Automated data collection and management at enhanced lagoons for wastewater treatment

Mémoire

Queralt Plana Puig

Maîtrise en génie des eaux
Maître ès sciences (M.Sc.)

Québec, Canada

Automated data collection and management at enhanced lagoons for wastewater treatment

Mémoire

Queralt Plana Puig

Sous la direction de:

Peter A. Vanrolleghem

Resumé

Les stations de mesure automatiques sont utilisées pour suivre et contrôler des usines de traitement des eaux usées. Ce suivi en continu à haute fréquence est devenu indispensable afin de réduire les impacts négatifs sur l'environnement car les caractéristiques de l'eau varient rapidement dans l'espace et dans le temps.

Toutefois, même s'il y a eu des progrès considérables, ces dernières années, de la technologie de mesure, les instruments sont encore chers. Aussi des problèmes de colmatage, d'encrassement ou de mauvaise calibration sont assez fréquents à cause du contact avec les eaux usées. La fiabilité des mesures en ligne et en continu est affectée négativement. Par conséquent, un bon entretien des instruments est essentiel, ainsi que la validation des données collectées, afin de détecter d'éventuelles valeurs aberrantes.

Dans le contexte de ce mémoire, en collaboration avec Bionest®, une méthodologie est proposée pour attaquer ces problèmes. Deux cas d'études en étangs aérés au Québec ont été considérés, avec l'objectif d'optimiser les activités d'entretien, de réduire les données non fiables et d'obtenir des grandes séries de données représentatives.

Abstract

Automated monitoring stations have been used to monitor and control wastewater treatment plants. Their capability to monitor at high frequency has become essential to reduce the negative impacts to the environment since the wastewater characteristics have an elevated spatial and time variability.

Over the last few years, the technology used to build these automatic monitoring stations, for example the sensors, has been improved. However, the instrumentation is still expensive. Also, in wastewater uses, basic problems like fouling, bad calibration or clogging are frequently affecting the reliability of the continuous on-line measurements. Thus, a good maintenance of the instruments, as well as a validation of the collected data to detect faults is required.

In the context of this thesis, in collaboration with Bionest®, a methodology has been developed to deal with these problems for two facultative/aerated lagoon case studies in Québec, with the objective of optimizing the maintenance activities, of reducing the fraction of unreliable data and of obtaining large representative data series.

Contents

List of Tables

List of Figures

List of Abbreviations

AC Alternating current

ADQATs Automatic data quality assessment tools

AMS Automated monitoring stations

APHA American Public Health Association

API Application programming interface

AR Autoregression

ASTM American Society of Testing and Materials

ATV-DVWK Abwassertechnischen Vereinigung - Deutscher Verband für Wasserwirtschaft und Kulturbau

BDSO Banque de données des statistiques officielles

BOD Biochemical oxygen demand

COD Chemical oxygen demand

CODs Soluble chemical oxygen demand

CUAHSI Consortium of Universities for the Advancement of Hydrologic Science, Inc.

CWA Clean Water Act

DO Dissolved oxygen

EC Environment Canada

EPA U.S. Environmental Protection Agency

EPA STORET U.S. Environmental Protection Agency, Storage and Retrieval

EU European Union

F/ALs Facultative/aerated lagoons

GEKO Gerätekontrolle

GEMStat Global Environment Monitoring System water quality database

GFCI Ground fault circuit interrupter

GUI Graphical user interface

IEEE Institute of Electrical and Electronics Engineers

ISO International Organization of Standardization

IWA International Water Association

LCL Lower control limit

LDO Luminescent dissolved oxygen

LED Light-emitting diode

LIMS Laboratory Information Management System

LWL Lower warning limit

MDDELCC Ministère du Développement durable, de l'Environnement et de la Lutte contre les changements climatiques

model*EAU* Canada Research Chair on Water Quality Modeling

mon*EAU* Automated Monitoring Station

N Nitrogen

NMKL Nordisk Metodikkomité for Næringsmidle

NTC Negative Temperature Coefficient

NTU Nephelometric turbidity unit

ODM Observations Data Model

R R programming language

RMSE Root-mean-square error

RSM30 River Side Monitoring System

SQL Structured Query Language

STORET Storage and Retrieval

TSS Total suspended solids

UCL Upper control limit

UPS Uninterruptible power supply

USGS U.S. Geological Survey

USGS NWIS U.S. Geological Survey, National Water Information System

UWL Upper warning limit

UV-VIS Ultraviolet-visible spectroscopy

WHO World Health Organization

WSDOT Washington State Department of Transportation

WWTP Wastewater treatment plant

Acknowledgments

During my studies in Québec City, I had the opportunity of meeting great people who have helped me in the past two years.

First of all, I wish to express my appreciation to my supervisor, Prof. Peter A. Vanrolleghem. Thank you to give me the opportunity to come back to Québec City. Your vast knowledge and experience helped me to get through my project.

I cannot forget Janelcy, a former postdoc, whom I have been working with. My sincere gratitude for your patience and help.

Also, I have to thank the Bionest company and its employees, especially Étienne and Coralie, to be supportive during my project.

And obviously, I have to mention the modelEAU team (both current and former members), it is always a pleasure to work with a hard-working team. It hasn't just been work, it has been a life experience... So, special thanks to all of you for everything.

In addition, I acknowledge the financial support of MITACS, to fund the project and my scholarship.

Finally, I want to thank my family, my parents and my brother. Even if you are far, you are close. Thank you for your support and encouragements.

Chapter 1

Introduction

Monitoring and control of wastewater treatment plants (WWTP) have become more important in the last few years to reduce the negative impacts and contamination of freshwater bodies, i.e. the control of the water quality of the effluent has received increased attention (Vanrolleghem and Lee, 2003). This special interest has been translated in regulations, such as the 91/271/EEC Directive in the EU, the CWA in the USA, and more recently in Canada, the SOR/2012-139.

In view of complying with these regulations, and considering that wastewater characteristics have an important spatial and time variability, automatic on-line monitoring of several parameters to determine the quality of the wastewater is evidently required (Bourgeois et al., 2001). Besides, the technology to satisfy these needs, like the use of sensors, has been improved with the target to obtain reliable continuous measurements, optimizing maintenance and costs of operation (Lynggaard-Jensen, 1999).

Sensors are essential to monitor and control WWTPs, providing on-line data (Jeppsson et al., 2002). For this, automated monitoring stations have been presented as the next generation of water quality monitoring systems. Rieger and Vanrolleghem (2008) proposed the monEAU concept (monitoring of water, "eau" in French) as one of these new monitoring systems, combining state-of-the-art technology and with the visions to provide flexibility to connect any sensor, to be installed anywhere and to be used for different monitoring purposes.

However, the new sensors are still quite expensive and they also still cause problems. Harremoës et al. (1993) identified these basic problems as fouling, bad calibration and reliability of the sensors. Over the last few years, the performance and the reliability of the sensors has improved, but there are still not a lot of water professionals that feel sufficiently confident to fully rely on them. Moreover, the current WWTPs are often not prepared for automated control systems or to take advantage of the on-line information (Campisano et al., 2013; Jeppsson et al., 2002).

These problems are present and common for any type of on-line sensor used in the WWTP chain due to anomalies in sensors, insufficient maintenance, severe environmental conditions or other external factors. Therefore, good maintenance to minimize these unreliable data remains a challenge, and so does the management of the high quantity of data without loosing their quality.

In Canada, 67 % of the WWTPs are aerated lagoon systems (Lafond, 2009). These aerated lagoons operate at a high hydraulic retention time (around 5 days) compared to conventional WWTP and they are designed to remove total suspended solids (TSS) and reduce the bio-chemical oxygen demand (BOD) (Metcalf and Eddy, 2003). However, in July 2012, a new regulation, named SOR/2012-139, was launched that is more restrictive for TSS and BOD removal and also imposes nitrogen removal when unionized ammonia exceeds 1.25 mg N/l (Government of Canada, 2012). Thus, there is a need to upgrade the current lagoons.

Bionest Technologies has developed a new technology, $KAMAK^{TM}$ (meaning "living lake" in Atikamekw), to tackle this necessity. With the objective to allow for lagoons with reduced hydraulic retention time, promote nitrogen removal and prevent the wash-out of nitrifiers, this technology consists of the integration of fixed-media in a lagoon subdivided in baffled reactors and clarification zones.

With the installation of the $KAMAK^{TM}$ media in aerated lagoons, the process becomes more intense, the reactions and the dynamics are faster, and the installation can suffer biofilm sloughing. To determine these variations, grab samples are not enough to reach a detailed understanding of the system. Thus, monitoring stations were installed to continuously collect data allowing the study of the system dynamics.

According to the monEAU vision given above, this thesis presents a methodology to deal with basic problems in getting reliable and useful data for aerated lagoon WWTP, including:

- How to maintain and keep the monitoring stations working properly
- How to handle data quality problems to provide the required or useful information
- How to manage the huge amount of data collected

Chapter 2

Literature review

Anthropogenic activities have a big influence on the environment, and during the past few decades, it has been increased due to the development of large urbanized areas and the industrial revolution and technological progress (Harremoës et al., 1993; Chapman, 1996). This effect can be translated as pollution to terrestrial ecosystems, freshwater and marine environments and the atmosphere.

Environmental monitoring concerns collecting and analyzing information on the state of the environment to identify any variation or trends over time (EC, 2012; Corbitt, 1990). Monitoring also comprises verifying whether policies and programs are having the desired results and activities are according to legislation.

Within the field of water, water quality deterioration has become a problem since it is a finite resource and fundamental to life. Besides, water pollution threatens development projects and makes wastewater treatment essential (Bartram and Ballance, 1996; Chapman, 1996). Therefore, there is a need to comprehend and assess water quality (Quevauviller et al., 2007). For that, reliable monitoring data is fundamental.

In this chapter, a water quality monitoring literature overview is presented and in particular its issues and challenges specially in case of wastewater treatment.

2.1 Need for water quality monitoring

Some WWTPs experience process disturbances, faulty design, overloading and inadequately trained operators (Vanrolleghem, 1994; Metcalf and Eddy, 2003). These facts affect the effluent quality and may lead to violate discharge limitations. However, the negative impacts to the environment can be reduced by incorporating monitoring and control strategies into the wastewater treatment operation. These monitoring activities can be conducted for many purposes. In case of a WWTP, EPA (2012a) and (Bartram and Ballance, 1996) propose the main purposes as follows:

- determine whether the WWTP is working in compliance with pollution regulations
- characterize waters along the WWTP and identify changes or trends in water quality over time
- identify current and forthcoming contamination problems in views of hazards prevention
- gather information to design a specific treatment program or to optimize WWTP operation
- respond to emergencies, such as floods or spills

Basically, proper control may allow achieving an optimum performance and avoiding operational problems (EPA, 1977; Olsson et al., 1998). Furthermore, it is known that wastewater is a huge problem for the environment thus, regulations have been launched to reduce its negative impacts and deterioration by the deleterious substances discharged by the WWTP effluents in freshwaters (Lovett et al., 2007; Government of Canada, 2012).

Existent regulations, like the 91/271/EEC Directive in the EU, the CWA in the USA, and the SOR/2012-139 in Canada, indicate the basic monitoring requirements of the WWTP depending on treatment capacity. The Council of the European Union (1991), the Government of Canada (2012) and the EPA (2015) mentioned that a monitoring plan has to ensure:

- an appropriate monitoring program to control some parameters (physical, chemical or biological)
- a proper analysis of samples by using standard methods
- an adequate frequency of monitoring to:
  - monitor discharges from wastewater treatment plants
  - assess the amounts and compositions of the wasted sludge

Summing up, monitoring and control of WWTPs have become essential to track and improve the performance of the system and to ensure that the water quality of the effluent agrees with the policies and regulations. Moreover, quality measurements are essential to make the correct decisions related to management of water resources, monitoring issues, biological quality, etc. (Quevauviller et al., 2007).

2.2 Water quality monitoring

Water quality monitoring is the activity to collect information about various water characteristics at determined points with a certain frequency (Bartram and Ballance, 1996; Chapman, 1996). Subsequently, the obtained information can be used to evaluate the physical, chemical and biological status of the water body as well as to understand its dynamics, identify pollution problems, establish trends and make necessary decisions (WHO, 1963). This global procedure is known as water quality assessment.

Generally, two different basic functions are established for water quality monitoring according to Dandy and Moore (1979); Karpuzcu et al. (1987): prevention and abatement. The objective of the prevention is to maintain the existing unpolluted or acceptable status of water quality, while, the goal of abatement is to control the system and reduce its pollution conditions (Harmancioglu et al., 1999).

Besides, behind the monitoring concept, it is possible to distinguish three different types of monitoring activities. Depending on long-term, short-term and continuous programs, Bartram and Ballance (1996); Chapman (1996) define these procedures as follows:

- *Monitoring* is a long-term activity, measurement and observation of the water quality for a specific purpose.
- *Surveys* are finite duration activities, intensive programs to measure and observe the quality of the water body for a specific purpose.
- *Surveillance* is a continuous activity with specific measurement and observation for the purpose of water quality management and operational activities.

In the specific case of WWTP, surveillance is the best practice to estimate the quality of the wastewater (Vanrolleghem, 2010). In particular, this type of monitoring permits to determine the status of the water quality, to assess the impact of long time changes, and to evaluate whether the WWPT operates according to the regulation's objectives (EC, 2003).

Furthermore, depending on the objectives of the measurements, two different approaches of water quality monitoring can be carried out in a WWTP (Thomas and Pouet, 2005):

- off-site measurements based on sampling and laboratory analysis
- on-site measurements with on-line measurement systems or field portable devices

According to the main needs for monitoring presented in section 2.1, Thomas and Pouet (2005) propose that the choice of appropriate measurement approach should be made following table 2.1.

Table 2.1: Criteria supporting the choice of monitoring approach in agreement with the needs (Thomas and Pouet, 2005).

	Off-site	On-site
Regulation compliance	✓	
Process control		✓
Hazards prevention		✓
Scientific knowledge	✓	✓

Normally, both monitoring approaches are combined to accomplish all needs presented. However, it seems that in case of wastewater quality monitoring, on-site measurements are preferable to better estimate the wastewater quality throughout time and space.

On-site and off-site measurements can be classified in three different types (Figure 2.1). For the in situ type, sensors are placed inside the water to be monitored (e.g. Beaupré (2010) and Kaelin et al. (2008)). The second type, the on-line configuration, the wastewater is pumped through a hydraulic loop in which the sensors are installed (e.g. van Griensven et al. (2000)).The third type is the simple measurement from a sample analyzed with portable sensors or standard methods in a lab (e.g. Berthouex and Hunter (1975)).

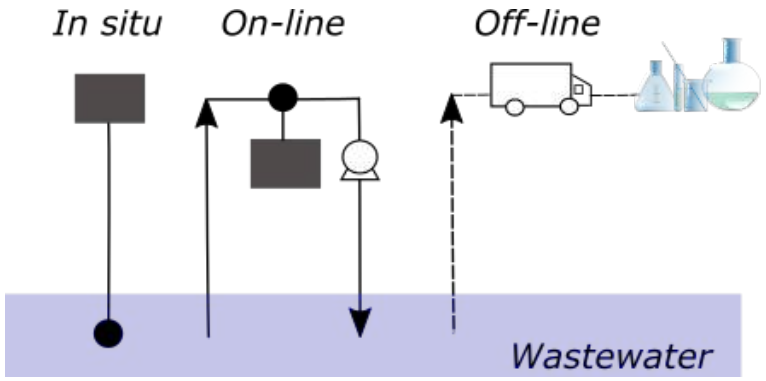


Figure 2.1: Schema of the different measurement types (Rieger, personal communication).

2.2.1 Monitoring program

As several regulations demand, a monitoring program has to be conducted to assess the discharges from WWTP, including such information as their flow and composition (Vanrolleghem and Lee, 2003; Jeppsson et al., 2002).

One of the main purposes of a monitoring program is to collect sufficient *good quality* data to evaluate spatial and/or temporal variations in water quality. Along with achieving a successful program to produce the expected information, Harmancioglu et al. (1998); Chapman (1996) propose some basic stages to be followed:

- Definition of the objectives of the monitoring
- Preliminary surveys and background information
- Monitoring design: location, parameters to measure, timing and frequency of monitoring, etc.
- Monitoring including field monitoring, hydrological monitoring and laboratory activities
- Data quality control
- Data storage, treatment and reporting
- Data interpretation
- Water management recommendations and decision making

Even if there is not a simple specific model for the structure of a monitoring program, it has to be consistent. Also, Chapman (1996) suggests that a program should be flexible to meet short-term objectives with the possibility of the program to be modified over longer periods to accommodate for new interests and priorities. It is recommended to review and update the program regularly according to the current needs. Moreover, once the program is prepared, it should be communicated to all participants (Bartram and Ballance, 1996). If the participants are not completely informed about the updates, the quality of the program will suffer.

2.2.2 On-line water quality monitoring

Water quality can change regularly over time and space, and to accurately characterize these variations of the water conditions, frequent and repeated measurements on-site are needed. According to Wagner et al. (2000), when the time between repeated measurements is sufficiently small[1], the resulting water quality record can be considered continuous.

In the context of continuous monitoring, when data collection is happening in real-time, this type of monitoring is also called on-line monitoring.

In this case of on-line water quality monitoring, multiple water quality probes and sensors configured into monitoring stations are used. They are placed throughout the wastewater treatment system and transmit usable information automatically and continuously in real-time. These stations are known as automated monitoring stations (AMS) and they are described in more detail in section 3.2.1.

