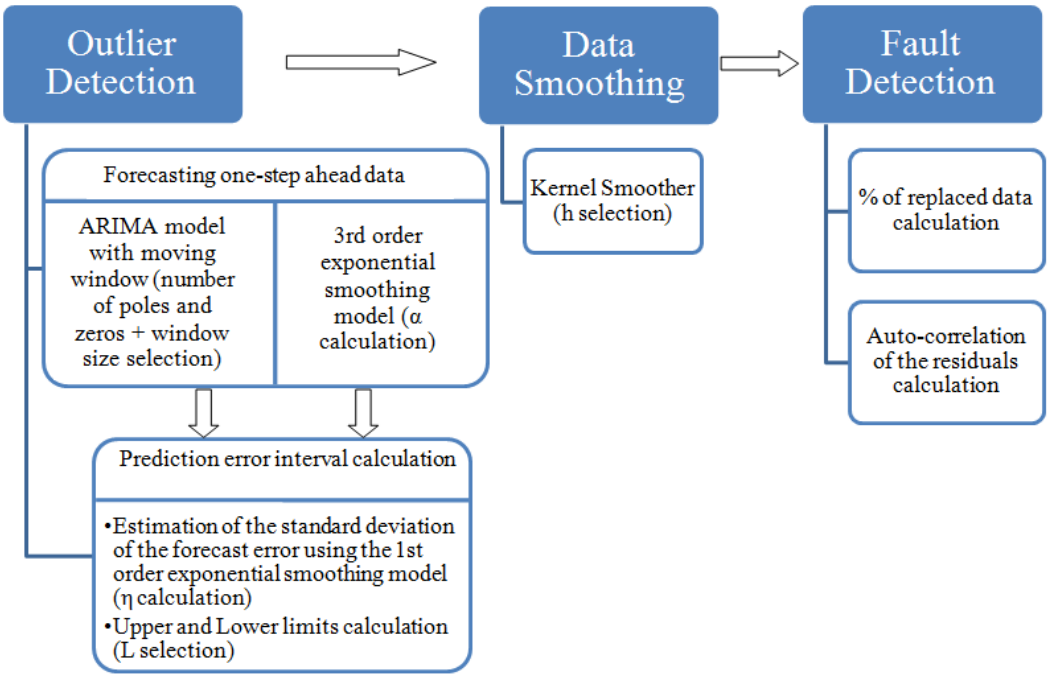

Figure 26 – Application of the moving window ARIMA model to the Univariate method

Figure 27 shows the RMSE changes as function of η for the forecasts made with the 3rd order exponential smoothing model. The η value that gives the minimum RMSE between the forecast mean absolute deviation and the absolute value of the one-step ahead forecast error is 0.17.

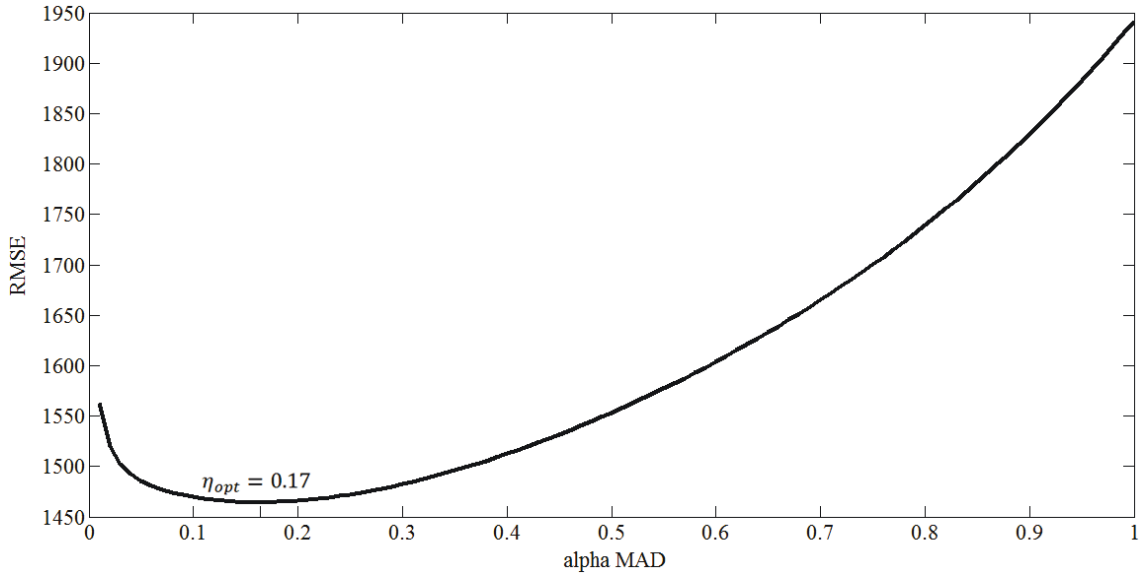

Figure 27 – η minimization

To tune L, different values of the this coefficient used for calculating the prediction error interval were tested for the Good data set to make the outlier detection more or less restrictive. Figure 28 shows that the outlier detection with $L = 3$ is too restrictive (for instance on February 18th) and leads to the incorrect rejection of real dynamics that are falsely detected to be outliers.