[1]The sampling frequency is considered sufficiently small when this sampling rate is at least twice of the frequency (Nyquist sampling theorem) which can be able to full reconstruct water quality changes (Olshausen, 2000).

2.2.3 Issues of on-line water quality monitoring

On-line monitoring programs have some challenges and associated limitations. A good monitoring program is not enough to obtain the desired and useful information about the wastewater conditions. It is also necessary to choose carefully among the available technologies and select the location with care (Dippenaar et al., 2000).

Due to the adverse environment in which they are installed, automated monitoring in wastewater systems is not used as much as in other processes. Rieger and Vanrolleghem (2008) present the basic limitations of the use of AMS: the lack of standardization, data quality problems, and insufficient flexibility of the stations.

Bourgeois et al. (2001) and Campisano et al. (2013) agree that the existing equipment is not well prepared for long-term on-line monitoring in a fouling environment (e.g. figures 2.2a and 2.2b). Thus, the instrumentation requires additional equipment for cleaning, regularly maintenance and the implementation of data fault detection methods. All this adds to the cost and leads to extra work.

(a) Matter accumulation attached on a sensor's cable.

(b) Matter accumulation on a sensor.

Figure 2.2: Examples of the fouling environment.

In contrast to grab or composite samples, automatic data acquisition systems are producing a huge amount of data with uncertain quality. Since manual data validation requires time and is very tedious, automatic data quality assessment tools (ADQATs) are necessary to validate the time series to be useful (Alferes et al., 2012). In that sense, poor quality data could drastically affect the results of their application, namely water quality models, WWTP control actions, decisions regarding design and operation, etc.

Summing up, the challenge associated with automated monitoring programs is to collect data that consistently represent water quality. This requires clear planning and protocols for data collection, and quality assurance and quality control. The quality of data collected by auto-

mated monitoring equipment is dependent on the methods used to handle, configure, operate, and maintain the equipment. For each station field procedures must be standardized and documented.

2.2.4 mon*EAU* project

mon*EAU* (monitoring of water, "eau" in French) is presented as the next generation of water quality monitoring networks. According to the limitations presented in section 2.2.3, Rieger and Vanrolleghem (2008) present the new vision of AMS with the following characteristics to be considered for its design:

- *A Flexible System.* A multi-objective monitoring network to be used for different purposes, wherever it is desired, with any type of sensor.
- *An Open and Modular System.* A system that can be adapted to the current needs, keeping the integrity of the robust framework.
- *A High Quality/Performance Database.* A platform to store the large data series is needed. Like the monitoring system, the database has to be robust and flexible enough to be adapted to any station application.
- *Remote Use.* Sometimes the access to the stations is not easy. The design of the system has to consider the minimization of maintenance requirements and energy demand, and remote access to the entire monitoring station including the sensors.
- *Automatic Data Quality Assessment.* Data evaluation is to be done with reference samples, sensor status/diagnosis data and time series information.
- *User-Friendly and User-Oriented Software Concept.* The required information is provided and visualized depending on the user level and the location.
- *Proactive and Flexible Maintenance Concept.* A protocol on maintenance and a schedule for the operators has to be provided. This furnished information is based on sensor self-diagnosis, the company or user experience and a proactive set of station-triggered experiments.

In this context, previous studies have been been conducted by the model*EAU* research group dealing with some of these characteristics. These works focused on sensor characterization and validation (Beaupré, 2010), data collection and management (Plana, 2013), and automatic data quality evaluation (Poirier, 2015; Saberi, 2015).

2.3 Data validation

On-line continuous monitoring is used to collect data at high frequency. However, the collected data can be beneficial only if it is accessible, accurate and reliable (Copp et al., 2010). This is the reason why an efficient monitoring system has to assure a good quality control and quality assessment (Alferes et al., 2013).

In case of a WWTP, due to the extreme environment where the sensors are put in, the raw data present some problems such as noise, missing values or systematic errors (Mourad and Bertrand-Krajewski, 2002). Detecting and replacing doubtful or wrong data is essential to avoid wrong decisions at the level of control, process modeling or planning of new treatment infrastructure (Thomann, 2008; Branisavljevic et al., 2010). Some data validation procedures have to be implemented before further use of the data.

Generally, the control of on-line sensors is done by comparing the on-line measurement data and and the values obtained from grab samples analyzed by reference methods (ATV-DVWK, 2000; Häck et al., 1999). This allows detecting out-of-control situations like drift, shift or outlier situations (See figure 2.3) (Thomann et al., 2002).

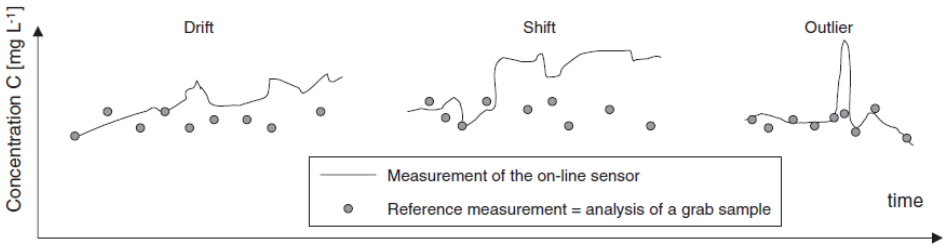


Figure 2.3: Out-of-control situations (Thomann et al., 2002)

Despite the fact that this methodology is often used and allows to verify the reliability of the data, it just permits to compare one single value of the situation at a determined moment. Thus, ADQATs for continuous validation will be more efficient and reliable to detect any bias (Rieger et al., 2004).

An example of a dynamic validation is the GEKO (*GEraeteKOntrolle*) monitoring software presented by Rieger et al. (2004). This software has been designed to meet an efficient monitoring by minimizing laboratory analysis and operator efforts, and detecting systematic errors as fast as possible.

Following the same idea, Alferes et al. (2013) proposed a combination of univariate methods as a dynamic validation to detect faults and outliers in the context of the mon*EAU* project. This methodology has been used in this thesis and is presented in more detail in section 3.3.

2.4 Water quality databases

On-line water quality monitoring generates large data sets. Thus, the tasks of storage, analysis and the interpretation of the collected data are crucial. Washington State Department of Transportation assure that the interpretation and evaluation of the data cannot be performed without good data storage and analysis (WSDOT, 2008).

To carry out these activities successfully, a water quality database has to be built (Peng et al., 2011). Such database should provide the simplicity to store and document all relevant data, and should be easy to use for further data evaluation (Rieger et al., 2004). However, to obtain a long-term database, there are some data management challenges. Camhy et al. (2012) and Holmes and Poole (1998) identify the following challenges:

- The variability of the collected raw data formats
- The database is continually growing and must be adaptable according the modifications of the monitoring program
- Data store and collection have to be adapted over time corresponding to the needs of high-performance storage and data access
- Personnel who is collecting and managing data is changing over time which can lead to some inconsistencies
- Archiving and documenting data sets like raw or validated data, as well as their meta-information (data about data)

Especially, the latter challenge is critical because without the storage of these metadata, the database has no meaning. Ideally, a database should answer at least the following questions:

- What has been measured?
- When has the value been measured?
- Where has the value been measured?
- How has the value been measured?
- Who has collected the value?
- Why has the value been measured?

Over the last few years, some water quality databases have been developed trying to tackle these challenges, e.g. the STORET database from EPA (2012b), the NWIS database from

USGS (2012), the GEMStat database from UNEP (2015) and the ODM database from CUAHSI (2015).

However, the storage and access to the metadata is still missing (Copp et al., 2010). Some authors, Camhy et al. (2012), Copp et al. (2010) and Rieger and Vanrolleghem (2008), affirm that metadata is a basic need for data evaluation and understanding, as well as for further studies and data exchange.

Furthermore, most of the published databases, in their design and in the way the metadata are stored, focus on measurement details, giving priority to data collection activities and data set characteristics rather than monitoring programs and locations (Sheldon et al., 2011).

Within the mon*EAU* concept, a robust, fast and flexible database is needed. Previously, Plana (2013) developed a first database to manage and store the data. Named dat*EAU*base, it consists in two different parts:

- A relational database built with MS Access.
- A user interface and tools to facilitate data entry, exporting and viewing data from the database, built with MS Excel and the R programming language.

The design of the dat*EAU*base deals with the different challenges presented above (See figure 2.4). It provides uniformity of the data format, sufficient flexibility for further modifications, relevant metadata storage, and documentation according to the monitoring program. However, the used softwares had several limitations with reference to their capacity.

2.5 Facultative aerated lagoons for wastewater treatment

After activated sludge, lagoon WWTPs are the second most popular natural system. For instance, in Canada, it represents 67 % of the number of WWTPs (Lafond, 2009). Generally, they are used to stabilize biodegradable organic matter (Grady et al., 2011). In some occasions, nitrogen and phosphorous removal can be observed but it occurs only in warm climates (Metcalf and Eddy, 2003).

The design of lagoons is simple, so low investment and operating costs are required. However, the understanding of the chemical and biochemical reactions is limited due to their complexity (Grady et al., 2011). Due to this complexity, lagoons sometimes present several issues such as a poor quality effluent and excess algae growth.

Typically, the lagoon is a basin excavated in the soil with sloped sidewalls (See figure 2.5).

Depending on the lagoon characteristics, such as depth and biological reactions, the system can be classified in three different types: aerobic, facultative/aerated and anaerobic (Grady

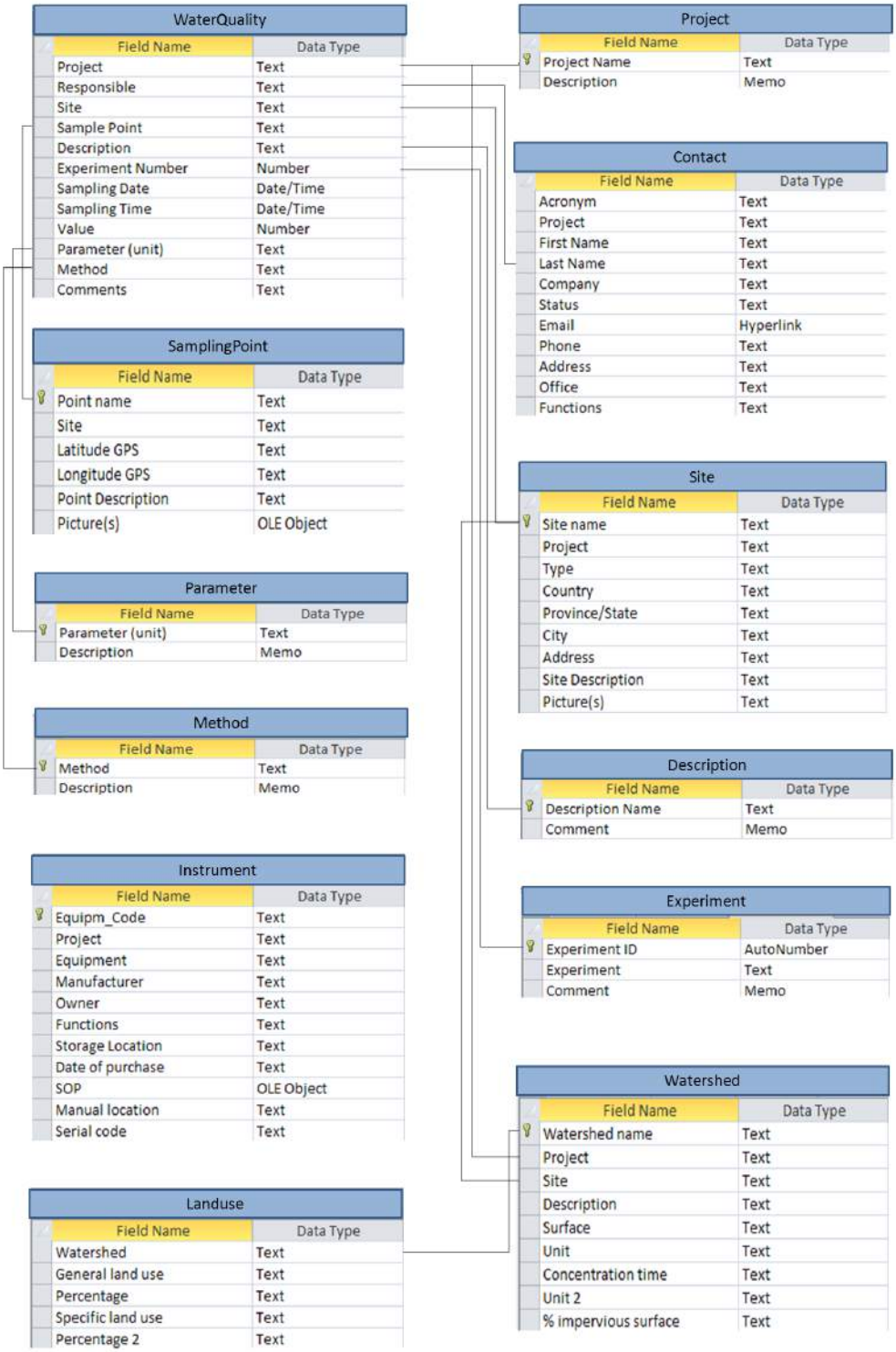


Figure 2.4: First dat*EAU*base model developed by Plana (2013).

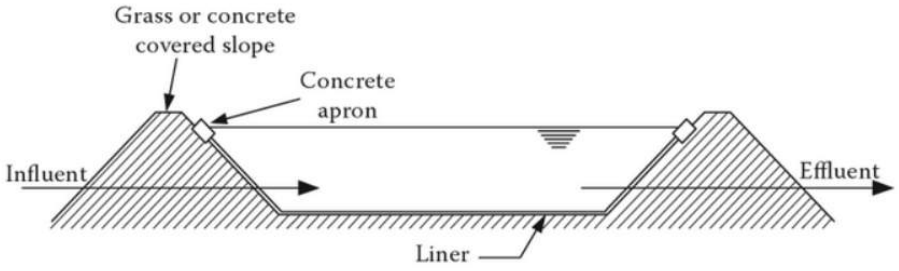
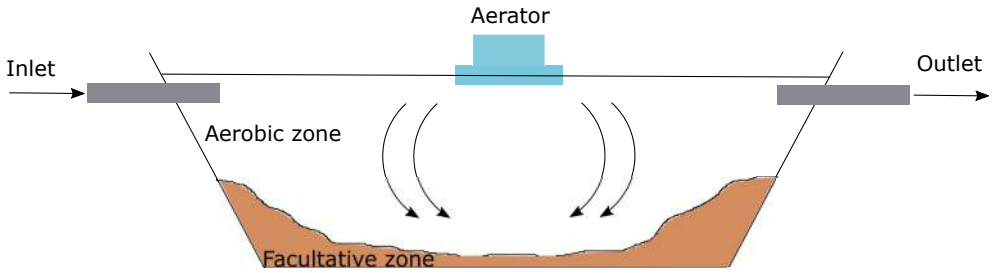


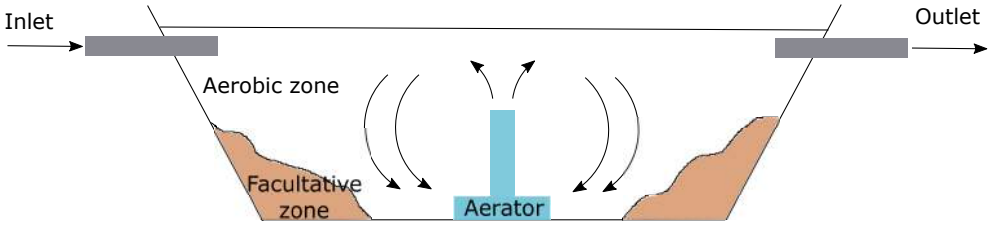
Figure 2.5: General schema of a lagoon (Grady et al., 2011).

et al., 2011; Vesilind, 2003). In the context of this thesis, partially mixed aerated lagoons, named facultative aerated lagoons (F/ALs), are studied.

In this system, aerators are installed at the bottom or on the surface to oxygenate the lagoon. However, in F/ALs, the provided air is insufficient to keep the lagoon completely mixed (Grady et al., 2011; MDDELCC, 2001). Thus, the biodegradable organic matter is stabilized by both aerobic and anaerobic processes (See figure 2.6a). Part of the suspended solids settle to the bottom of the basin where anaerobic digestion is occurring. The other part of the suspended solids and the soluble organic matter are oxidized in the aerobic zones (MDDELCC, 2001).



(a) Schema of a F/ALs with floating surface aerators.



(b) Schema of a F/ALs with submerged diffused aeration.

Figure 2.6: Schemas of a facultative/aerated lagoons.

F/ALs systems are able to remove from 50 to 200 mg/L of TSS and the overall BOD removal rate is from 0.5 to 0.8 d^{-1} (Metcalf and Eddy, 2003). The hydraulic retention time is typically between 4 and 10 days.

In the province of Québec, 37 % of the 540 F/ALs are overloaded (BDSO, 2014). Moreover, since these systems are installed outside and the retention time is long (around 20 days in Québec, (MDDELCC, 2001)), under winter conditions, the organic matter degradation becomes slower and the effluent load is higher. Thus, overloaded lagoons have to be upgraded, e.g. their treatment capacity needs to be increased.