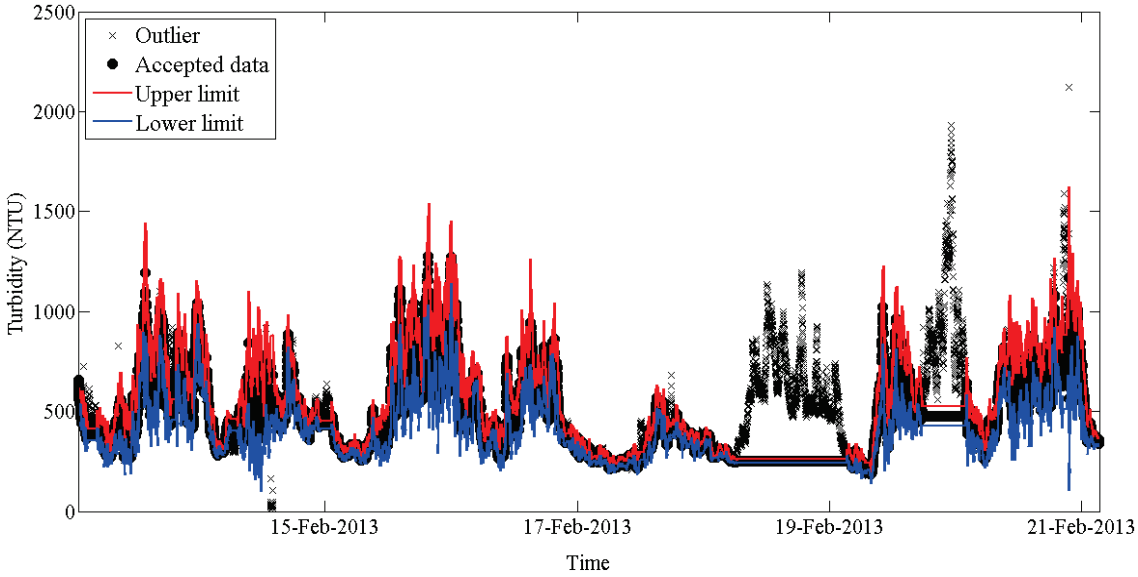

Figure 28 – Outlier detection with ARIMA10[2,2] – L = 3 – Bad tuning effects

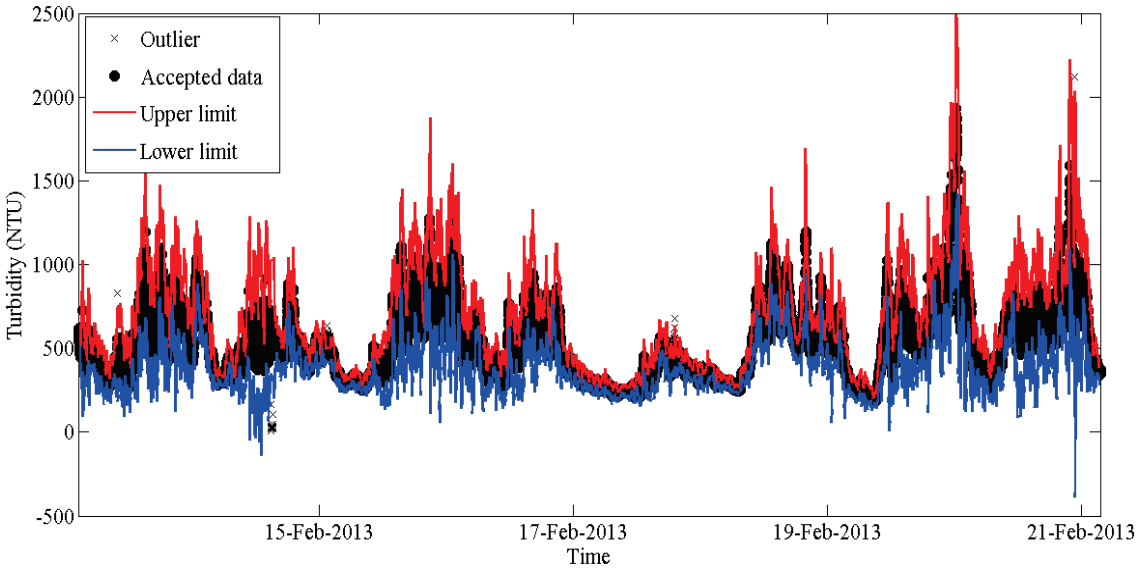

Figure 29 – Outlier detection with ARIMA10[2,2] – L = 5 – Good tuning

It should be noted that due to time limitations the backward reinitialisation algorithm, that prevents the outlier detection of getting out of control with the 3rd order exponential smoothing model (explained in section 3.2.1), is not adapted to the outlier detection with the moving window ARIMA model. In order to compare the two models in the univariate

method, this algorithm was thus disabled for the exponential smoothing model to provide the same situation for both models.

To smooth the data with the kernel smoother in the second block of Figure 26, the only parameter that had to be tuned, was h. As discussed in section 3.2.2, this parameter was set to 10 using the previous experience with similar data and according to the further application of the data. This selection leads to a good trade-off between the large variance and the large bias of the smoothed data.

Figure 30 shows the accepted data (when the forecasts are made with the ARIMA10[2,2]) and the smoothed data with $h = 10$.

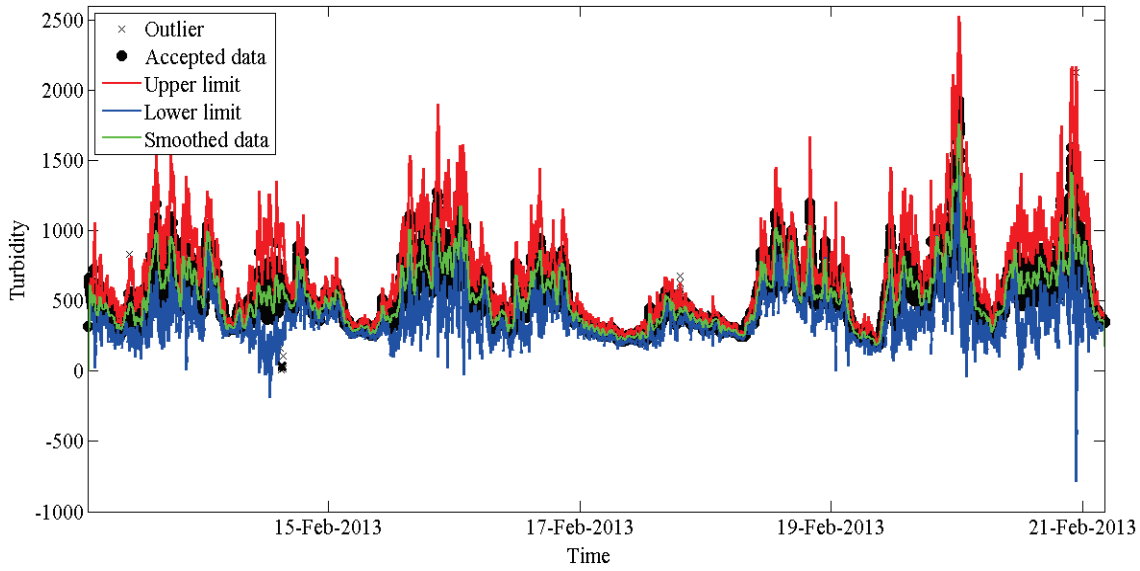

Figure 30 – Accepted and smoothed data with L = 5

The last step in the univariate method shown in Figure 26 is the detection of faults in the data. In order to compare the efficiency of the two models, the moving window ARIMA model and the 3rd order exponential smoothing model, in detecting the probable faults in the data series, two of the data features (introduced in Table 1) were taken into account. The "percentage of replaced data" and the "Auto-correlation of the residuals" will be calculated with accepted data and smoothed data produced by the two models for different data sets. The reason for selecting these two data features is that they give good measures of the fit of the model to the system or the goodness of the models in detecting outliers.

The percentage of replaced data represents the fraction of the raw data that is detected to be outliers for the windows of the data from $k-h$ to $k+h$. The auto-correlation of the residuals shows the fit of the model to the raw data or the randomness of the residuals. The data features will be calculated for accepted and smoothed data produced by the two models and will then be considered with their acceptability limits to detect faults.

In practice, the acceptability limits are defined according to the knowledge of the expert about the characteristics of the variables, the measurement sensors and their locations and finally the decision of the user according to the further application of the data. In this project in order to provide the tools to compare the performance of the two models, the following limits are considered to classify the data features as valid or invalid. If the percentage of replaced data exceeds the limit of 20%, the data for that period is considered doubtful. If the failure of the runs test exceeds the 95% confidence interval (i.e. the residuals are non-randomly correlated) the data for that period is also considered doubtful. If there is a period for which the two tests vote for the doubtfulness of the data, the data for that period is considered invalid.

The percentage of doubtful data and invalid data will be calculated with both models for the different data sets. An expert can then make a decision about the efficiency of the models according to his/her knowledge about the data. If a model leads to a large percent of doubtful data, while the expert knows that the period is clean and does not include many faults, we can include that the model is not efficient in detecting outliers or faults.

Figure 31, Figure 32 and Figure 33 show the application of the univariate method with the ARIMA10[2,2], ARIMA30[1,1] and 3rd order exponential smoothing models respectively for the calibration data set shown in Figure 21 with $L = 5$. In all of these figures, the first plot shows the flow data. This figure is included for the reader to better understand certain dynamics observed in the turbidity data. The second and third plots show the turbidity dynamics while the upper (red) and lower (blue) confidence limits are removed in the third figure to show more clearly the detection of outliers. The green line represents the smoothed data. The fourth and fifth plots show the two data features, the percentage of

replaced data and the auto-correlation of the residuals (applying the run test). The straight lines in the run test plots show the 95% confidence interval for the data to be independent.