To increase the treatment capacity of a F/ALs system, different possibilities exist (Gerardi and Lytle, 2015; Metcalf and Eddy, 2003):

- Change the lagoon's hydraulics. A completely mixed system or the addition of baffles are the most common processes to reduce dead zones and decrease hydraulic retention time. However, wash-out of the biomass can occur as a result and more energy is needed.
- Integration of fixed-film. Adding inert media, on which biofilm develops, prevents the wash-out of the nitrifiers and a complete nitrification and CBOD removal can be achieved. And as a result more energy is needed.

Bionest Technologies has developed a fixed-film media, $KAMAK^{TM}$, to augment the treatment capacity of F/ALs reducing the hydraulic retention time, promoting nitrogen removal and avoiding biofilm sloughing even under the winter conditions occurring in Québec.

2.6 Problem statement

In the context of this thesis, the following research challenges have been identified:

- *Measurement in raw wastewater*: at the inlet of the wastewater system, clogging and fouling is typical due to the high pollution load of the water. It produces the hardest measurement conditions and it affects the reliability of the measurements.
- *Environmental conditions*: since the study concerns continuous monitoring, the installation has to be prepared for dry and wet weather conditions, large temperature changes and seasonal loads. In particular, for the use of sensors under winter conditions and considering that the lagoon system is outside with long retention times, proper installation will be critical to keep the sensors ice-free and to ensure a suitable temperature in the whole monitoring system. Also, it is important to guarantee safe access to the technicians and researchers for maintenance.
- *Optimize the laboratory analyses to make it cost-time effective*: Reduce the number of analyses without jeopardizing validation of the sensors
- *Data management and storage*: Due to the huge amount of collected data, there is a lack of tools to manage and store the data adequately. The existing databases do

not have enough capacity and their design do not allow complete, organized storage of the metadata. Furthermore, this lack of useful tools affects the accessibility to the information and reduces the use of the data for further studies.

2.7 Objectives

According to the mon*EAU* research line and the issues of water quality monitoring, the main objective of this thesis is to combine and improve upon the previous studies performed by the model*EAU* research group by Beaupré (2010), Plana (2013), Poirier (2015) and Saberi (2015) to assure a good operation of the monitoring stations in WWTP followed by robust procedures of data validation.

Furthermore, in the context of this thesis, a new application scenario has been studied: F/ALs under really cold conditions where two monitoring stations are installed to evaluate the system performance, one at the inlet and another one at the outlet of the F/ALs. Thus, the approaches developed so far had to be adapted.

To minimize unreliable data due to sensor failure, lack of maintenance, environmental conditions or any other factor, the specific objectives are translated into:

- Collect and interpret daily, weekly and yearly *good* water quality time series, which entails:
  - Designing and testing monitoring stations working on the new F/ALs scenario under really cold conditions
  - Establishing a maintenance procedure of the monitoring stations
  - Establishing on-line data validation procedures (on the new F/ALs scenario)

Moreover, to sustain easy management and use of the large amounts of collected data and metadata and to achieve the goals presented above, the design and the further development of the dat*EAU*base developed by Plana (2013) has been included as another objective.

Chapter 3

Materials and Methods

This section will describe the case study site, the materials that make up the stations to collect the water quality data, as well as all methods used to obtain and store quality data.

3.1 Site description

As mentioned in section 2.3, this thesis has focused on an enhanced lagoon wastewater treatment scenario. Below, the two case studies are presented.

3.1.1 Wemotaci

The first implementation was located in a reserve called Wemotaci on the north shore of the Saint-Maurice river in the Mauricie region of Québec, Canada (Coordinates: 47°54'25"N 73°47'00"W) (Figure 3.1).

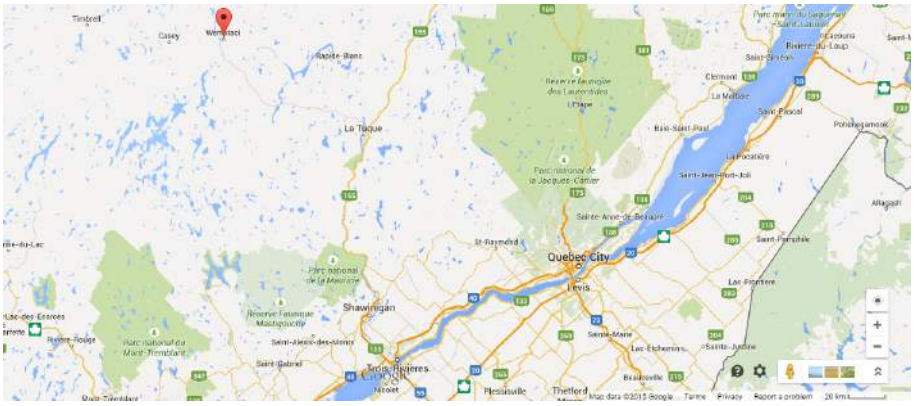


Figure 3.1: Location of the Wemotaci case study site.

The sewer system in Wemotaci is combined. All wastewater going to the WWTP is residential. And its characteristics agrees with the typical untreated domestic wastewater composition

proposed by Metcalf and Eddy (2003): 120 - 400 mg/l of TSS and 110 - 350 mg/l of DBO_5.

The WWTP consists in two aerated lagoons in series treating a maximum daily flow of 866 m^3/d (Figure 3.2). The corresponding volumes of each lagoon are 6525 m^3 for the biggest one, and 4352 m^3 for the smallest one. Thus, the retention time of these lagoons is 12.5 days. On both lagoons, aerators are installed at the bottom.

Specifically, the study was located at the smallest one (see figure 3.2). This lagoon was divided in two halves by a watertight membrane to compare a standard lagoon and a lagoon enhanced with BIONEST media (see section 2.5). To evaluate the performance of the system, two monitoring stations were installed, one on the inlet and the other on the outlet (see figures 3.2 and 3.3).

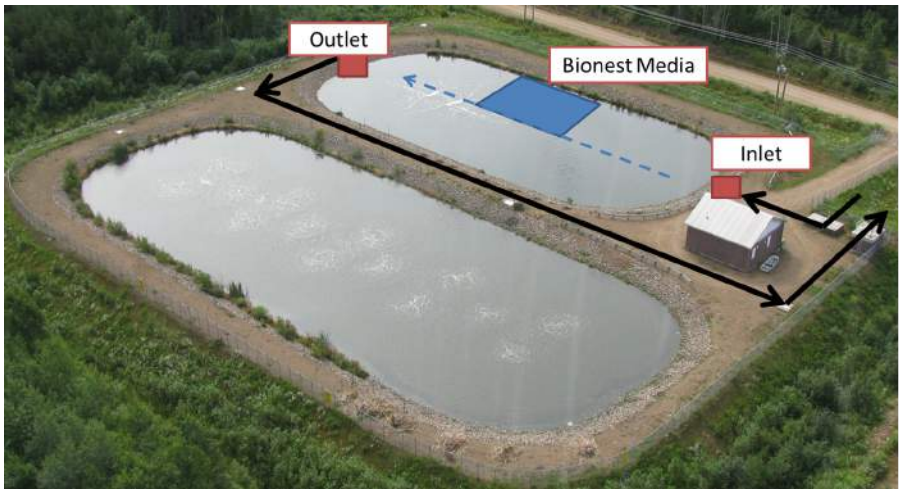


Figure 3.2: Location of the RSM30 stations (red squares) in the Bionest media (blue square) enhanced lagoons of Wemotaci.

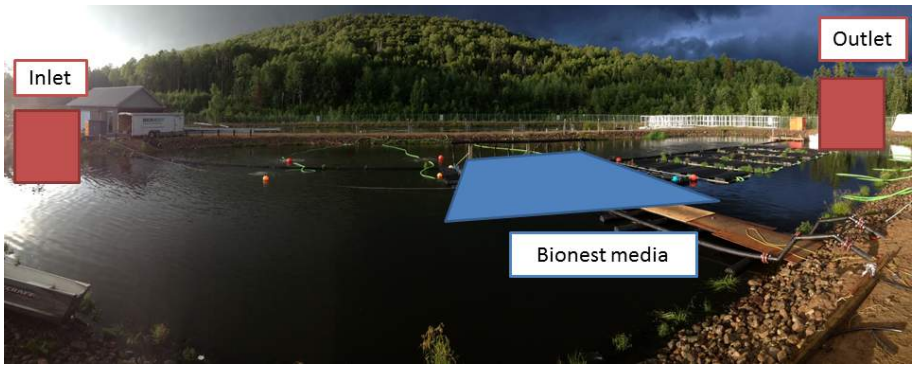


Figure 3.3: Wide angle picture of the enhanced lagoon of Wemotaci.

The experienced period of this installation was between September 2013 and April 2014, so it allowed studying the operation of the monitoring stations under cold climate conditions (e.g.

figure 3.4).

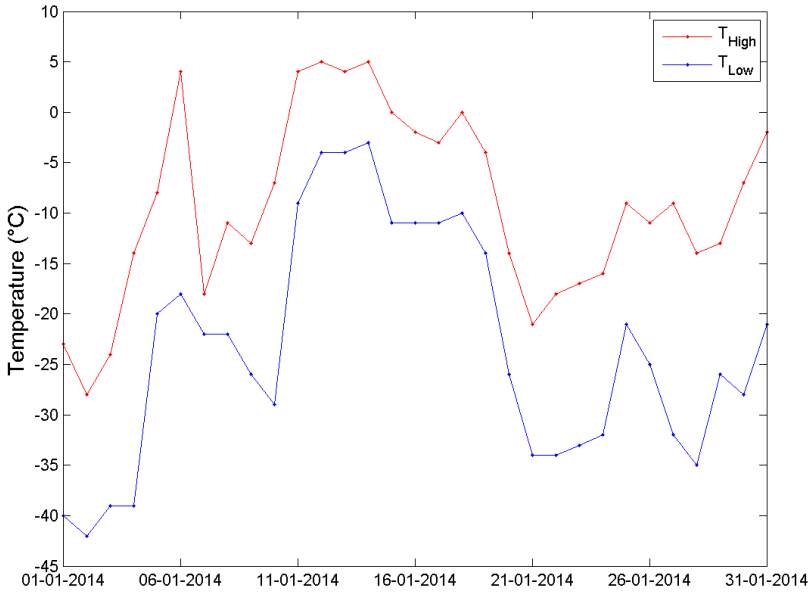


Figure 3.4: High and low daily air temperatures in Wemotaci during January 2014.

3.1.2 Grandes-Piles

The second case study was also located in the Mauricie region, more specifically in the regional county municipallity of Mékinac, called Grandes-Piles (Coordinates: 46°41'12"N 72°44'33"W) (Figure 3.5).

Figure 3.5: Location of the Grandes Piles case study site.

As in Wemotaci, the sewer system in Grandes-Piles is combined. And the wastewater origin is basically residential. Also its composition agrees with the standard municipal characteristics

proposed by Metcalf and Eddy (2003).

The WWTP consists in one circular concrete aerated lagoon separated by a baffle in two different cells with aerators installed at the bottom of the lagoon (see figure 3.6). Its total volume is 1564 m^3 and it treats 129 m^3/d. Thus, the retention time is about 12 days.

The Bionest media is installed in the middle of the first cell. The RSM30 stations were installed at the inlet and at the end of the media permitting to study its performance (see figure 3.6).

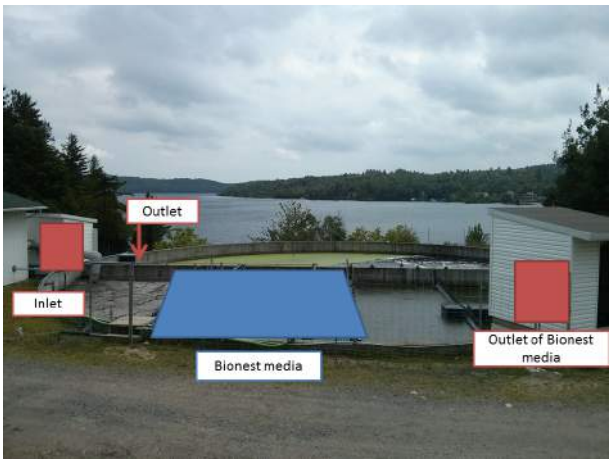


Figure 3.6: Location of the RSM30 stations (red squares) to study the performance of the Bionest media (blue square) in Grandes Piles.

This second installation was set up at the end of November 2014. It is still working to obtain a yearly series of data for further studies.

3.2 Installation of RSM30 monitoring stations

In this section, the equipment used for on-line monitoring is presented, as well as the maintenance required to keep the equipment working properly.

3.2.1 Automated monitoring stations

The applied RSM30 (Primodal) monitoring stations are a versatile design of water side monitoring equipment according to the monEAU vision (Rieger and Vanrolleghem, 2008): pursuing flexibility to be deployed wherever it is wanted, to connect different sensors from several manufacturers and to obtain the desired output.

The used monitoring stations are built as in figure 3.7. The main part of the installation includes sensors placed into the water (in situ measurements), and the station in which the sensors control and data storage are housed. The station is installed inside a cabin to protect it from extreme environmental conditions and vandalism.

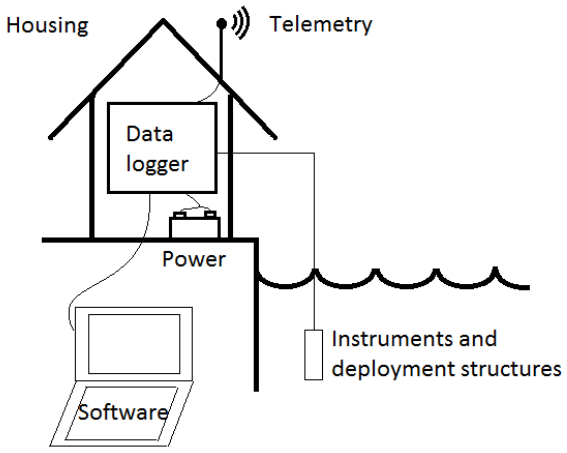


Figure 3.7: Automated monitoring station installation.

3.2.2 RSM30 station description

An example of a monitoring station is the equipment developed by Primodal Systems, named RSM30. The RSM30 is a water quality monitoring station that registers, transmits real-time data and analyses the data in real-time using its custom-designed PrecisionNow software (Primodal, 2015c).

The RSM30 is encased in a secure NEMA 4X rated fiberglass enclosure designed for a range of diverse environmental conditions and ease-of-deployment (Figure 3.8). The stations are equipped with a computer where the software to store the data runs, a controller, a router to provide remote access to the stations and a climate controller to mitigate extreme temperature fluctuations (Figure 4.5). Also, an uninterruptible power supply (UPS) is installed to provide emergency power when there is a short power failure. Moreover, equipment safety is assured through surge and GFCI (Ground Fault Circuit Interrupter) protection inside the unit.

Additionally, next to the temperature controller incorporated inside the RSM30, another temperature controller has been installed into the cabin, given the extreme cold climate conditions that the stations were exposed to during the winter.

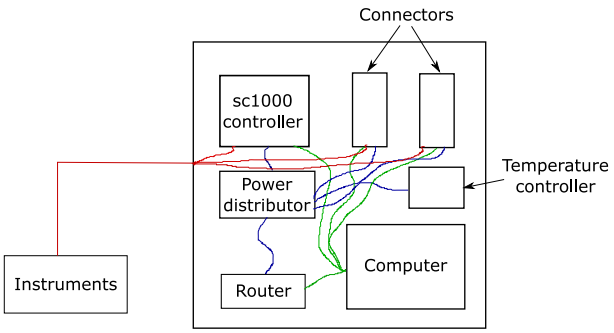
3.2.3 Softwares

The RSM30 system includes a robust software framework serving as the mainstay of the stations and server network permitting the simple connection of various modules through a specified API (Application Programming Interface) (Copp et al., 2010).

Some modules provide basic functionality like data input or output but the main purpose for this framework structure is the capability to incorporate new developments or to implement data evaluation modules (Rieger and Vanrolleghem, 2008; Copp et al., 2010). Thus, robust

Figure 3.8: Installed Primodal RSM30 cabin.

(a) Installed equipment inside the Primodal RSM30.



(b) Schema of the equipment inside the Primodal RSM30.

Figure 3.9: Equipment inside the Primodal RSM30 station.

operation is combined with the required flexibility. Figure 3.10 represents the monEAU concept.

In these stations, measurement, meta, LIMS (Laboratory Information Management System) and log data from sensors are controlled by Primodal's own PrecisionNow BaseStation software (Primodal, 2015c) and ana::pro software (Advanced Process Software) provided by the sensor manufacturer s::can (s::can, 2012).

PrecisionNow software

The PrecisionNow software is provided by Primodal (Hamilton, Canada) and is used to configure sensor inputs and data evaluation modules, to establish the communication with the

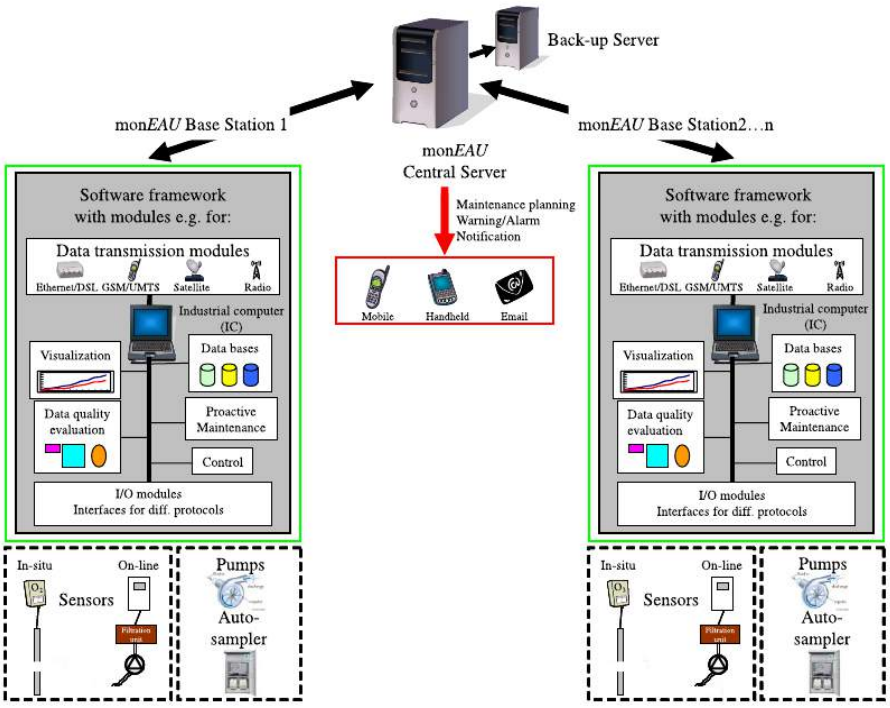


Figure 3.10: Set-up of the mon*EAU* water quality monitoring network (Rieger and Vanrolleghem, 2008).

central server and the data visualization system (Figure 3.11). Current and historical time series data can also be visualized inside the software as single or as multiple time series on one graph (see figures 3.12 and 3.13). Also, the collected data can be exported in different file formats such as text, xml, csv or Excel files (Primodal, 2015c).

This software provides flexibility to add, remove or change sensors rapidly and easily. Furthermore, sensors from different manufacturers can be configured on it (Primodal, 2015b).

Each Base Station can operate alone or as part of a monitoring network, as presented on figure 3.10. Every unit is fully equipped to operate in isolation, store and visualize measurements and meta- data, as well as it has the communication tools to transmit the data to a central storage location (Central Server) (Primodal, 2015b). Moreover, the RSM30 has sufficient capacity to store years of data.

The Central Server is programmed to retrieve data from any Base Stations within the network. This functionality permits an automated comparison of the data from multiple locations in real-time minimizing the effort needed for post-processing and the manual comparison of data.

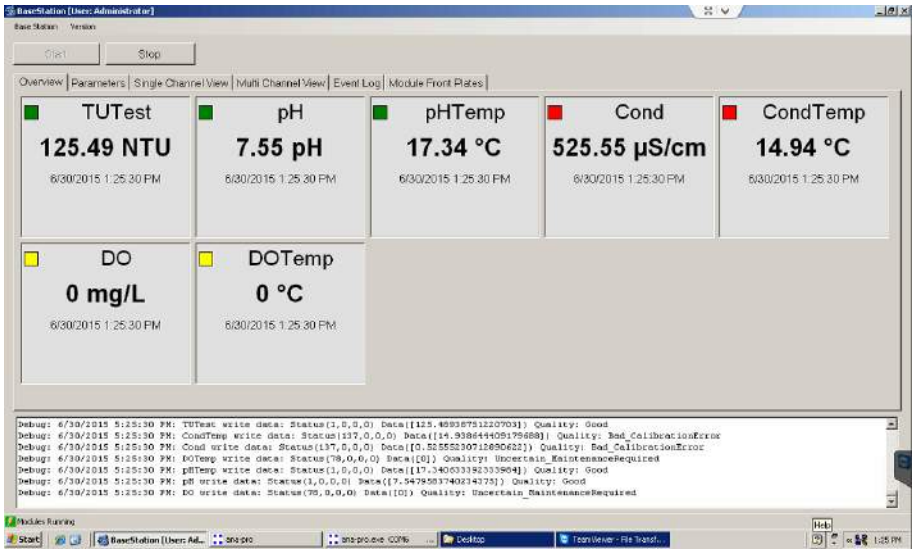


Figure 3.11: BaseStation software graphical user interface (Primodal, 2015a). The small squares in green indicate the sensor is working correctly, in yellow indicate a warning problem on the sensor, and the red indicate an error on the sensor.

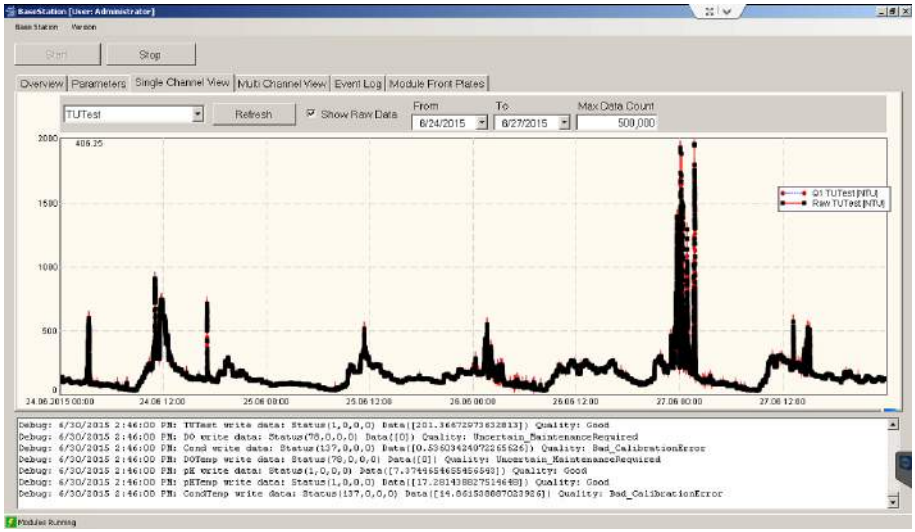


Figure 3.12: Influent TSS time series on a graph on BaseStation software interface (Primodal, 2015a). Data collected at Grandes-Piles.

ana::pro software

The ana::pro software was created by s::can (Wien, Austria) to offer numerical and graphical data and advanced multiparameter process visualization of s::can sensors (Figure 3.14). It offers advanced spectral analysis, derivative and delta spectra access, an autocalibration module, data-logger access, automatic or manual transfer of data, interpretation of measurements, off-line data analysis, interfaces for data transfer and automatic verification (s::can, 2006).

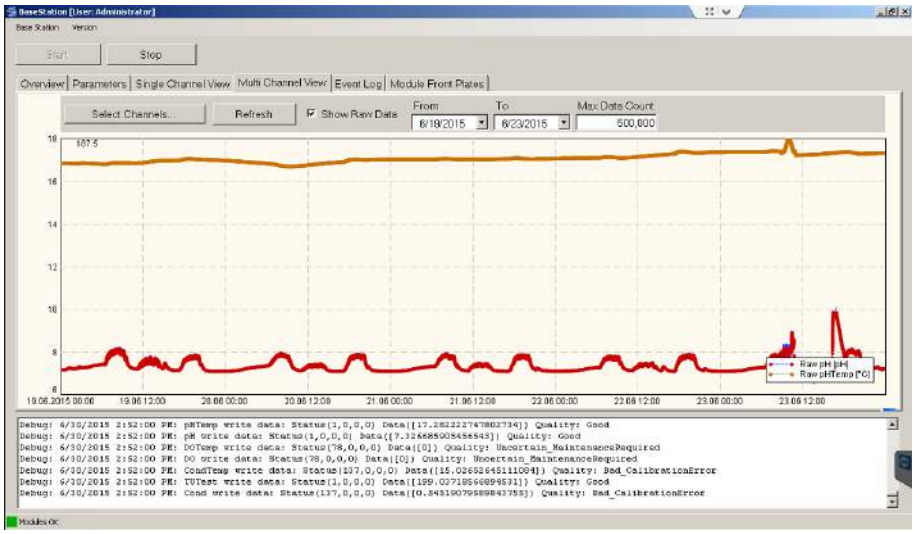


Figure 3.13: Influent temperature and pH time series on a graph on BaseStation software interface (Primodal, 2015a). Data collected at Grandes-Piles

In contrast to the PrecisionNow software, the ana::pro software is specially developed for the operation of all s::can probes, in particular spectrometer probes that require more complex applications. Additionally, it allows the use of the s::can dissolved oxygen probes, the s::can ammonium probe and other sensors distributed by s::can.

Figure 3.14: ana::pro software interface for two different s::can sensors (e.g. spectro::lyser and ammo::lyser) (s::can, 2006).

3.2.4 Installed sensors

A short overview of the sensors used in this implementation is presented below. Only two different brands of sensors have been used, even if the RSM30 system is flexible for all types of sensors.

Each station includes six sensors to measure the water quality parameters pH, conductivity, turbidity, ammonia, TSS, DO, temperature, nitrates and COD. These sensors and some characteristics are presented briefly in table 3.1.

pH sensor

The pH digital differential sensor is manufactured by Hach (Figure 3.15). The principle of operation of this pH sensor is to measure the concentration of protons as -log[H^+] by evaluating an electrical potential in mV between the glass electrode and the reference electrode, similar to the potential between the two plates of a capacitor (Hach, 2006b). Also, the glass electrode acts as a transducer that coverts chemical energy into electrical energy producing a potential proportional to the pH value. Moreover, to automatically compensate pH measurements for temperature variations, an integral NTC (Negative Temperature Coefficient) 300 ohm thermistor is installed.

Figure 3.15: Hach pH sensor used in the RSM30 stations.

The manufacturers do not provide a self-cleaning for the pH sensors. However, compressed air has been installed to prevent sensor's fouling due to the sever environment. In this application, the air is activated every 5 minutes.

Conductivity sensor

The conductivity sensor used is manufactured by Hach (Figure 3.16). The principle of the measurement of the inductive conductivity probe is made by passing an AC current through a toroidal drive coil which induces a current in the electrolyte solution. This induced solution current produces a current in a second toroidal coil. The amount of current induced in the second coil is proportional to the solution conductivity (Hach, 2008).

Neither or self-cleaning system nor compressed air are provided by the manufacturer or by the researchers for the conductivity sensor.

Table 3.1: Summary table of the used sensors.

Sensor	**Brand**	**Model**	**Parameter**	**Units**	**Range**	**Operating T**
pH	Hach [1]	DPD1R1	pH	-	-2 to 14	-5 °C to 70 °C
			Temperature	°C	-5 to 70	
Turbidity	Hach	LXV423.99.00100	Turbidity	NTU	0 to 4000	0 °C to 40 °C
			TSS	mg/L	0.001 to 500000	
Conductivity	Hach	3727E2T	Conductivity	μS/cm	0 to 2000000	-10 °C to 200 °C
			Temperature	°C	-10 to 200	
DO	Hach	5790000	DO	mg/L	0 to 20	0 °C to 50 °C
			Temperature	°C	0 to 50	
ammo::lyser	s::can [2]	E-532-PRO-PH-075	Temperature	°C	0 to 60	0 °C to 60 °C
			pH	-	2 to 12	
			K^+	mg/L	0.02 to 1000	
			NH_4^+	mg/L	0.02 to 1000	
spectro::lyser	s::can	5A-1035-485p0t01-sNO	TSS	mg/L	depending on the optical pathlength	0 °C to 45 °C
			NO_3^-	mg/L	depending on the optical pathlength	
			COD_{total}	mg/L	depending on the optical pathlength	
			$COD_{solbule}$	mg/L	depending on the optical pathlength	

[1] Hach, Loveland, CO, United States
[2] s::can, Wien, Austria

Figure 3.16: Hach Conductivity sensor used in the RSM30 stations.

Dissolved oxygen sensor

A Luminescent Dissolved Oxygen (LDO) sensor produced by Hach is used for the dissolved oxygen concentration determination (Figure 3.17). This sensor is specially designed for municipal and industrial wastewater applications (Hach, 2006a).

The basis of the LDO sensor is to transmit blue light from a LED to the sensor surface. Between the flashes of blue light, a red LED is flashed on the sensor and used as an internal reference. The blue light excites the luminescent material and when the material relaxes it emits red light. The time for the red light to be emitted is measured. Increased oxygen in the sample decreases the time it takes for the red light to be emitted. The time measurements correlate to the oxygen concentration.

Figure 3.17: Hach LDO sensor used in the RSM30 stations.

The LDO sensor is only installed at the outlet, since the concentration of the organic matter is high. Thus, the consumption of the DO is also high and its concentration is around 0 mg/L.

Like the pH sensor, even if the LDO is not equipped by automatic cleaning, compressed air is provided to prevent matter attachments. Also, the air is activated every 5 minutes.

Turbidity sensor

The turbidity is measured by the Solitax sensor (Figure 3.18). Manufactured by Hach, this probe can measure the two correlated variables turbidity and TSS. By the way, only one of them can be displayed and collected.

Figure 3.18: Hach Solitax sensor used in the RSM30 stations.

The turbidity parameter is measured by the light-scattering principle. It is based on the measurements of the infrared light scattered sideways over an angle of 90°(Hach, 2013).

To determine TSS values, internally, there is a turbidity-TSS relation curve that automatically converts from the turbidity measurements.

In this case study, only turbidity measurements have been collected. However, to be able to compare on-line measurements with laboratory measurements, a calibration curve has been developed for each sensor following the procedure presented by Bertrand-Krajewski et al. (2007).

Finally, this sensor is equipped with a wiper that is used as a self-cleaning system to periodically remove some particles and matter attached on the measurement length. The wiper is controlled by the sc1000 controller and is activated every 5 minutes (Detailed in section 3.2.5).

spectro::lyser sensor

The spectro::lyser sensor is manufactured by s::can (Figure 3.19). This probe is available to measure different parameters depending on its application. In case of wastewater applications, the measured parameters are nitrate, Chemical Oxygen Demand (COD), soluble Chemical Oxygen Demand (CODs) and Total Suspended Solids (TSS).

The spectrometer probes work according to the principle of UV-VIS spectrometry over a certain wavelength range (190-720 nm or alternatively, but not used here: 190-390 nm) (s::can,

Figure 3.19: s::can spectro::lyser sensor used in the RSM30 stations.

2007b). Substances contained in the medium to be measured weaken a light beam that transmits through this medium. The light beam is emitted by a lamp, and after contact with the medium its intensity is measured by a detector over a range of wavelengths. Each molecule has a particular (set of) wavelength(s) at which light is absorbed. The concentration of light-adsorbing substances contained in the sample determines the extent of the absorption at a particular wavelength. The wavelengths used for the determination of the concentration if the different pollutant characteristics are selected internally by a proprietary algorithm included in the ana::pro software.

The optical path length of the probe is fixed and cannot be varied. Depending on the application, it can be chosen to be 5 mm, 2 mm, 1 mm or 0.5 mm by inserting a piece allowing to shorten the path length (Figure 3.20). In this application, the path lengths used are 2 mm for the inlet and 5 mm for the outlet.

Figure 3.20: Inserting piece to reduce the path length.

Furthermore, this sensor has a self-cleaning system controlled by the ana::pro software. Compressed air is injected periodically to remove possible attached matter. In this application, the air clean is activated every 5 minutes.

ammo::lyser sensor

The ammo::lyser sensor, manufactured by S::can, monitors the concentration of ammonium and potassium ions (Figure 3.21). The measuring principle is based on ion selective electrodes using membranes that are porous for one specific ion-type (s::can, 2007a).

Figure 3.21: s::can ammo::lyser sensor used in the RSM30 stations.

A robust ion selective membrane in the ammonium electrode separates the ammonium ions from the water. Afterwards, to compensate automatically for possible remaining cross-sensitivities, the ammo::lyser is equipped with sensors for temperature and pH, and a potassium electrode.

Like the spectro::lyser probe, the ammo::lyser is equipped with a compressed air self-cleaning system also controlled by ana::pro providing compressed air every 5 minutes.

3.2.5 Controller

All Hach sensors are connected to a sc1000 Multi-parameter Universal Controller. It is a state-of-the-art controller system with the possibility to use it directly with 8 sensors or in a network with several other sc1000 controllers to accommodate many more sensors and parameters (Hach, 2012).

The controller consists of a display module and one or more probe modules. The display module is a touch-screen display with a user-friendly interface. In normal operation the touch screen displays the measured values for the probes selected (Figure 3.22). Each probe module can be configured with relays, analog outputs, analog or digital inputs, and digital fieldbus cards.

The display offers different display modes and a pop-up toolbar:

- **Measured value display**: The controller identifies the connected probes and displays the associated measurements.
- **Graph display**: Displays time series with the measured values as graphs.

Figure 3.22: Interface and color touch-screen display of the sc1000 controller (Hach, 2012).

- **Main menu display**: Software interface for setting up parameters and settings of a device, probe and display module (Figure 3.23).
- **Pop-up toolbar**: The pop-up toolbar provides access to the sc1000 controller and probe settings and is normally hidden from view.