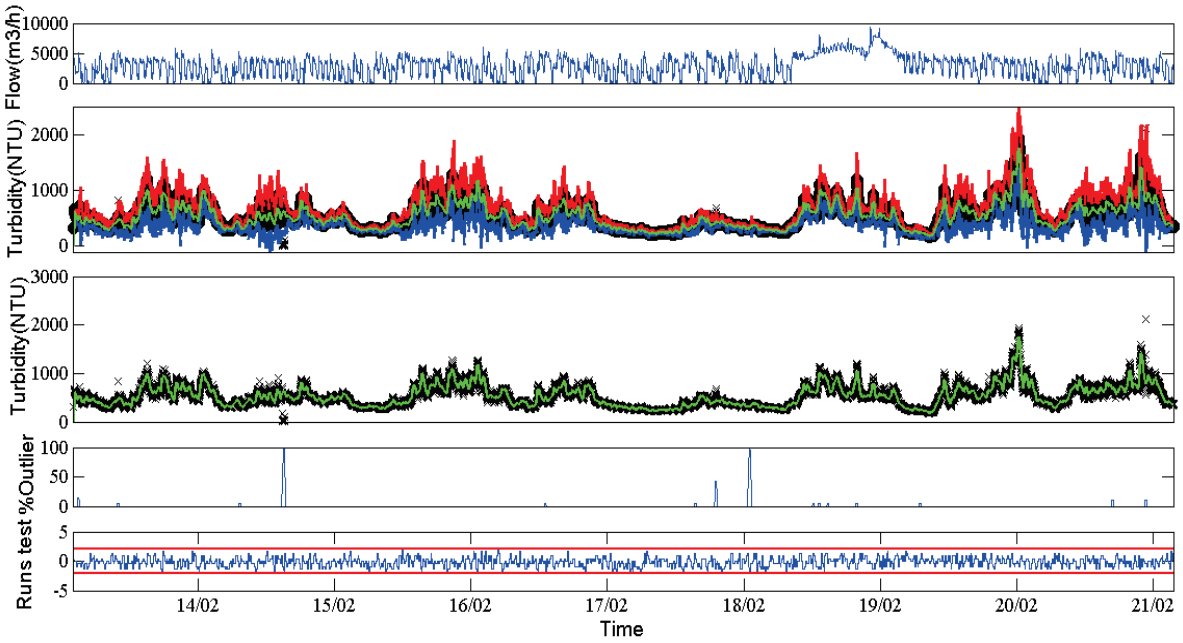

Figure 31 – Univariate method with the ARIMA10[2,2] model – Calibration data set (explanation of the figure, see text)

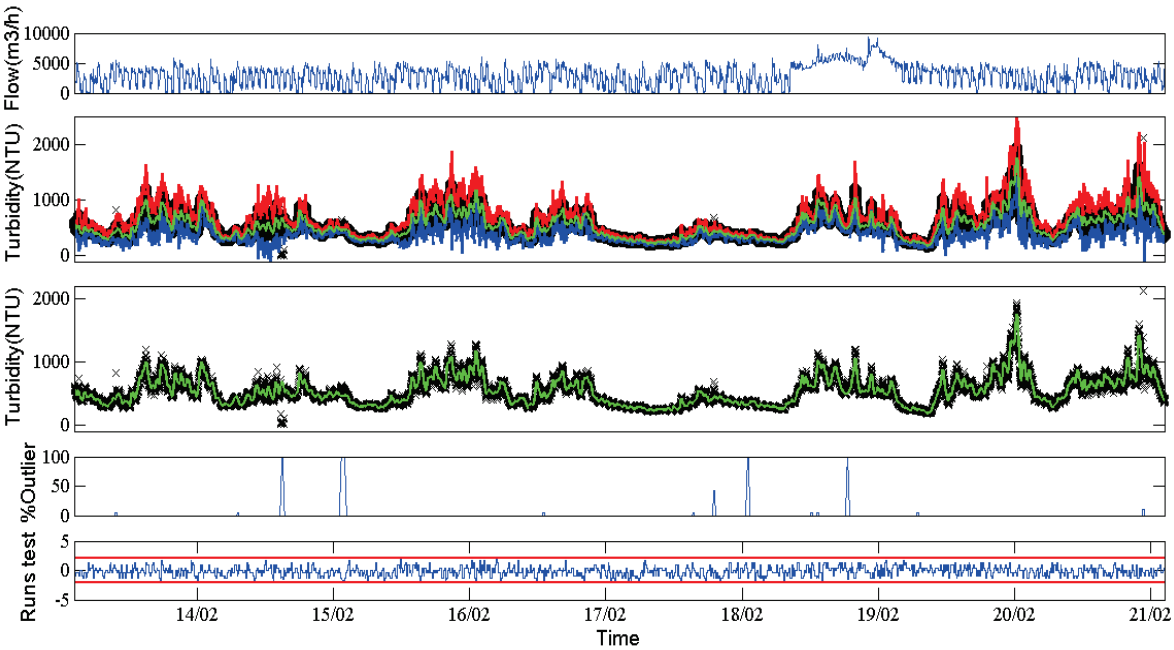

Figure 32 – Univariate method with the ARIMA30[1,1] model – Calibration data set (explanation of the figure, see text)

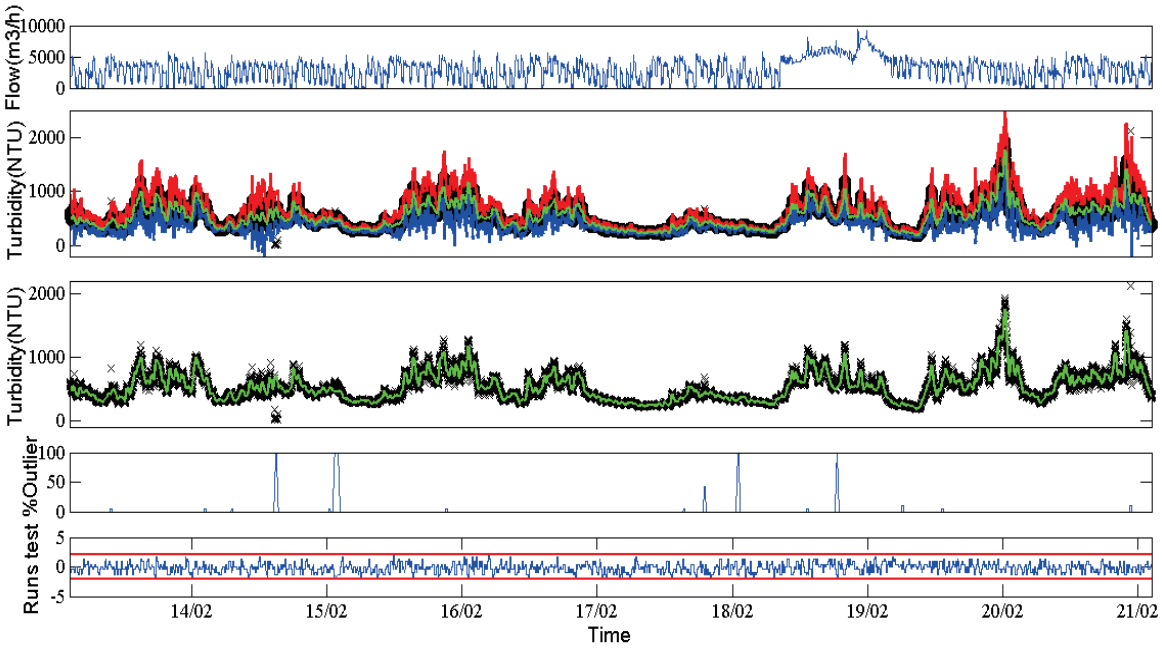

Figure 33 – Univariate method with the 3rd order exponential smoothing model – Calibration data set (explanation of the figure, see text)

Table 4 gives the results of the analysis of the fault detection tests for the calibration data set. It can be observed that the results of the ARIMA30[1,1] and the 3rd order exponential smoothing models are similar with very slight differences. As was previously known, the calibration data set is clean and does not have many outliers or any faults. Therefore, both models perform correctly in detecting outliers and faults in the calibration phase. However, the ARIMA10[2,2] model detects less outliers and thus less doubtful data. The reason may be that this model was selected according to its best fit to the calibration data set and thus may be overfitted to the data.

Table 4 – Fault detection for calibration data set

Model	% Replaced data	% Doubtful data	% Invalid data
ARIMA10[2,2]	0.8571	1.0285	0
ARIMA30[1,1]	1.7571	1.9285	0
3rd order exponential smoothing	1.8685	2.0399	0

4.3.2. Validation of the univariate method

After calibrating the parameters of the models and the univariate method according to the calibration data set, the performance of the univariate method in detecting outliers and faults should be also validated on the basis of data sets different from the calibration data set. Similar to the calibration step the percentage of doubtful data and invalid data detected by both models will be calculated and the results will be compared.