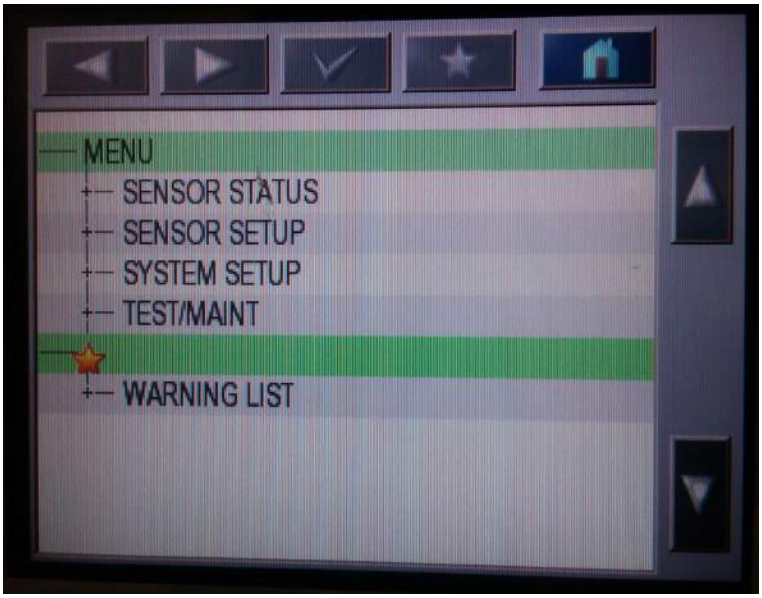


Figure 3.23: Main menu display of the sc1000 controller (Hach, 2012).

3.2.6 Maintenance and operation

The stations require periodical control and maintenance visits. Supervision of the operation of the stations is essential to assure the quality and reliability of the measurements. For that a functional check might be required for one of the following reasons:

- Routine functional check
- Suspicion of fouling of the measuring windows of the light-based sensors and electrodes
- Change of location where the probe is deployed or change in type of probe connected
- Suspicion of probe malfunction
- Changes in operational conditions

During the execution of a functional check, the following actions proposed by Plana (2013) have to be performed.

- Checking the actual status and the functionality of the different probes
- Checking the trueness and precision of the readings to describe the reliability of the sensor (ISO, 1994).
- Checking the historical status or system stability [3]
- Checking unintentional modifications of the measuring settings caused by unauthorized access or remote control
- Checking the probe's mounting

To carry out these functional check activities, a specific protocol for cleaning, validation and calibration has been established in this thesis for each sensor (See section 4.2.2). The frequency by which the activities have to be realized depends mainly on the type of application.

Cleaning protocol

Fouling due to the adhesion of silt, clay or biofilm on the sensors is one of the main problems when working with AMS. Despite automated cleaning systems (wiper, air cleaning) installed on some sensors, manual cleaning is still needed.

Normally, the cleaning is done using a soft, wet cloth (Kim Wipe) and distilled water. In case there is still some attached matter on the sensor, a kitchen degreaser and a hydrochloric acid (3 %) solution can be used. The procedure is detailed in Boudreau (2011).

[3] The stability is known as the ability of a sensor to keep its performance characteristics for a relatively long period of time (ISA, 1982).

Validation protocol

The reliability of the measured values is another important characteristic for on-line measurements. Different error sources can act simultaneously on the output: bad calibration, membrane deterioration and sensor desconfiguration are the most common problems. When that happens, the accuracy of the data decreases. To detect these errors, an evaluation of the sensor measurements is required periodically to detect these errors.

Data validation procedures used in this thesis are detailed in section 3.3.

Calibration protocol

Regular calibrations are required due to the sensor errors presented above (Section 3.2.6). To verify whether calibration is necessary, the displayed readings should be compared with the values of a reliable comparison method (Section 3.3.1). In case of a significant difference between the laboratory values and the readings of the sensor following this method, a calibration has to be performed.

pH, conductivity and LDO sensors are calibrated from the sc1000 controller. On the other hand, the spectro::lyser and the ammo::lyser are calibrated from the ana::pro software.

3.3 Data validation procedures

On-line measurements can provide detailed information on the system dynamics in water bodies. However, assuring data quality is a challenge. As mentioned before (Section 2.3 and 3.2.6), the accuracy and reliability of these on-line measurements is highly important to assure good water quality monitoring (Thomann et al., 2002).

In this thesis a combination of univariate off-line and on-line methods, presented by Alferes et al. (2013), have been used to evaluate and validate the collected data (Figure 3.24). On one side, off-line analysis are applied to detect systematic errors and bad calibration (Alferes et al., 2013; Thomann et al., 2002). On the other side, on-line analysis is used to detect unusual situations from the sensors like drift, shift, bias, or outliers (Yoo et al., 2008).

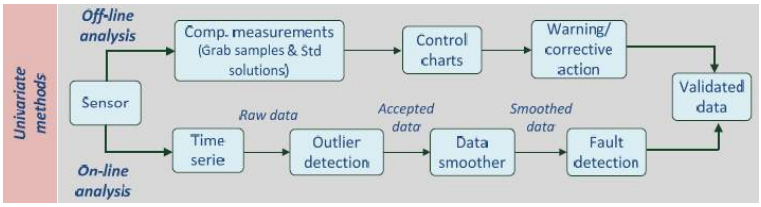


Figure 3.24: Univariate methods for water quality data assessment (Alferes et al., 2013).

In this thesis, the abnormal situations have been classified as outliers and faults. Outliers are

understood as single value of a set of values which is inconsistent with the other values of the data set (ISO, 1994). Faults, on the other hand, are considered when the trueness and precision of the sensor decrease and a bias, a drift, a complete failure or a precision degradations of a set of values is detected.

3.3.1 Off-line analysis for fault detection

The most common off-line control of on-line sensors consists of using concentration values measured in grab samples (Thomann et al., 2002). Measuring these samples with a reference method, (ISO, 2003), allows to detect systematic and gross errors. For that, control charts are applied to compare on-line and lab measurements and determine out-of-control situations and estimate how the system is working (Montgomery, 2008).

Control charts

The first theory of control charts was proposed by Walter A. Shewhart (Montgomery, 2008), hence their name Shewhart control charts.

Montgomery (2008) proposes a general model for control charts. A typical control chart is based on three control lines (Figure 3.25): the central line, the upper control limit (UCL), and the lower control limit (LCL). The central line determines the general level of the process, and the UCL and LCL are used to judge statistically whether the process is operating in control or not.

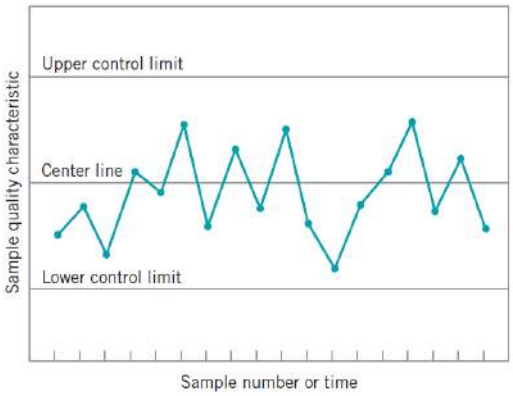


Figure 3.25: Model of a typical control chart (Montgomery, 2008).

Before a control chart is built, several observations have to be collected. Afterwards, the estimation of the average level and the control limits can be plotted on the control chart (Duncan, 1967; Montgomery, 1980). The procedure applied to develop the control charts is:

1. Analyze that the data (x) meets the specified criteria:

- Data should usually be distributed normally around an average.
- Measurements need to be independent of one another.

2. Calculate the mean of the observed data ($\overline{x}$).

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{3.1}$$

The average delimits the central line:

$$Center\ line = \overline{x} \tag{3.2}$$

3. Calculate the standard deviation (σ_w) of the data.

$$\sigma_{\overline{x}} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{3.3}$$

4. Calculate the upper and lower control limits (UCL, LCL) as:

$$UCL = \overline{x} + L\sigma_{\overline{x}} \tag{3.4}$$

$$LCL = \overline{x} - L\sigma_{\overline{x}} \tag{3.5}$$

Generally, for the limit-lines, the L parameter is arbitrarily taken as 3. This is typically called the three-sigma control limit.

Additionally, apart from the control limits, in some other monitoring projects, such as these of van Griensven et al. (2000) and Thomann et al. (2002), warning limits are also calculated and applied on the control charts. Thus, the upper warning limit (UWL) and the lower warning limit (LWL) are determined as:

$$UWL = \overline{x} + L\sigma_{\overline{x}} \tag{3.6}$$

$$LWL = \overline{x} - L\sigma_{\overline{x}} \tag{3.7}$$

For warning limits, L is arbitrarily taken as 2.

5. Plot the *Central line*, the *UCL* and the *LCL* lines and graph the sample number or time (x-axis) versus measurements (y-axis) similar to figure 3.25.

6. Finally, evaluate the measured value to see if the process is out-of-control.

It is possible to observe out of control points when there is an indication that the process has changed (Berthouex, 1989). Also, some criteria are established to decide whether the system is out-of-control (e.g. at least one point outside of the control limits and two of three consecutive points outside the warning limits but inside the control limits (Montgomery, 2008; Nelson, 1984, 1985)).

$\overline{x}$ and $\sigma_{\overline{x}}$ have to be estimated from preliminary samples (x_i) when the process is thought to be in control (Montgomery, 2008). These estimations should be done with at least 20 samples.

In this particular case, the control charts used are based on standard solutions and grab samples measured with a reference method according to ISO (2003). The utility of them is to detect when it is necessary to clean and calibrate the sensors (NMKL, 1990). A specific methodology has been developed for this specific type of control charts based on the method explained above. Standard methods for this type of control chart can be found in APHA (1995) and ASTM (1990).

The steps established to develop the control charts in this thesis are:

1. Analyze if the data (x) is normally distributed.
2. Calculate the difference between the measured value (x_i) by the sensor and the value of the standard solution or the measured value by the reference method (μ).

$$d_i = \mu - x_i \tag{3.8}$$

3. The center line is defined in 0 since it is the expected value of the difference between the measured value and the lab value.

$$Center\ line = 0 \tag{3.9}$$

4. Calculate the standard deviation ($\sigma_{\overline{d}}$)

$$\sigma_{\overline{d}} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} d_i^2} \tag{3.10}$$

5. Calculate the upper and lower control limits (UCL, LCL) using the formulas:

$$UCL = +L\sigma_{\overline{d}} \tag{3.11}$$

$$LCL = -L\sigma_{\overline{d}} \tag{3.12}$$

   The L parameter is taken as 2, corresponding to a 95 % of probability that the value can be accepted.

6. Graph the control chart, by drawing the different lines and plotting the differences between measured values and the reference value (y-axis) versus the sample time (x-axis).
7. Evaluate the graph to see if the process is out-of-control.

In this case, it is established that when a point is out of the boundaries, the system is out-of-control, and a calibration is required. If after several calibrations, the system still persists out-of-control, the sensor should be send to the company for repair or replacement.

3.3.2 On-line analysis for fault detection

Regarding the on-line analysis, the methodology applied is based on the automatic data quality evaluation of Alferes et al. (2013) that permits to detect outliers and faults. This data assessment procedure is based on an univariate method, and it consist in three different consecutive parts (Figure 3.24): outlier detection, data smoothing and fault detection.

Outlier detection

The outliers detection method is based on the forecasting of the expected value and the comparison of the new collected value with this forecast. Using autoregression (AR) models, it is possible to identify the AR forecasting model considering historical data. Afterwards, the model is projected into the future to obtain the forecast (Alferes et al., 2013).

Outliers are identified by comparing the measured values with the forecast values taking into account their dynamic prediction error determined by the standard deviation of the forecast error. A third-order exponential smoothing model is used to estimate the forecast values, whereas a first order smoothing model is used to calculate the forecast error. First, the forecast model is explained.

At time T, the forecast value at the next time step $(T+1)$ can be calculated following (Alferes et al., 2013):

$$\widehat{x}_{T+1} = \widehat{a}_T + \widehat{b}_T + \frac{1}{2}\widehat{c}_T \tag{3.13}$$

where $\widehat{a}_T$, $\widehat{b}_T$ and $\widehat{c}_T$ are the coefficients of the model. They can be computed using the first, second and third exponentially smoothed statistics (S_T, $S_T^{[2]}$ and $S_T^{[3]}$). These smoothed statistics can be calculated as:

$$S_T = \alpha x_T + (1-\alpha) S_{T-1} \tag{3.14}$$

$$S_T^{[2]} = \alpha x_T + (1-\alpha) S_{T-1}^{[2]} \tag{3.15}$$

$$S_T^{[3]} = \alpha x_T + (1-\alpha) S_{T-1}^{[3]} \tag{3.16}$$

where α is the smoothing constant that minimizes the residuals between the model and a representative set of error-free calibration data. Small α values give more weight to the historical data, giving a slow response. In a similar way, large α values provide more weight to the current observed data leading to a faster response.

Afterwards, the coefficients of the model (Equation 3.13) are calculated by:

$$\widehat{a}_T = 3S_T - 3S_T^{[2]} + S_T^{[3]} \tag{3.17}$$

$$\widehat{b}_T = \frac{\alpha}{2(\alpha-1)^2}\left[(6-5\alpha)S_T - 2(5-4\alpha)S_T^{[2]} + (4-3\alpha)S_T^{[3]}\right] \tag{3.18}$$

$$\widehat{c}_T = \left(\frac{\alpha}{\alpha-1}\right)^2\left(S_T - 2S_T^{[2]} + S_T^{[3]}\right) \tag{3.19}$$

Alferes et al. (2013) and Saberi (2015) propose the following algorithm for outlier detection:

1. Select a "Good data" set without any important number of faults. This selection is based on personal experience and preliminary statistical analysis to assure that the system is in control (e.g. control charts).

2. Calibrate the model to determine the smoothing constant α.

   According to the 3^{rd} order exponential smoothing method, the root-mean-square error (RMSE) between the observed values and the forecast values is calculated for α values ranging from 0.01 to 1 (with an interval of 0.01). The minimum RMSE value points to the best α value.

3. Set the initial values of S_1, $S_1^{[2]}$ and $S_1^{[3]}$ equal to x_1.

4. Calculate $\widehat{a}_1$, $\widehat{b}_1$ and $\widehat{c}_1$ according to the equations 3.17, 3.18 and 3.19.

5. Calculate the one-step-ahead forecast data at time 1, $\widehat{x}_2$, with the equation 3.13.

6. Having the observed value x_T, calculate S_T, $S_T^{[2]}$ and $S_T^{[3]}$ for the subsequent steps T according to the equations 3.14, 3.15 and 3.16

7. Calculate the model parameters $\widehat{a}_T$, $\widehat{b}_T$ and $\widehat{c}_T$ with the equations 3.17, 3.18 and 3.19.

8. Calculate the one-step forecast data $\widehat{x}_{T+1}$ with equation 3.13.

9. Calculate the one-step-ahead error $e_T(1)$ using:

$$e_T(1) = x_T - \widehat{x}_T \tag{3.20}$$

   where x_T is the observed value and $\widehat{x}_T$ is the predicted value at time T calculated at time $T-1$.

10. Estimate the variance of the forecast error σ_e^2 through the estimation of the mean absolute standard deviation Δ using:

$$\widehat{\sigma}_{e,T} = 1.25\widehat{\Delta}_T \tag{3.21}$$

Considering that the forecast error is a normal distribution, the approximation to estimate its standard deviation is $\sqrt{\frac{\pi}{2}}$ which is 1.25. $\widehat{\Delta}_T$ is calculated by:

$$\widehat{\Delta}_T = \alpha_{std}|e_T(1)| + (1-\alpha)\widehat{\Delta}_{T-1} \tag{3.22}$$

where α and α_{std} are the smoothing constants of the forecast variable and of the standard deviation of the error, respectively.

11. Calculate the prediction interval $xlim$ as:

$$xlim = \widehat{x}_T \pm K\widehat{\sigma}_{e,T} \tag{3.23}$$

where K is a proportional constant that can be adjusted depending on the need of more or less restrictive limits. For small values of K, the limits are more restrictive.

12. Repeat steps 6 to 11 until all observed data have been treated.

13. Identify the outliers by comparing the observed data and the prediction error interval.

    a) If x_T is outside the prediction interval, it is considered to be an outlier and it is replaced by the forecast value ($\widehat{x}_T$) (Figure 3.27).

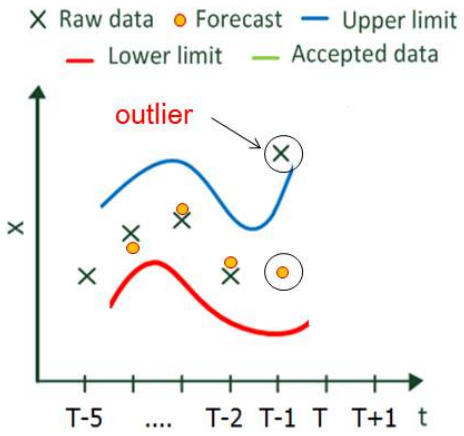


Figure 3.26: Method to detect outliers with an outlier just detected (Alferes et al., 2013).

14. Create a new *accepted data* series.

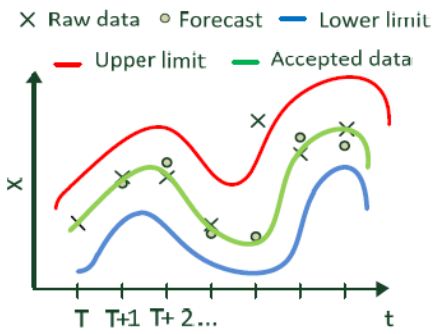


Figure 3.27: Method to detect outliers with the detected outlier replaced (Alferes et al., 2013).

Note that this outlier detection procedure in the univariate method also comes with a backward reinitialization algorithm to detect when the system is out of control. It means that this algorithm is able to detect when several consecutive points are considered as outliers (Saberi, 2015). When that happens, the algorithm reinitiates in the backward time direction to recuperate the lost data until the starting point of the out of control situation. Then, the algorithm changes back to the forward direction. This backward reinitialization algorithm permits to skip the data that have caused the out of control situation.