Two validation data sets were selected. One is related to the rain event occurring on the 30th of January and the other one is related to the rain of the 26th of April. The univariate method with the ARIMA10[2,2] model was applied to the validation data sets and the results are shown in Figure 34 and Figure 35.

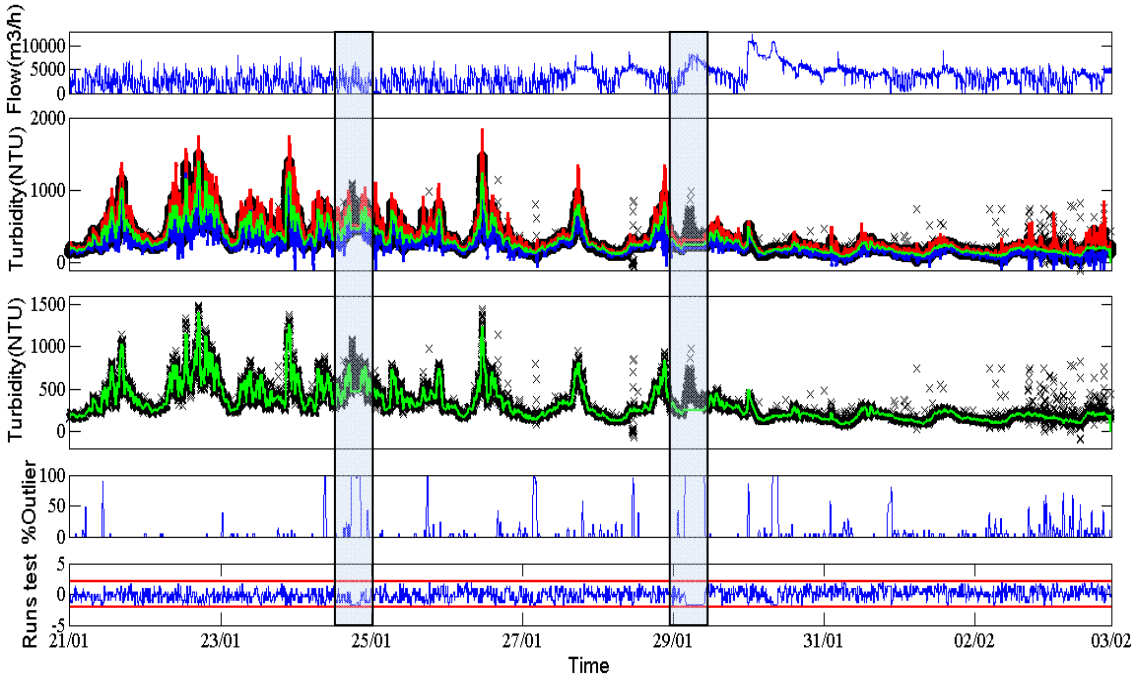

Figure 34 – Univariate method with ARIMA10[2,2] model – Validation data set of January 30th (explanation of the figure, see text)

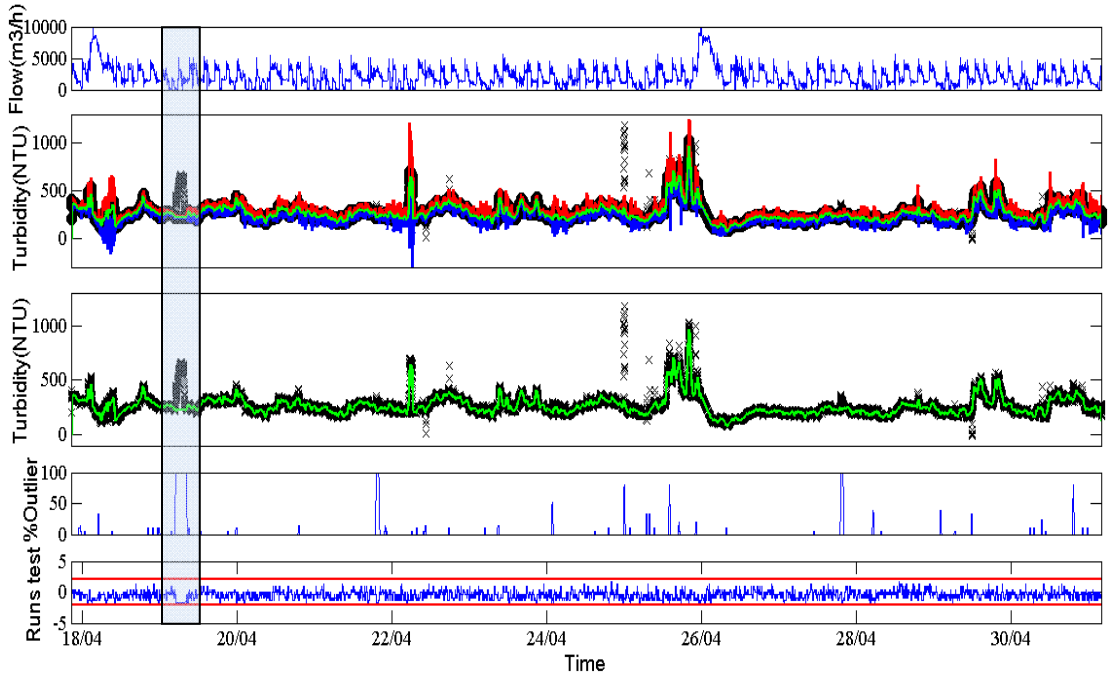

Figure 35 – Univariate method with the ARIMA10[2,2] model – Validation data set of April 26th (explanation of the figure, see text)

In Figure 34 and Figure 35, a few periods are detected and marked in which the real dynamics are rejected as outliers. As previously discussed in section 4.2.2, the reason for this behavior of the ARIMA10[2,2] model can be that very few data points are taken in the moving window or the model is overfitted to noise. In section 4.2.2, the ARIMA10[2,2] was selected according to its fit to the calibration data set. However, for the validation data sets, no satisfactory results are obtained. Therefore, the ARIMA30[1,1] model, which led to a trade-off between the acceptable fit to the calibration data set and the simpler structure with more data points in the moving window, was also applied to the univariate method. The results are shown in Figure 36 to Figure 39.

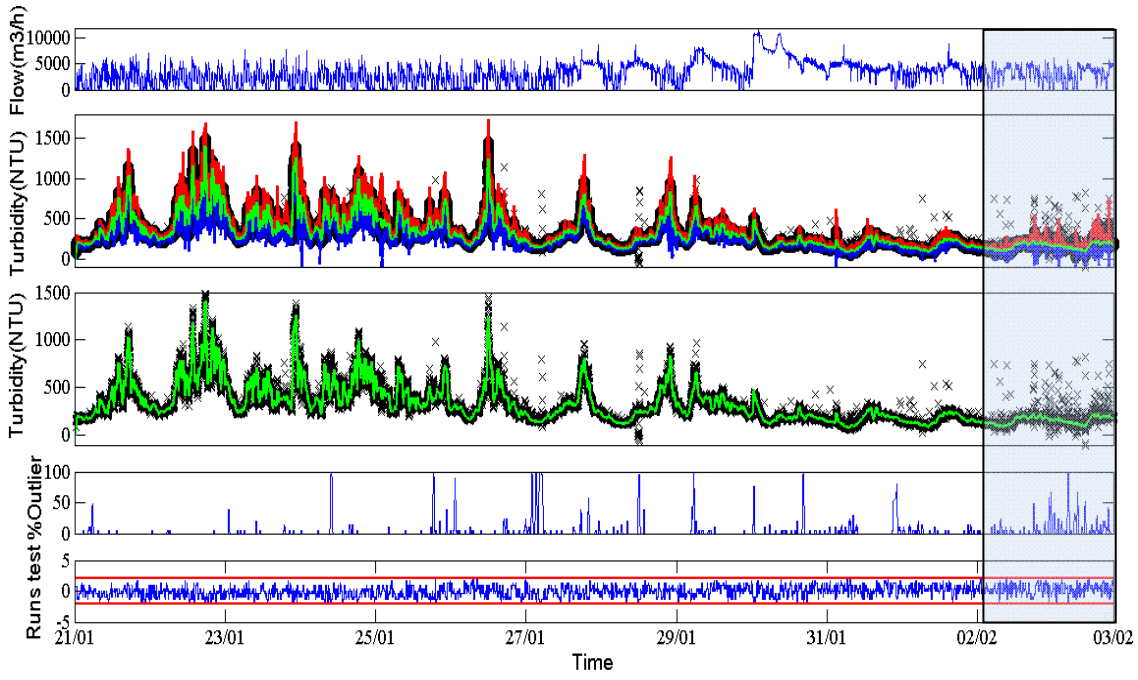

Figure 36 – Univariate method with the ARIMA30[1,1] model – Validation data set of January 30th (explanation of the figure, see text)

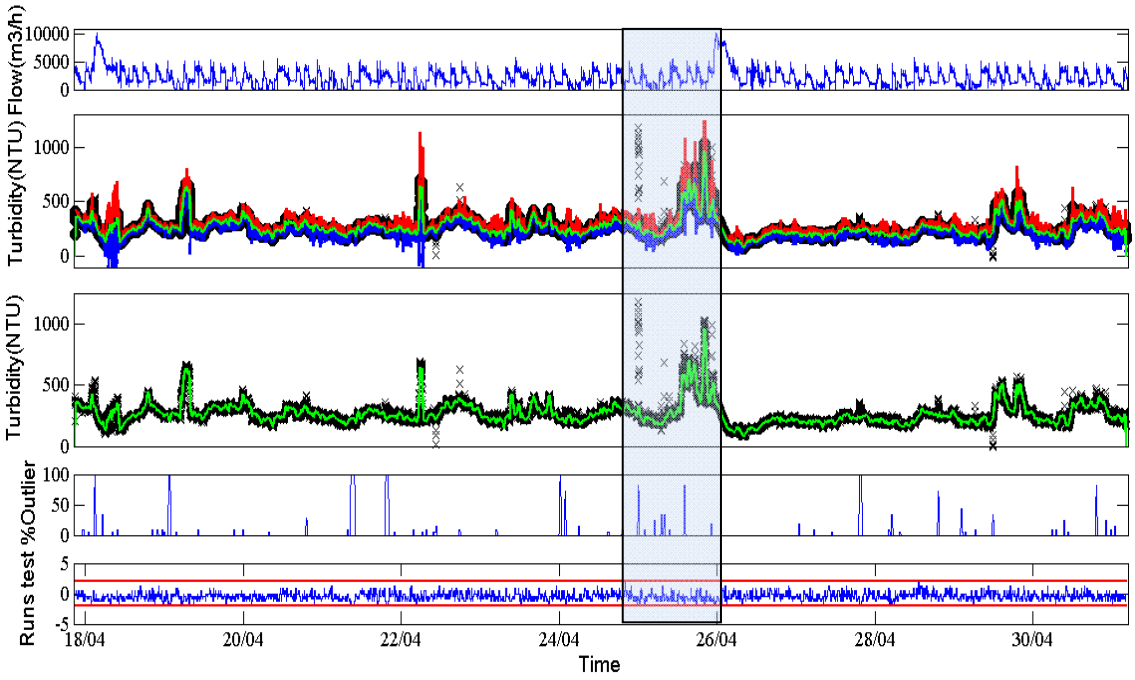

Figure 37 – Univariate method with the ARIMA30[1,1] model – Validation data set of April 26th (explanation of the figure, see text)

It can be observed that the problematic periods of the ARIMA10[2,2] model do not exist for the ARIMA30[2,2].

A closer look of the periods marked in Figure 36 and Figure 37 are shown in Figure 38 respectively in Figure 39. In these figures it can be observed that the moving window ARIMA model performs correctly in the detection and replacement of the outliers.

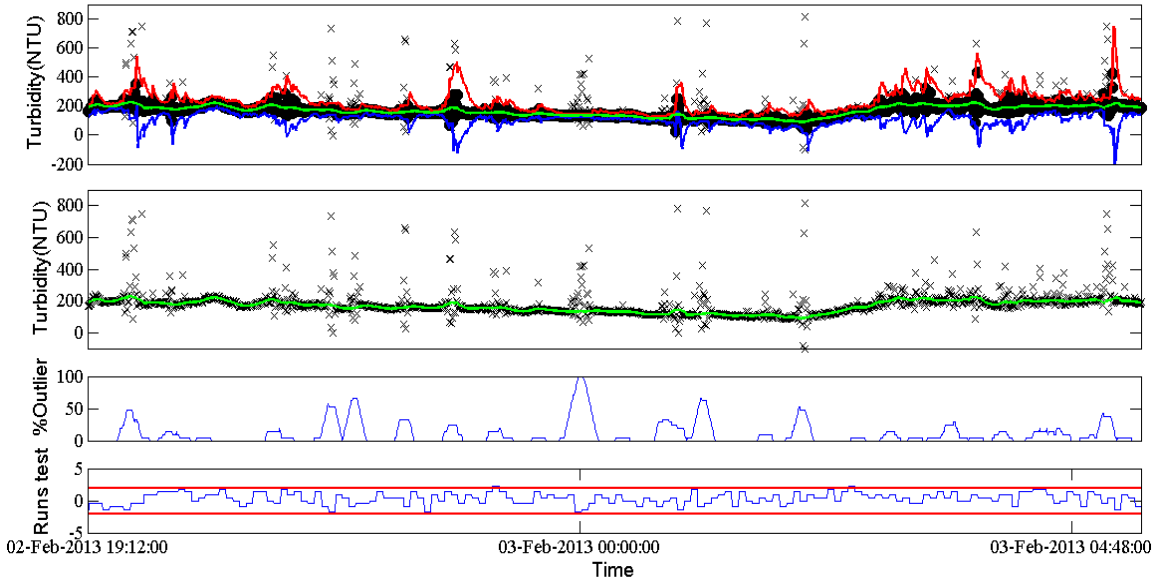

Figure 38 – Closer look of the outlier detection of the period marked in Figure 36 (explanation of the figure, see text)

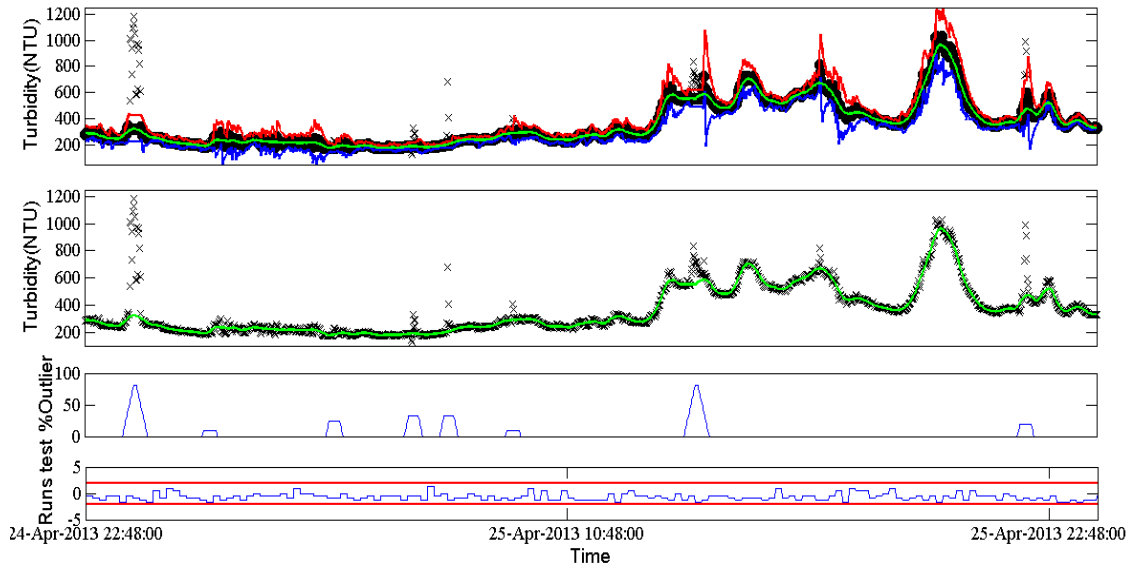

Figure 39 – Closer look of the outlier detection of the period marked in Figure 37 (explanation of the figure, see text)

Considering the results of the application of the ARIMA30[1,1] model to the univariate method, we can conclude that the model is successful in getting adapted to the dynamics in rain periods while detecting the outliers in dry periods.