Data smoothing

After the outliers have been removed and replaced and the new *accepted data* set has been created, the next step is to smooth the data to decrease the noise that interferes with the data and can lead to wrong results in the fault detection step.

The method used to smooth the data is a Kernel smoother with a proper bandwidth (Alferes et al., 2012). This methodology consists in fitting a real-valued Kernel function on the noisy data by using nonparametric regression (Takahama and Sakai, 2009; Wand and Jones, 1994). Its main objective is to estimate the regression function using a weighted average of the raw data. Schimek (2013); Aydin (2007); Cai (2001) propose to use the weighted Nadaraya-Watson approach.

The kernel estimate of the function $\widehat{y}_h$ can be expressed as:

$$\widehat{y}_h(x_0) = \sum_{i=1}^{n} W(x_0, x_i; h) \cdot y(x_i) \tag{3.24}$$

where $\widehat{y}_h(x_0)$ is the estimated value of the observed point at x_0, n is the number of observed points, $W(x_0, x_i; h)$ is the weighting function, $y(x_i)$ are the observations at the x_i points and h is the bandwidth which determines the number of neighboring points around x_0 used to estimate the value.

According Nadaraya-Watson regression estimation, the weighting function is obtained by:

$$W(x_0, x_i; h) = \frac{K\left(\frac{x_0 - x_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x_0 - x_i}{h}\right)} \tag{3.25}$$

where the $K(x)$ is the kernel function. Normally, this function is selected to be nonnegative, symmetric about zero, continuous and twice differentiable (Aydin, 2007). Some kernel function alternatives are presented by Schimek (2013); Takahama and Sakai (2009); Aydin (2007). In this thesis, the Gaussian kernel function is proposed by Alferes et al. (2012):

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \tag{3.26}$$

in this specific case, $x = \frac{x_0 - x_i}{h}$.

Notice that the selection of smoothing parameter, h, is much more important than the kernel function itself due to the predictions of the kernel regression (Hardle, 1990). This parameter permits to control the smoothness or roughness of the estimates. However, it has to be carefully determined since it may lead to under-smoothing or over-smoothing situations (Saberi, 2015).

Once the accepted data has passed through the kernel smoother, the noise has been reduced. This new time series will be called *smoothed data*.

Fault detection

The last step of the data quality assessment is the fault detection. Alferes et al. (2013) propose to calculate some data features over the *smoothed data* and their acceptability limits based on raw data series at each time T (Figure 3.28). Table 3.2 details the different data features used.

After the estimation of the data features and their acceptability limits (Figure 3.28), the final step is to validate the data series:

- *Valid*: all tests are passed
- *Doubtful*: some tests have failed, some further analysis has to be done
- *Not valid*: all tests or the most important tests are failed

[4] Difference between the accepted data and the smoothed data

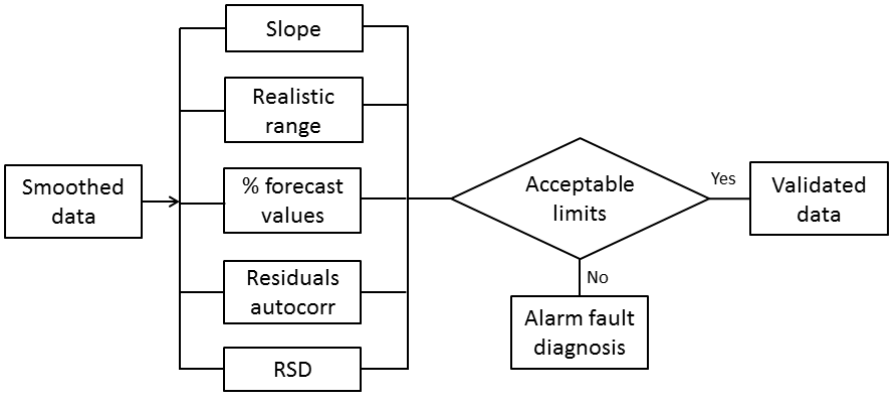


Figure 3.28: Method for fault detection (Alferes et al., 2013).

Table 3.2: Fault detection data features (Alferes et al., 2013).

Feature	**Definition**	**Purpose**
Slope	Slope between two successive points in the smoothed data	Assess the dynamics of the data, need to determine gradients and sudden changes
Locally realistic range	Range where values are commonly observed in a specific position	Determine the inaccuracy of the data in the expected range
% replaced data	Fraction of measured values replaced by the forecast values (outliers)	Assess the goodness of the smoothed data and its data features
Autocorrelation of the residuals	Run test to the residuals [4] to determine their correlation (Dochain and Vanrolleghem, 2001)	Determine if residuals are randomly distributed
Residuals' standard deviation	Standard deviation of residuals	Determine the data variance

When a fault has been detected on the *smoothed data* set, it is not considered any further, neither on the *validated data* set nor for further analysis.

For supplementary information about the test and implementation of the on-line analysis methodology presented in this section the reader is referred to Saberi (2015).

3.4 Databases: data storage and management

In the context of the mon*EAU* project, a structural and relational database is required to assure the good storage and management of the data. The database, named dat*EAU*base, consists in two parts:

- The set linked relational database tables that contain data for a given project.
- The user interface, which is a website accompanied by a user's guide to facilitate data entry, exporting and viewing.

3.4.1 Relational database

As the purpose of the database is to store information in a useful way, it is comprised of multiple related tables that contain records and fields. The fields describe the type of information stored, and the records are the items in the database. Another requirement of the database is a big storage capacity.

To accomplish these needs, MySQL was chosen as an open source relational database management system. It is a widely used database system in the open source sector. Kofler and Kramer (2005); DuBois and Online (2013) enumerate the reasons for its use:

- It is fast, and its speed is increasing after many improvements.
- It is easy to learn and use. There is a user interface that facilitates the development of databases in a simple way and that is much less complex than larger systems.
- It is highly documented on the Internet and books.
- It can run on several operating system such as Windows, Linux, Mac OS x, etc.
- It supports Structured Query Language (SQL), which is the standard language for current database systems.
- It is used for web application development. It can be easily accessed and queried directly using: C, C++, Perl, Java, PHP, Python and VB computer languages.
- It is open source.

3.4.2 User interface

To protect the database and to manage the data safely, a graphical interface for the users has been created. The requirements of the user interface are that it should introduce, export and explore the data on the database in a user-friendly way. Also, it should allow for a dynamical application and be accessible to different users at the same time.

Looking at the list of programming languages to create an interface that can be used to access and query the MySQL database, the Python language is the one that better suits these needs as a powerful tool supporting heavy calculus and graphics (Camhy et al., 2012; Ceder et al., 2010).

Python can be used as a scripting language for web applications. Some web frameworks support Python for the development, design and maintenance of complex web applications. To create these web applications, one of these high-level Python web frameworks, called Django, has been used (Alexis and Bersini, 2012; Holovaty and Kaplan-Moss, 2009).

It has been designed to build dynamic and interesting web sites as quickly as possible. It also provides the basic patterns and shortcuts for frequent programming tasks.

Regarding the accessibility of the database, Django takes care of security, user authentication and content administration, providing a secure way to manage user accounts and passwords and avoiding many common security mistakes (Holovaty and Kaplan-Moss, 2015).

Finally, this web framework is easy, quick and flexible in terms of adaptation to the requirements.

Chapter 4

Results

In this chapter, the developed tools, the methodology established to optimize the maintenance tasks, and some examples of collected high frequency data and their interpretation are presented.

4.1 dat*EAU*base

As mentioned in section 2.4, tools to manage and store collected data and especially metadata are needed. In the context of this thesis, an improved dat*EAU*base has been developed. Following the work of Plana (2013), a modular structure has been used (see figure 4.1). As discussed in section 3.4, the softwares selected to develop the dat*EAU*base have been MySQL for the database development and Django for the graphical user interface (GUI).

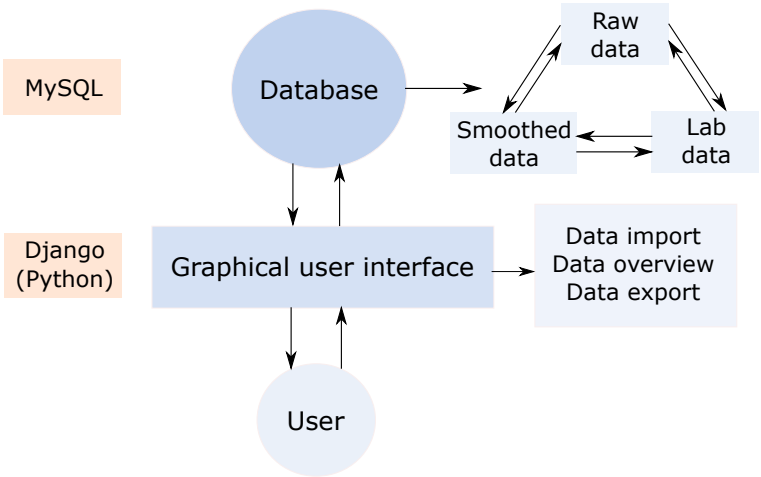


Figure 4.1: Modular design of the dat*EAU*base.

MySQL has been selected because, compared to MS Access, it is an open source software that offers a large storage capacity, and provides more flexibility for database design and data

queries. Also more options are offered for the tables design and the relations between the tables.

Furthermore, Django has been chosen as a Python web framework because Python is a powerful programming language able to design a comprehensive GUI and to support large data series at the same time. Comparing Python with R or MS Excel, it offers more flexibility and capacity.

In the following sections, both the database and the GUI are explained in detail. Also, a discussion of their expectations and their design is proposed.

4.1.1 Database

The database has been designed to store all relevant data, including the raw data, the filtered data, the lab measurements and all corresponding metadata. Storing raw, filtered and lab data together has been considered important since all of them are related and essential to obtain validated data series as mentioned in section 3.3.

In case of the metadata, for further analysis and for the interpretation of the measurements data, a deep study has been conducted on the database design and which information will be relevant. Firstly, a draft structure with the main fields to be included in the database has been proposed trying to answer the different questions presented in section 2.4 (see figure 4.2). Afterwards, every field has been broken down in several tables. The final design proposed for the database is presented in figure 4.3. This structure is based on 23 different tables that are interrelated.

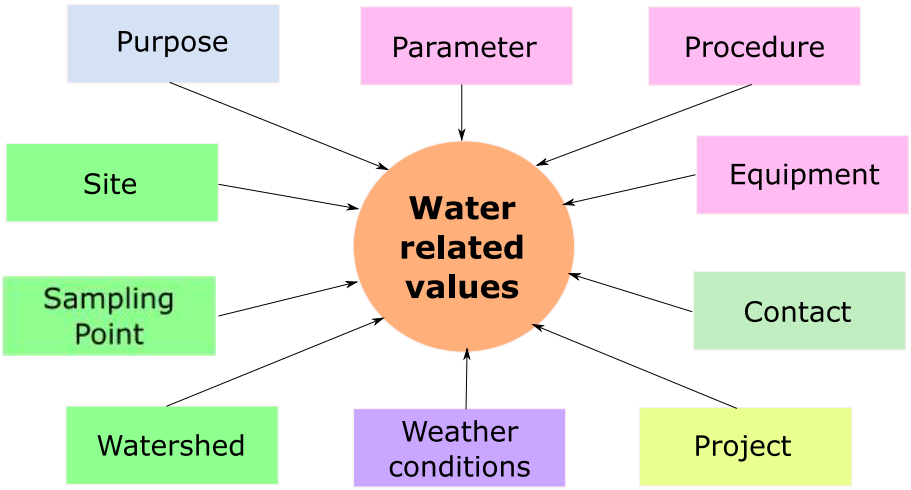


Figure 4.2: dat*EAU*base structure.

In comparison to the dat*EAU*base structure proposed Plana (2013), this new structure provides more flexibility and more information about weather conditions, the purpose of the

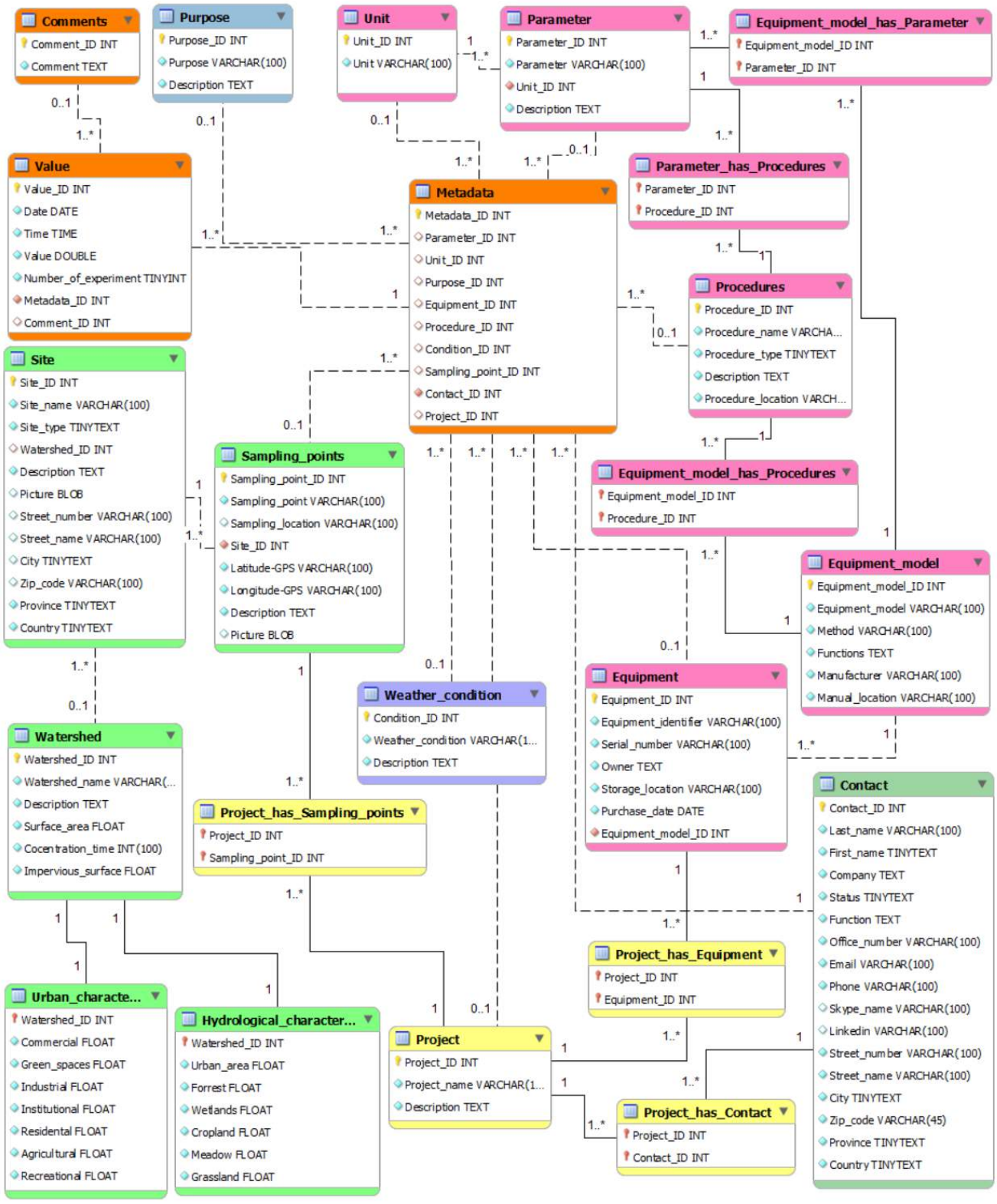


Figure 4.3: dat*EA U*base model with the links between the tables. Details for different groups of tables are given in figures 4.4 - 4.7.

measurements, the watershed characteristics, the equipment, and the procedures. Also, it optimizes the data storage since any combination of metadata is recorded with a number identification and linked with the *Value* table.

Every table contains a primary key (indicated as a yellow filled key symbol in figure 4.3), which is a column that uniquely identifies each row in a table. The links between the tables are made through the primary keys. The relation between tables can be either optional (designed as dotted lines in figure 4.3) or mandatory (designed as continuous lines in figure 4.3) and there are three different types of links:

- **1:1** - One row of the first table is related to only one row of the second table
- **1:n** - One row of the first table is related to multiple rows of the second table. In the opposite direction, each row of the second table is related to only one row of the first table. And it is represented as 1.* to figure 4.3.
- **m:n** - Each row of the first table can be related to multiple rows of the second table and also in the opposite direction.

In the cases of 1:1 and 1:n relations, the primary key of one table becomes a column in the other table and this new column is called a foreign key (indicated as red diamond in figure 4.3) because it is not a primary key in this table but it is for the other. In the case of a m:n relation between two tables, the two primary keys of both tables are added as a new separate table (indicated by red filled keys). This table is connected to those tables with a 1:n link.

Furthermore, in figure 4.3, filled diamonds indicate that the field cannot be null, while empty diamonds mean that the field can be null.

In figure 4.3, the orange tables are the main tables of the database, called primary tables (see enlarged structure in figure 4.4). In these tables, the values and the corresponding metadata are stored. The *Metadata* table is directly or indirectly linked to all other lookup tables for detailed information of each field in this table.

Each value stored in the dat*EAU*base is unique because it has a time stamp and can have a comment, next to a whole range of metadata that explain which parameter the value pertains to, its units, the purpose of the measurement, the equipment used to measure it, according to which procedure, where, under what conditions, by whom and in which project.