Similar results related to the application of the 3rd order exponential smoothing model to the univariate method are demonstrated in Figure 40 and Figure 41.

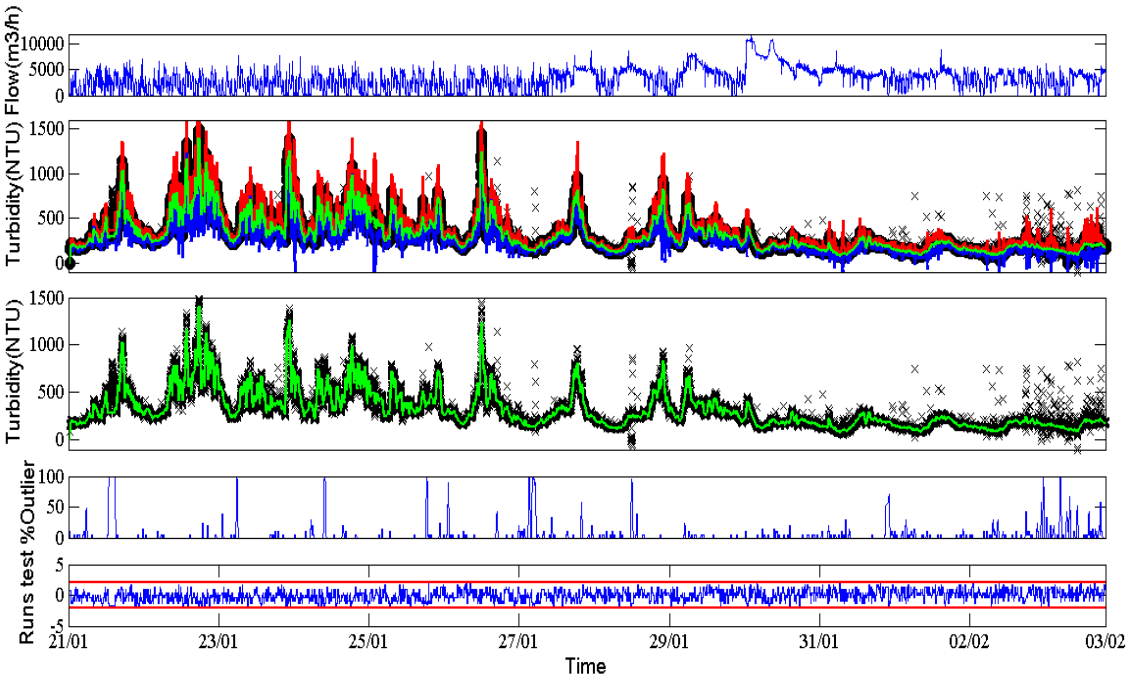

Figure 40 – Univariate method with the 3rd order exponential smoothing model – Validation data set of January 30th (explanation of the figure, see text)

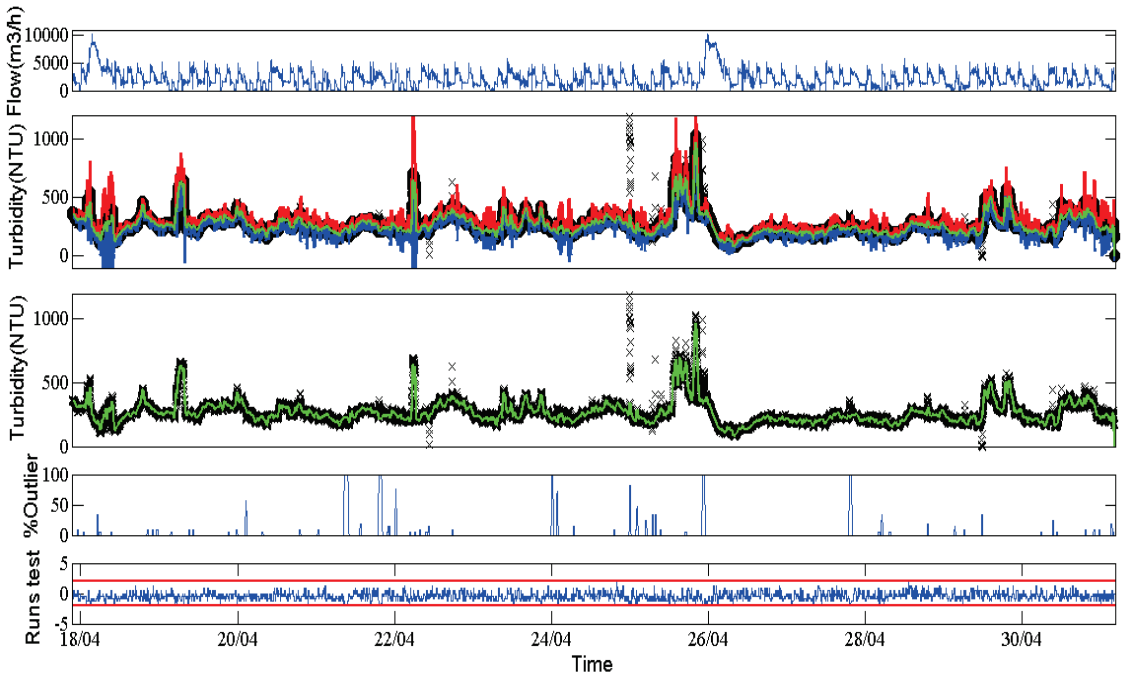

Figure 41 – Univariate method with the 3rd order exponential smoothing model – Validation data set of April 26th (explanation of the figure, see text)

Table 5 shows the results of the analysis of the fault detection tests with the ARIMA10[2,2], ARIMA30[1,1] and 3rd order exponential smoothing models for the validation data sets.

Table 5 – Fault detection for validation data sets

Model / Rain event	ARIMA10[2,2]			ARIMA30[1,1]			3rd order exponential Smoothing		
	% Replaced data	% Doubtful data	% Invalid data	% Replaced data	% Doubtful data	% Invalid data	% Replaced data	% Doubtful data	% Invalid data
January 30th	8.0014	8.4854	0.070586	4.7948	5.4805	0.0202	5.1578	5.662	0
April 26th	3.1878	3.1878	0	3.6052	3.6052	0	3.1721	3.1721	0

Comparison of the results in Table 5 shows that the ARIMA30[1,1] model and the 3rd order exponential smoothing model lead to similar fault detection performance. As discussed before, the ARIMA10[2,2] model rejected more data as outliers in the January 30th data set. That is also the reason why we notice the difference especially in the percentage of replace

data for this period. Therefore the ARIMA30[1,1] is finally selected as the alternative model of the outlier detection system.

4.4. Discussion

In this chapter the exponential smoothing models and the way they give forecasts were revisited. Considering the empirical way these models are employed to detect outliers, the aim was to propose a more systematic approach to find a model to describe the system.

Accordingly, in the calibration step the ARIMA models were fitted to a moving window data of the turbidity time series and the quality of the one-step ahead forecasts were compared to the ones made by the exponential smoothing models according to the RMSE criterion. An ARIMA model structure was found with two poles and two zeros and 10 data points in the moving window that led to the least RMSE. However, in order to prevent the overfit of the model to the system, the ARIMA30[1,1] model, that provided a good trade-off between an acceptable performance and a simple structure with a larger window size, was selected as an alternative model of the system.

The identified moving window ARIMA models, ARIMA10[2,2] and ARIMA30[1,1], were then used in the univariate method to detect outliers and faults in the turbidity time series in comparison with the 3rd order exponential smoothing model. The results showed that while at the calibration step all the models performed similarly, the performance of the ARIMA10[2,2] degraded slightly for the validating data sets. Therefore, the ARIMA30[1,1] was selected as the model of the system.

Tuning of different steps of the univariate method is really important to obtain the satisfactory performance. It was observed that in tuning of the method the previous experience of the expert in working with similar systems and his/her knowledge about the further application of the data is of great importance.

Using the moving window concept with the ARIMA model helped getting adapted to the varying dynamics. The results in the RMSE calculation for one-step ahead forecasts

demonstrated the improvement made by this approach compared to the exponential smoothing model. In addition, this model was successful in rejecting outliers while following the fast turbidity dynamics during rain events. However, according to the calculation of data features like the percentage of replaced data and the auto-correlation of the residuals, not much difference could be noticed in their application to the univariate method for outlier rejection and fault detection compared to the exponential smoothing model. This can be related to different reasons.

One of the possible reasons is that in the univariate method only the one-step ahead forecasts are taken into consideration. For further steps ahead the ARIMA model with moving window is expected to perform considerably better and may show better performance in detection of outliers and faults. Since the objective of the univariate method application has always been to study the one-step ahead forecasts, the ARIMA approach for further steps ahead predictions was not tested.

The other reason may be that for the estimation of the standard deviation of the forecast error, according to which the prediction error interval is calculated, a 1^{st} order exponential smoothing model is fitted to the one-step ahead forecast errors. Due to time limitations, only the same approach with the ARIMA model to detect outliers was tested. If another method to calculate the confidence interval of the forecasts more compatible with the new forecasting method would have been used, different results may have been obtained.

After all, the performance of the two models in the univariate method has been compared for only a few periods of data in specific situations. The two models may perform differently in different situations such as collected periods with faster dynamics or with more outliers and faults. On the other hand, it is hard to exactly know whether a data point is an outlier or a fault. More reliable comparison results may be obtained if one can test the method for artificially produced data sets to which specific outliers and faults are added deliberately.

Due to the application of the moving window approach, which repeats the model identification at each time step, the calibration and even validation of the new method imposes a heavier calculation load compared to the exponential smoothing method.

However, in the online application of the method, this task will not impose any limitation since at each time step the new model will be identified as the new data point emerges. For this case, identification of the new model according to the recent data set will no longer be time consuming.

5. CONCLUSIONS AND FUTURE WORK

In this research project, different steps of the univariate method proposed by Alferes et al. (2012) for detecting outliers and faults in the data provided by the mon*EAU* water quality monitoring stations were studied. The focus in this project was mostly on the outlier detection step. To detect outliers in the univariate method, Alferes et al. (2012) proposed fitting a 3rd order exponential smoothing model to the system, projecting it into the future and calculating one-step ahead forecasts. The exponential smoothing models and the way they produced the one-step ahead forecasts were revisited. Some mathematical developments allowed to discuss the empirical way the model was fitted to the system and the way the one-step ahead forecasts were calculated.

Accordingly, the objective was defined to propose a more systematic approach in calculating the one-step ahead forecasts. To do so, the ARIMA model with moving window approach was proposed.

The turbidity time series collected from the inflow to the primary clarifier at the inlet of a wastewater treatment plant in Copenhagen was selected as the case study time series. Different ARIMA model structures and window sizes were tested and the quality of their forecasts was compared by using the RMSE criterion. Finally, an ARIMA model with one pole and one zero with 30 data points in the moving window, that led to a trade-off between the acceptable fit to the calibration data set and the simpler structure, was proposed as the alternative model of the water quality monitoring system. This resulted in around 30% of improvement in the accuracy of the forecasts compared to the 3rd order exponential smoothing model.

Using the moving window approach with the ARIMA model imposed a heavier calculation load at the calibration and even validation steps in comparison with the exponential smoothing model. However, in the online application of the approach, identification of the new model according to the recent data set will no longer be time consuming.

Consequent to finding the moving window ARIMA model structure, outliers and faults in the time series were detected by the univariate method and the results were compared to the

ones detected by the 3rd order exponential smoothing model. It was observed that the two models behave similarly in detecting outliers and faults according to the specified criteria.

It was discussed that tuning of different steps of the univariate method was very effective in the method. To have the proper tunings, expert knowledge about the further application of the data is important and his/her previous experience working with similar systems is very helpful.

The objective of using the moving window approach was to adapt the model to the varying dynamics. The results of the RMSE calculation for the one-step ahead forecasts with the moving window ARIMA model at the calibration step proved the improvement made by this model compared to the exponential smoothing model. However, according to the specified criteria, the expected improvement in detection of outliers and faults was not observed. One of the possible reasons is that the categorization of the data as outlier or faulty can hardly be done with 100% confidence. Models can be compared with more reliability for artificial data sets with deliberately added outliers and faults. In addition, in the univariate method only the one-step ahead forecasts are taken into consideration while the moving window ARIMA model is expected to perform considerably better in forecasting for further steps ahead and may then also show better performance in detection of outliers and faults when applied to the univariate method. Above all, calculation of the prediction error interval according to which the data is classified as outlier or not is done by estimating the standard deviation of the forecast error using the 1st order exponential smoothing model also in the proposed ARIMA-based method. Using another approach to calculate the confidence interval of the forecasts, more compatible with the new forecasting model, may lead to different results.

For the future, it is first suggested to theoretically analyze the performance of the proposed moving window ARIMA approach using a simulated data set or a test bench.

It is also proposed to develop an alternative method to calculate the prediction error interval, by using a method to generate confidence intervals from an adaptive estimate of the variance of forecast errors.

Another suggestion is to add the backward reinitialisation algorithm proposed by Alferes et al. (2012) to the outlier detection step with the moving window ARIMA model. However, application of a similar procedure with the moving window data might not be as straightforward as it seems. Instead, it is proposed to consider an adaptive approach that modifies the value of L, the coefficient to make the prediction error interval more or less restrictive, according to the rate of change in the standard deviation of the forecast error.

It should be noted that the turbidity time series, which was measured at a specific location in a water system, was selected as the case study water quality parameter according to which the ARIMA model with moving window was calibrated. However, water quality parameters have different dynamics and properties. In addition, different parameters are not measured with the same measurement instrument and at the same sampling rate. Therefore, if one would like to apply the approach to other water quality parameters in future, all conducted tests should be repeated, possibly leading to a new model structure and window size.

Since water quality time series are often highly correlated, it is proposed to continue the work on the use of multivariate methods besides the univariate to draw useful information from the set of correlated data.

Using the RMSE criterion may not be the most systematic approach to select the best moving window ARIMA model structure. It is suggested to follow an alternative approach to find the order of the model or the window size and compare the results with the current model structure.

Finally, the Multi Model Filtering Algorithm (Maldonado et al., 2010), explained in section 2.2.3, could be applied to the system. This method can be useful for the case study system that exhibits various behavioral modes and may lead to very interesting results. In the calibration phase, to determine different states of the system, the range of variations in the

moving window ARIMA model parameters can be monitored and a limited number of possible model combinations can be considered. Subsequently, a bank of Kalman filters can be designed for the set of identified models according to which the states of the system can be estimated. By using Bayes' rule, the conditional probability of each model to represent the data can be calculated at each time step. The model that leads to the highest probability can be selected to represent the data and according to that the one-step ahead forecasts can be generated.

BIBLIOGRAPHY

Z.H. Abu-el-zeet, V.M. Becerra & P.D. Roberts. (2002) Combined bias and outlier identification in dynamic data reconciliation. Computers & Chemical Engineering, 26(6), 921-935.

D. Aguadoa & C. Rosen. (2008) Multivariate statistical monitoring of continuous wastewater treatment plants. Engineering Applications of Artificial Intelligence, 21(7), 1080–1091.