For example, on June 15, 2015 at 10:53:00, the conductivity was measured in mS/cm with the Conductivity_001 sensor under dry conditions, with the purpose of calibrating it according to ISO-15839 methodology, at the inlet of the Grandes-Piles F/AL by Plana for the project mon*EAU*. And because the calibration was unsuccessful, so a comment was added giving more details.

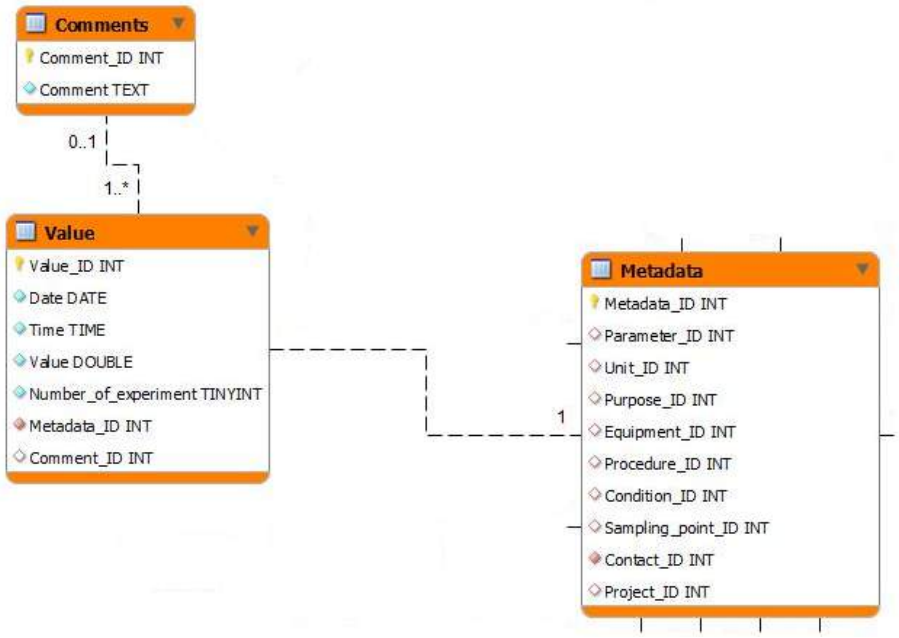


Figure 4.4: Metadata, value and comments tables of the dat*EAU*base.

The number of experiments in figure 4.4 refers to the number of replicate measurements. For on-line measurements, it will always be 1. And for lab analysis, it may be 1, 2 or 3.

The lookup tables have been divided in six different blocks identified by different colors. In the pink tables, all information about the equipment and procedures is stored, as well as which parameters can be measured with the equipment, the measurement procedures and the units used (see enlarged structure in figure 4.5).

For example, the conductivity is the measure of the ion-activity in water and it is measured in mS/cm. This parameter is measured with the sensor Conductivity_001 corresponding to the Hach's model 3727E2T with serial number 25692696 based on the inductive conductivity measurement principle. For further information, the corresponding manual can be found at location PLT-2659. Currently, the sensor is installed in the Grandes-Piles F/AL for on-line measurements and for its maintenance, the followed procedure is SOP_049_Conductivity stored in room PLT-2659.

The tables in green contain all information about the sampling location. More specifically, information about the site and the identification of the sampling points are detailed. Also, urban and hydrological characteristics are collected (see enlarged structure in figure 4.6).

For example, the conductivity sensor is installed at the inlet of the Grandes-Piles F/AL. This F/AL address is 267-303 5e Av, Grandes-Piles, G0X 1H0, in the province of Québec in Canada and the coordinates of the inlet are 46°41'04"N 72°42'59"W. The watershed on which the F/AL

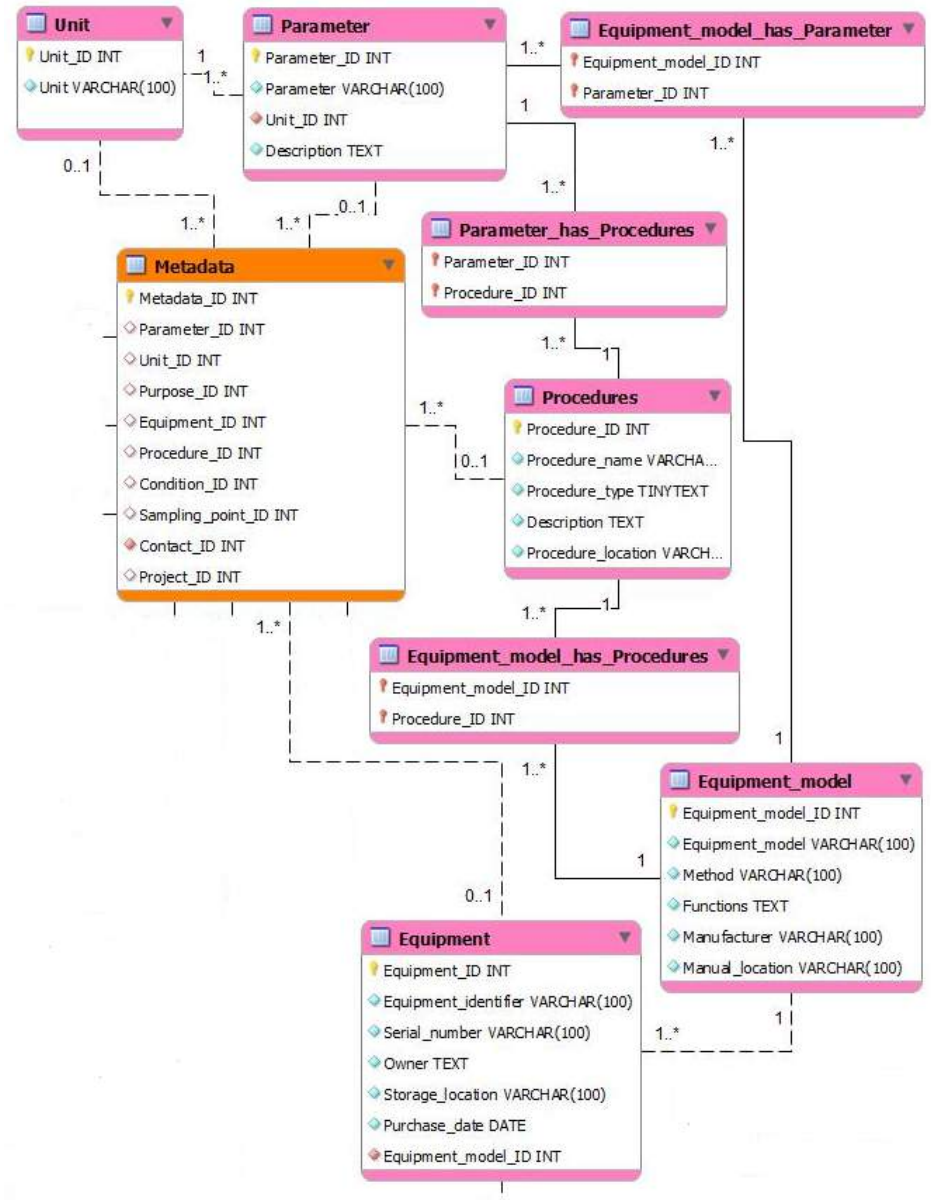


Figure 4.5: Instrumentation information tables of the dat*EAU*base.

is located is the Saint-Maurice river. This watershed is 43,300 km^2, one of the most important tributaries of the Saint-Lawrence river, its concentration time is 35 d and the impervious surface percentage, 42. The urban characteristics of the watershed are 1.5 % of commercial area, 54.25 % of green spaces, 0.75 % of industrial area, 13.5 % of residential area, 22 % of agricultural area and 8 % of recreational area. And the hydrological characteristics are 17 % of urban area, 39 % of forest area, 21 % of wetlands, 12 % of croplands, 8 % of meadow and 3 % of grassland.

The tables in yellow contain the information about the projects and their links to other fields

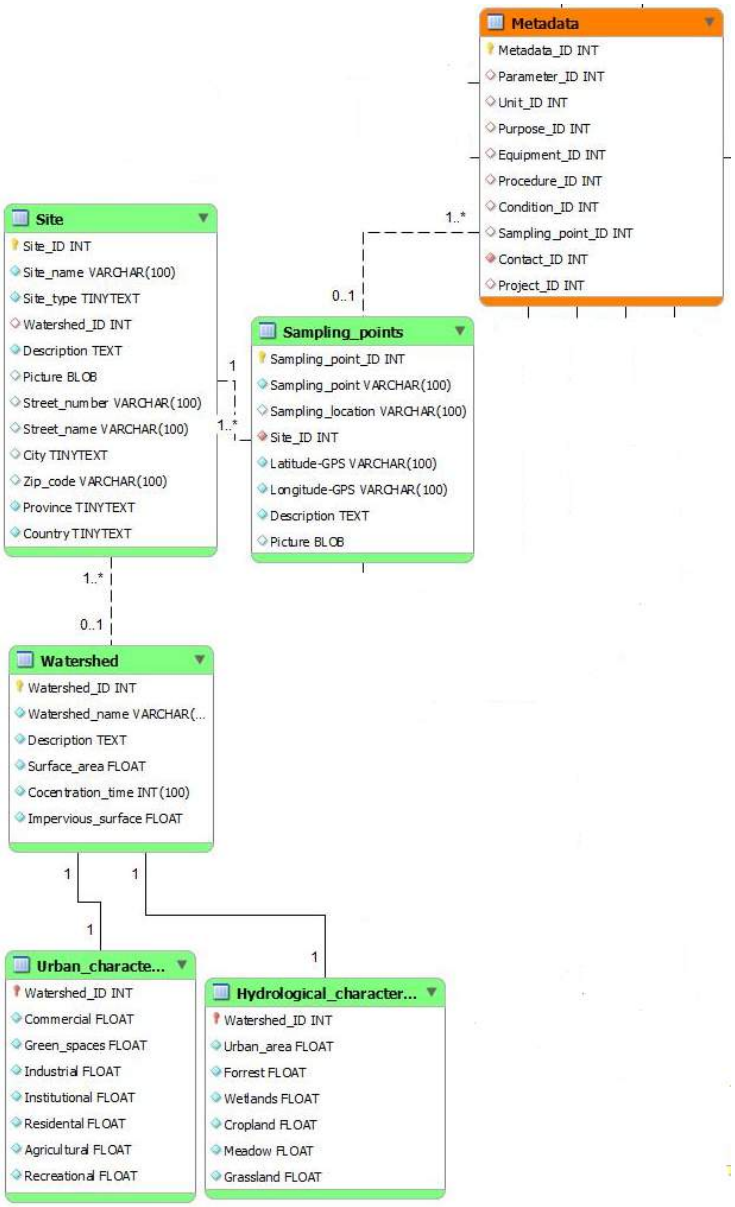


Figure 4.6: Sampling location information tables of the dat*EAU*base.

of the database (see enlarges figure 4.7). These linking tables contain the information about who is working in a project, where a project takes place and which equipment is used, and viceversa, in how many projects someone is working, in how many projects a location is used, and in how many projects an equipment is used.

For example, the mon*EAU* project is the project for which the AMS are currently installed at the inlet of the Grandes-Piles F/AL, the equipment used on it is the Conductivity_001, ph_002, ammo::lyser_001 sensors. The people involved are Alferes, Plana and Vanrolleghem.

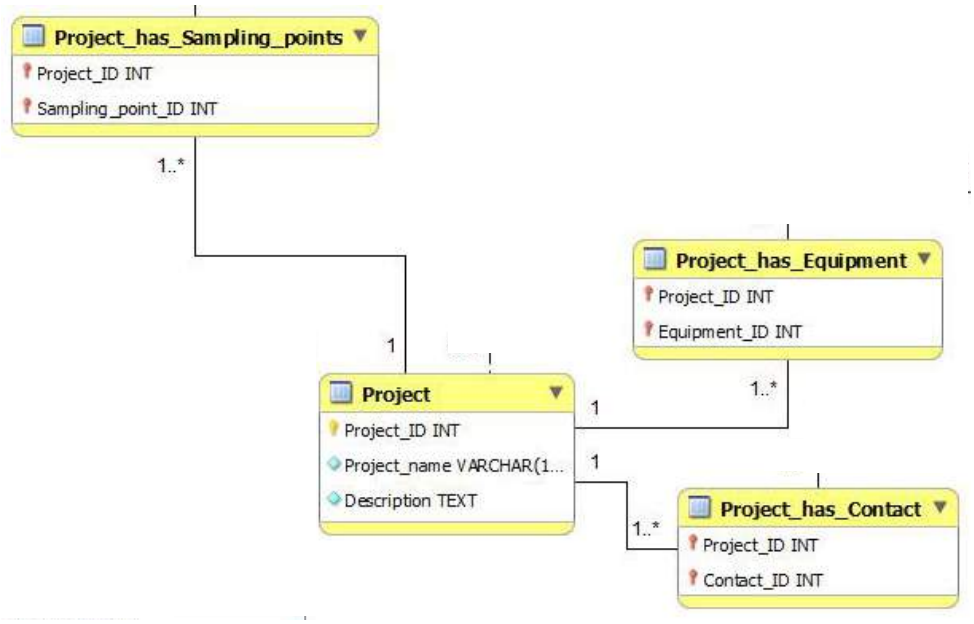


Figure 4.7: Project information table and its relation with other tables of the dat*EAU*base.

In the dark green table, all information of the people involved in a project is stored. In the blue table, the purpose of the measurement is detailed. Finally, in the purple table, the weather information is stored and related with the metadata table.

With this structure and the metadata included in the database it is possible to answer at least the questions presented in section 2.4:

- What has been measured?
- When has the value been measured?
- Where has the value been measured?
- How has the value been measured?
- Who has collected the value?
- Why has the value been measured?

This information permits easy understanding of each value stored in the dat*EAU*base for further analysis and use . The extraction of the desired information is done by simple requests. Generally, the algorithm followed is:

1. Select an existing database in SQL schema.

2. Fetch the data from a database table. This command returns data in the form of result table called result-set.

3. Combine records from two or more tables in a database.

4. Extract only those records that fulfill a specified criterion. For example, a certain period of data.

5. Sort the result-set by one or more columns.

For example, according to this procedure, the MySQL code to query a value with some metadata is:

```
USE dateaubase;

SELECT Date, Time, Value, Number_of_experiment, Comment, Parameter, Unit, Purpose,
    Equipment_identifier, Procedure_name, Weather_condition, Sampling_point,
    Sampling_location, Site_name, Site_type, Watershed_name, Last_name, Project_Name

FROM Value

LEFT JOIN Comments ON Value.Comment_ID = Comments.Comment_ID
LEFT JOIN Metadata ON Value.Metadata_ID = Metadata.Metadata_ID
LEFT JOIN Parameter ON Metadata.Parameter_ID = Parameter.Parameter_ID
LEFT JOIN Unit ON Metadata.Unit_ID = Unit.Unit_ID
LEFT JOIN Purpose ON Metadata.Purpose_ID = Purpose.Purpose_ID
LEFT JOIN Equipment ON Metadata.Equipment_ID = Equipment.Equipment_ID
LEFT JOIN Procedures ON Metadata.Procedure_ID = Procedures.Procedure_ID
LEFT JOIN Weather_condition ON Metadata.Condition_ID =
    Weather_condition.condition_ID
LEFT JOIN Sampling_points ON metadata.sampling_point_ID =
    Sampling_points.Sampling_point_ID
LEFT JOIN Site ON Sampling_points.site_ID = Site.Site_ID
LEFT JOIN Watershed ON Site.Watershed_ID = Watershed.Watershed_ID
LEFT JOIN Contact ON Metadata.Contact_ID = Contact.Contact_ID
LEFT JOIN Project ON Metadata.Project_ID = Project.Project_ID

WHERE Value.Date between '2014-07-13' and '2014-07-13'
and Value.Time between '10:00:00' and '10:00:25'

order by Value_ID;
```

The result of this request is presented in a table like:

Table 4.1: Result of the query described in the text.

Date	Time	Value	Number_ of_ experiment	Comment	Parameter	Unit	Purpose	Equipment_ identifier
2014-07-13	10:00:00	15.034	1		Conductivity	mS/cm	Sensor-testing	Tetra-Con700IQ_001
2014-07-13	10:00:05	17.398	1		Conductivity	mS/cm	Sensor-testing	Tetra-Con700IQ_001
2014-07-13	10:00:10	17.258	1		Conductivity	mS/cm	Sensor-testing	Tetra-Con700IQ_001
2014-07-13	10:00:15	17.401	1		Conductivity	mS/cm	Sensor-testing	Tetra-Con700IQ_001
2014-07-13	10:00:20	17.399	1		Conductivity	mS/cm	Sensor-testing	Tetra-Con700IQ_001
2014-07-13	10:00:25	34.562	1	Sensor was not cleaned well	Conductivity	mS/cm	Sensor-testing	Tetra-Con700IQ_001
2014-07-13	10:00:00	25	1		Ammonium	mg/l	Measur-ment	ammo::lyser
2014-07-13	10:00:05	25.3	1		Ammonium	mg/l	Measur-ment	ammo::lyser
2014-07-13	10:00:10	25.2	1		Ammonium	mg/l	Measur-ment	ammo::lyser
2014-07-13	10:00:15	25.4	1		Ammonium	mg/l	Measur-ment	ammo::lyser
2014-07-13	10:00:20	3	1	Sensor was not in the immersed	Ammonium	mg/l	Measur-ment	ammo::lyser

Table 4.2: Result of the query described in the text (continued).