H. Akaike. (1974) A new look at the statistical model identification. Automatic Control, IEEE Transactions on, 19(6), 716-723.

J. Alferes, A. Lynggaard-Jensen, T. Munk-Nielsen, S. Tik, L. Vezzaro, A. Kumari Sharma, P. Steen Mikkelsen & P.A. Vanrolleghem (2013a) Validating data quality during wet weather monitoring of wastewater treatment plant influents. In: Proceedings of the 86th Annual Water Environment Federation Technical Exhibition and Conference (WEFTEC 2013). Chicago, United States, October 4-9 2013.

J. Alferes, P. Poirier, C. Lamaire-Chad, A.K. Sharma, P.S. Mikkelsen & P.A. Vanrolleghem (2013b) Data quality assurance in monitoring of wastewater quality: Univariate on-line and off-line methods. In: Proceedings of the 11th IWA Conference on Instrumentation, Control and Automation. Narbonne, France, September 18-20 2013.

J. Alferes, P. Poirier & P.A. Vanrolleghem (2012) Efficient data quality evaluation in automated water quality measurement stations. In: Proceedings of the International Environmental Modelling and Software Society (iEMSs). Leipzig, Germany, 2012.

S. Bai, J. Thibault & D.D. McLean. (2006) Dynamic data reconciliation: Alternative to Kalman filter. Journal of Process Control, 16(5), 485-498.

M. Basseville. (1988) Detecting changes in signals and systems- A survey. Automatica, 24(3), 309-326.

I. Ben-Gal (2005) Outlier Detection. In Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers. Kluwer Academic Publishers, Boston, USA.

P.M. Berthouex & G.E. Box. (1996) Time series models for forecasting wastewater treatment plant preforemance. Water Resources, 30(8), 1865-1875.

J.L. Bertrand-Krajewski, S. Barraud & B. Choca (2000) La mesure de l'impact environnemental des systèmes d'assainissement : Exemple de l'observatoire de terrain en hydrologie urbaine (OTHU). In: Proceedings of the Actes du 3°Congrès Universitaire de Génie Civil. Lyon, France, 27-28 juin 2000.

M.V. Bijnen & H. Korving (2008) Application and results of automatic validation of sewer monitoring data. In: Proceedings of the 11th International Conference on Urban Drainage. Edinburgh, Scotland, United Kingdom

G. Bloch, M. Ouladsine & P. Thomas. (1995) On-line fault diagnosis of dynamic systems via robust parameter estimation. Control Engineering Practice, 3(12), 1709-1717.

N. Branisavljevic, Z. Kapelan & D. Prodanovic. (2011) Improved real-time data anomaly detection using context classification. Journal of Hydroinformatics, 13(3), 307-323.

A. Campisano, P.J. Cabot, D. Muschalla, M. Pleau & P.A. Vanrolleghem. (2013) Potential and limitations of modern equipment for real time control of urban wastewater systems. Urban Water Journal, 10(5), 300-311.

S.C. Chan, Z. G. Zhang & K.W. Tse (2005) A new robust Kalman filter algorithm under outliers and system uncertainties. In: Proceedings of the Circuits and Systems, ISCAS 2005. IEEE International Symposium on. IEEE, 23-26 May 2005.

R. Conejo, E. Guzmán & J.L. Pérez-de-la-Cruz. (2007) Knowledge-based validation for hydrological information systems. Applied Artificial Intelligence, 21(8), 803–830.

D. Dochain & P.A. Vanrolleghem (2001) Dynamical Modeling and Estimation in Wastewater. IWA Publishing, London, UK.

M.A. Gandhi & L. Mili. (2010) Robust kalman filter based on a generalized maximum-likelihood-type estimator. Signal Processing, IEEE Transactions, 58(5), 2509-2520.

F.P. Garcia, D.J. Pedregal & C. Roberts. (2010) Time series methods applied to failure prediction and detection. Reliability Engineering & System Safety, 95(6), 698-703.

M.C. Hau & H. Tong. (1989) A practical method for outlier detection in autoregressive time series modelling. Stochastic Hydrology and Hydraulic, 3(4), 241-260.

D.J. Hill & B.S. Minsker. (2010) Anomaly detection in streaming environmental sensor data: A data-driven modeling approach. Environmental Modelling & Software, 25(9), 1014–1022.

K.W. Hipel & A.I. Mcleod. (1978) Preservation of the rescaled adjusted range: 2. Simulation studies using Box-Jenkins models. Water Resources Research, 14(3), 509-516.

P.J. Huber. (1964) Robust estimation of a location parameter. Annals of Mathematical Statistics, 35(1), 73-101.

H. Liu, S. Shah & W. Jiang. (2004) On-line outlier detection and data cleaning. Computers and Chemical Engineering, 28(9), 1635–1647.

L. Ljung (1987) System Identification: Theory for the User. Prentice-Hall Inc., Engelewood Cliffs, New Jersey, USA.

M. Maldonado, A. Desbiens, R. Del Villar R & R. Aguilera. (2010) On-line estimation of frother concentration in flotation processes. Canadian Metallurgical Quarterly, 49(4), 435-446.

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

D.C. Montgomery, L.A. Johnson & J.S. Gardiner (1990) Forecasting and Time Series Analysis 2nd Edition. McGraw-Hill Inc., New York, USA.

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

R. Pastres, S. Ciavatta & C. Solidoro. (2003) The Extended Kalman Filter (EKF) as a tool for the assimilation of high frequency water quality data. Ecological Modelling, 170(2), 227–235.

A. Patcha & J.M. Park. (2007) An overview of anomaly detection techniques: Existing solutions and latest technological trends. Computer Networks, 51(12), 3448–3470.

R.K. Pearson. (2002) Outliers in process modeling and identification. Control Systems Technology, IEEE Transactions on, 10(1), 55-63.

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

J. Rissanen. (1978) Modeling by shortest data description. Automatica, 14(5), 465-471.

M.G. Schimek (2013) Smoothing and Regression, Approaches, Computation and Application. John Wiley & Sons, New York, USA.

A.F.M. Smith & M. West. (1983) Monitoring renal transplants: An application of the multiprocess Kalman filter. Biometrics, 867-878.

T. Takahama & S. Sakai (2009) A comparative study on kernel smoothers in differential evolution with estimated comparison method for reducing function evaluations. In: Proceedings of the Evolutionary Computation, 2009. CEC'09. Congress on. IEEE, 2009

J.W. Taylor. (2010) Triple seasonal methods for short-term electricity demand forecasting. European Journal of Operational Research, 204(1), 139-152.

M. Thomann. (2008) Quality evaluation methods for wastewater treatment plant data. Wat. Sci. Tech., 57(10), 1601-1609.

M. Thomann, L. Rieger, S. Frommhold, H. Siegrist & W. Gujer. (2002) An efficient monitoring concept with control charts for on-line sensors. Water Science & Technology, 46(4-5), 107-116.

J.A. Ting, A. D'Souza & S. Schaal (2007) Automatic outlier detection: A Bayesian approach. In: Proceedings of the Robotics and Automation, 2007 IEEE International Conference on. IEEE, 2007. Rome, Italy

V. Venkatasubramanian, R. Rengaswamy, K. Yin & S.N. Kavuri. (2003a) A review of process fault detection and diagnosis - Part I: Quantitative model based methods. Computer & Chemical Engineering, 27(3), 293-311.

V. Venkatasubramanian, R. Rengaswamy, K. Yin & S.N. Kavuri. (2003b) A review of process fault detection and diagnosis: Part II: Qualitative models and search strategies. Computers & Chemical Engineering, 27(3), 313-326.

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

M. West. (1981) Robust sequential approximate Bayesian estimation. Journal of the Royal Statistical Society. Series B (Methodological), 43, 157-166.

A.S. Willsky. (1976) A survey of design methods for failure detection in dynamic systems. Automatica, 12(6), 601-611.

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