Procedure_name	Weather_condition	Sampling_point	Sampling_location	Site	Site_type	Watershed_name	Last_name	Project_name
ISO15839:2003							Alferes	monEAU
ISO15839:2003							Alferes	monEAU
ISO15839:2003							Alferes	monEAU
ISO15839:2003							Alferes	monEAU
ISO15839:2003							Alferes	monEAU
ISO15839:2003							Alferes	monEAU
SOP:Measuring ammonium	Dry-weather	Outlet	Primary_clarifier	Quebec-West	WWTP	Cheveau	Maruéjouls	retEAU
SOP:Measuring ammonium	Dry-weather	Outlet	Primary_clarifier	Quebec-West	WWTP	Cheveau	Maruéjouls	retEAU
SOP:Measuring ammonium	Dry-weather	Outlet	Primary_clarifier	Quebec-West	WWTP	Cheveau	Maruéjouls	retEAU
SOP:Measuring ammonium	Dry-weather	Outlet	Primary_clarifier	Quebec-West	WWTP	Cheveau	Maruéjouls	retEAU
SOP:Measuring ammonium	Dry-weather	Outlet	Primary_clarifier	Quebec-West	WWTP	Cheveau	Maruéjouls	retEAU

Another query example is to find out where an equipment is located and for which project it is used. The corresponding MySQL code is:

```
USE dateaubase;

SELECT Equipment_identifier, Equipment.storage_location, Project_name

FROM equipment

LEFT JOIN project_has_equipment ON equipment.Equipment_ID =
    project_has_equipment.equipment_ID
LEFT JOIN project ON project_has_equipment.project_ID=project.Project_ID

WHERE equipment.equipment_identifier='TetraCon700IQ_001';
```

And the table result is like:

Table 4.3: Result of the query described in the text.

Equip-ment_identifier	Stor-age_location	Project_name
TetraCon700IQ_001	PLT-1234	monEAU

When comparing this database with other databases, one can conclude that it is innovative by emphasizing metadata storage, providing flexibility for possible future improvements and its large storage capacity.

In summary, it is a useful tool to manage large amounts of data without loosing information and the data's quality.

For a detailed explanation of the contents of each table, the links between them, and some examples of the content of each table, the reader is referred to annex A.

4.1.2 Graphical user interface

A GUI has been designed to introduce data to the database, to explore data contained in the database, and to export data from the database in a user-friendly way without the need to know the MySQL language. As mentioned in section 3.4.2, the Python web framework, Django, has been used.

While programming the GUI, it is possible to define format restrictions for the different fields. This promotes the uniformity and consistency of the stored data. For example, the value field has to be a number, the date format has to be 'YYYY-MM-DD'. If the data format does not

correspond to the predefined format, the GUI does not allow to introduce the data into the database.

Also, it allows to protect the database by giving different users different privileges. For example, data introduction to the value and the metadata tables can be done by any user, whereas the data introduction to other lookup tables can only be done by the administrator or an identified superuser. Also, it is possible to only allow access to a certain information to a certain user. For example, giving the permission to an external user to consult some information from a project, an equipment or a sampling point.

Furthermore, easy access and facilities to share the data are other facilities that the GUI offers. Since it is a web framework, any allowed user can access the dat*EAU*base from anywhere with internet connection.

To sum up, in a research group, where everybody has data from different projects, and people are changing continuously, a GUI helps to get used easily to manage the data in a proper format, as well as to consult historical data.

However, lack of time prevented the GUI to be completed and only the superuser interface is available at the moment (see figure 4.8).

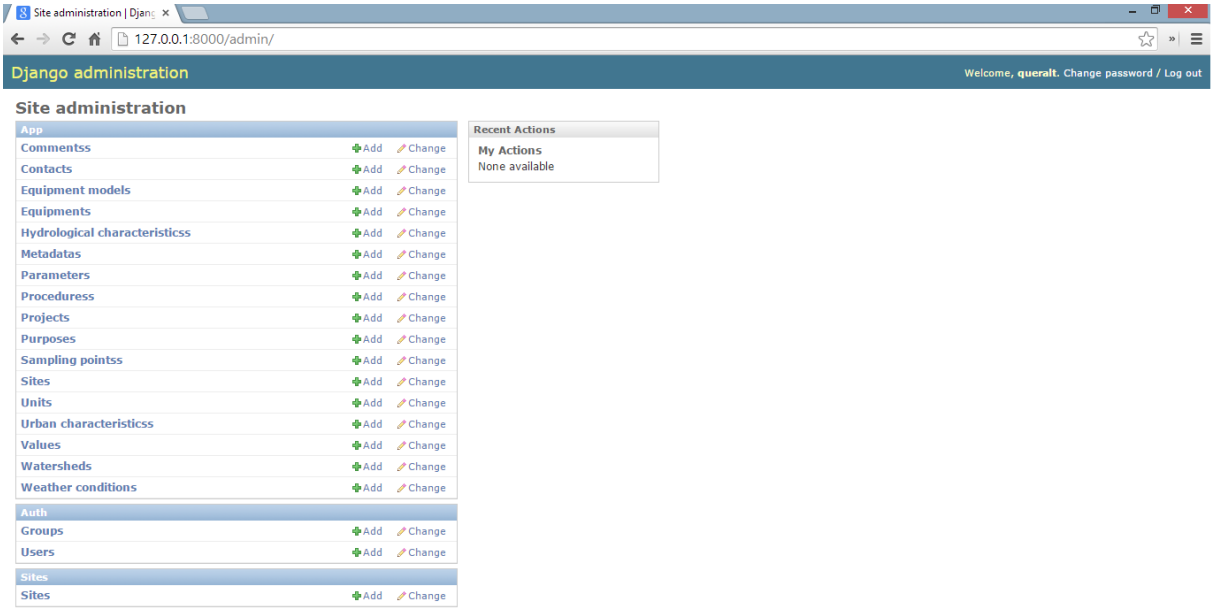


Figure 4.8: Superuser GUI of the dat*EAU*base.

4.2 Making monitoring stations work properly

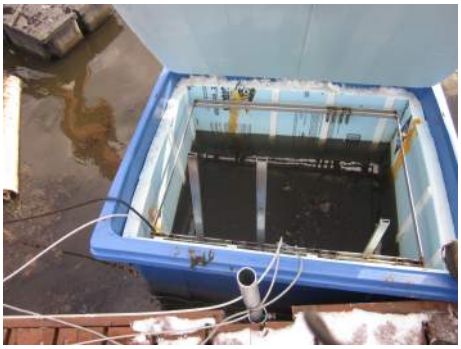
For a successful use of AMS, several challenges related to raw wastewater characteristics and environmental conditions have to be overcome as presented in section 2.6. In this section, a methodology to maintain and keep safe the AMS is developed to deal with some of the challenges, like fouling and bad calibration.

4.2.1 Installation

For in situ type measurements, the sensors are installed inside the water to be monitored (see figure 2.1). In both case studies, supports for the sensors had to be made and customized to each location.

In Wemotaci, at the inlet, the sensors were installed inside a small housing (see figure 4.14a), where in a hole on the floor, all supports and sensors were installed (see figure 4.9a). At the outlet, the sensors were installed outdoors, inside a container that protected the sensors and held the supports (see figure 4.9b).

(a) Sensors support at the inlet of the Wemotaci lagoon.



(b) Sensors support at the outlet of the Wemotaci lagoon.

Figure 4.9: Different sensor supports made for the Wemotaci F/AL.

In Grandes-Piles, at the inlet, the sensors were installed in a basin where the wastewater arrives before being pumped to the F/AL. The supports were attached to the basin as presented in figure 4.10a. At the outlet, the system is similar to the one at the inlet at Wemotaci, i.e. there is a hole in the floor where the sensors are instaled (see figure 4.10b).

Nevertheless, in Québec, there is a big difference between winter conditions and summer conditions. For that, the installation has to be prepared for cold and warm temperatures, and also for dry and wet weather. Moreover, not only it is necessary to protect the instrumentation, but one must also assure safe working conditions for the technicians and researchers.

At low temperatures, the water temperature at the outlet of F/ALs can be 0 °C or even below. Since most of the sensors cannot work below 0 °C (see table 3.1) and part of the sensor is above the water phase (air or ice), a heater was installed during the winter period at the outlet of

(a) Sensors support at the inlet of the Grandes-Piles F/AL.

(b) Sensors support at the outlet of the Grandes-Piles F/AL.

Figure 4.10: Different sensor supports made for the Grandes-Piles F/AL.

both case studies to keep the area where the sensors were installed ice-free (e.g. figure 4.11). Figure 4.12 proves that at low temperatures, by keeping the water around the sensors ice-free with the heater, the sensor continued working.

Figure 4.11: Heater installed in the outlet container at Wemotaci.

Also, under winter and wet conditions, cable connections outdoors suffer because water may infiltrate and freeze. To prevent this, a protection cage is needed. The one that was installed in Wemotaci F/ALs is shown in figure 4.13.

Finally, to protect the technicians and researchers from the severe cold, a place to work safely has to be available. In Wemotaci, the housings where the stations were installed was not big enough to comfortably work in. Thus, two tents were installed around the housing as presented in figure 4.14a. Afterwards, thanks to the experience in Wemotaci, two small houses were built in Grandes-Piles with enough space to house the stations and the sensors, as well as to work inside (see figure 4.14b).

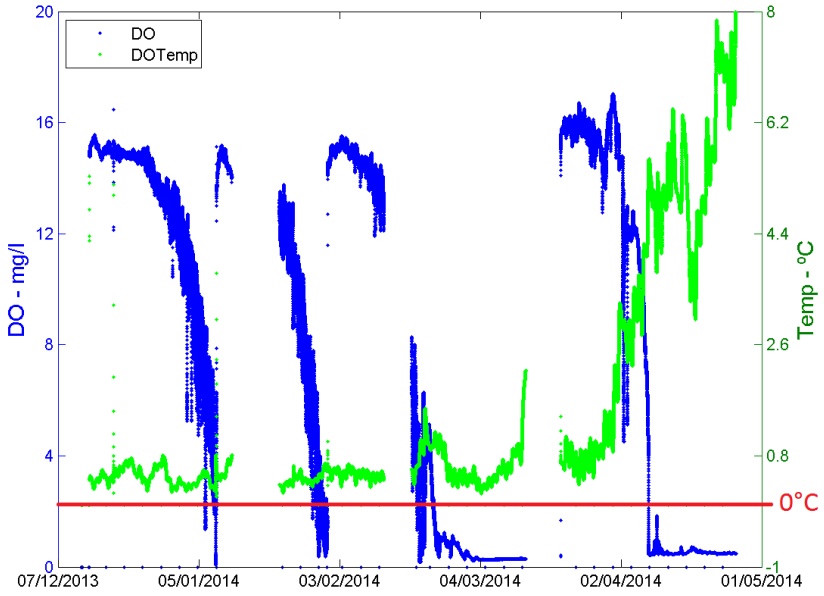


Figure 4.12: Example of low temperature at the Wemotaci lagoon case study.

Figure 4.13: Protection for the cable connections at the Wemotaci F/AL.

4.2.2 Maintenance protocol

Since the water is heavily polluted, a strict maintenance schedule is required to collect good quality data. Experience has shown that this maintenance schedule is not according the one proposed by the manufacturers.

Two different maintenance procedures are proposed depending on whether the validation and the calibration of the sensor have to be done on site by a technician or researcher, or they

(a) Tent installed with a butane on the station's house at the inlet of the Wemotaci F/AL.

(b) Small house built at the inlet of the Grandes-Piles F/AL.

Figure 4.14: Housing to work safely under severe environmental conditions.

can be done remotely from another computer. In case of on site validation and calibration, the procedure to be followed is presented in figure 4.15. This procedure is applicable to the pH, conductivity, Solitax, LDO and ammo::lyser sensors. In case of the spectro::lyser sensor, for which the validation and calibration can be done remotely, the procedure followed for maintenance activities is shown in figure 4.16.

Depending on the location of the sensors, a cleaning and control check have to be done more or less frequently. In this case study, two different schedules have been proposed: one for the more heavily polluted affluent of a F/AL (see table 4.4) and another one for the clean effluent of a F/AL (see table 4.5).

Table 4.4: Schedule of cleaning, validation and calibration proposed for a F/ALs inlet.

Sensor	**Cleaning**	**Validation**	**Calibration**
pH	Weekly	Weekly	Monthly
Turbidity	Every 2 weeks	Monthly	Yearly
Conductivity	Weekly	Weekly	Monthly
ammo::lyser	Weekly	Weekly	Monthly
spectro::lyser	Weekly	Weekly	Local every 2 months

Table 4.5: Schedule of cleaning, validation and calibration proposed for a F/ALs outlet.

Sensor	**Cleaning**	**Validation**	**Calibration**
pH	Every 2 weeks	Every 2 weeks	Every 6 weeks
Turbidity	Monthly	Monthly	Yearly
Conductivity	Every 2 weeks	Every 2 weeks	Every 6 weeks
DO	Every 2 weeks	Every 2 weeks	Every 6 weeks
ammo::lyser	Every 2 weeks	Weekly	Monthly
spectro::lyser	Every 2 weeks	Weekly	Local every 2 months

The cleaning frequency is based on the observed differences between the measured value before

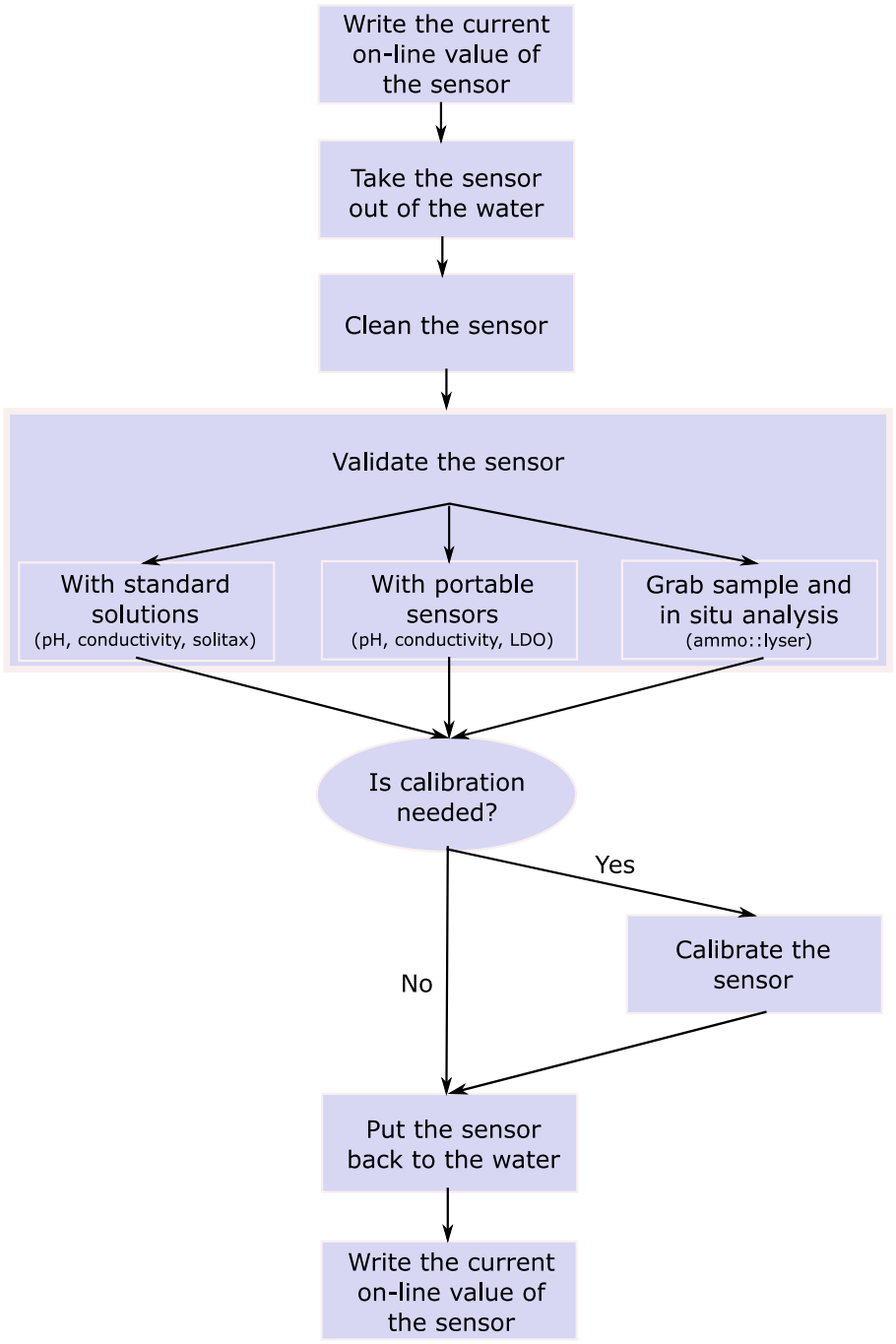


Figure 4.15: Schema of the maintenance protocol for on site calibration and validation for the pH, conductivity, LDO, Solitax and ammo::lyser sensors.

and after the cleaning action. To determine whether this difference is significant, control charts have been built following the same procedure presented in section 3.3.1 [1] (e.g. figures 4.17 and 4.18). To fix the limits several values with a significance level lower than 10 %.

[1]The procedure to build the control charts to evaluate the cleaning effect is detailed in section B.

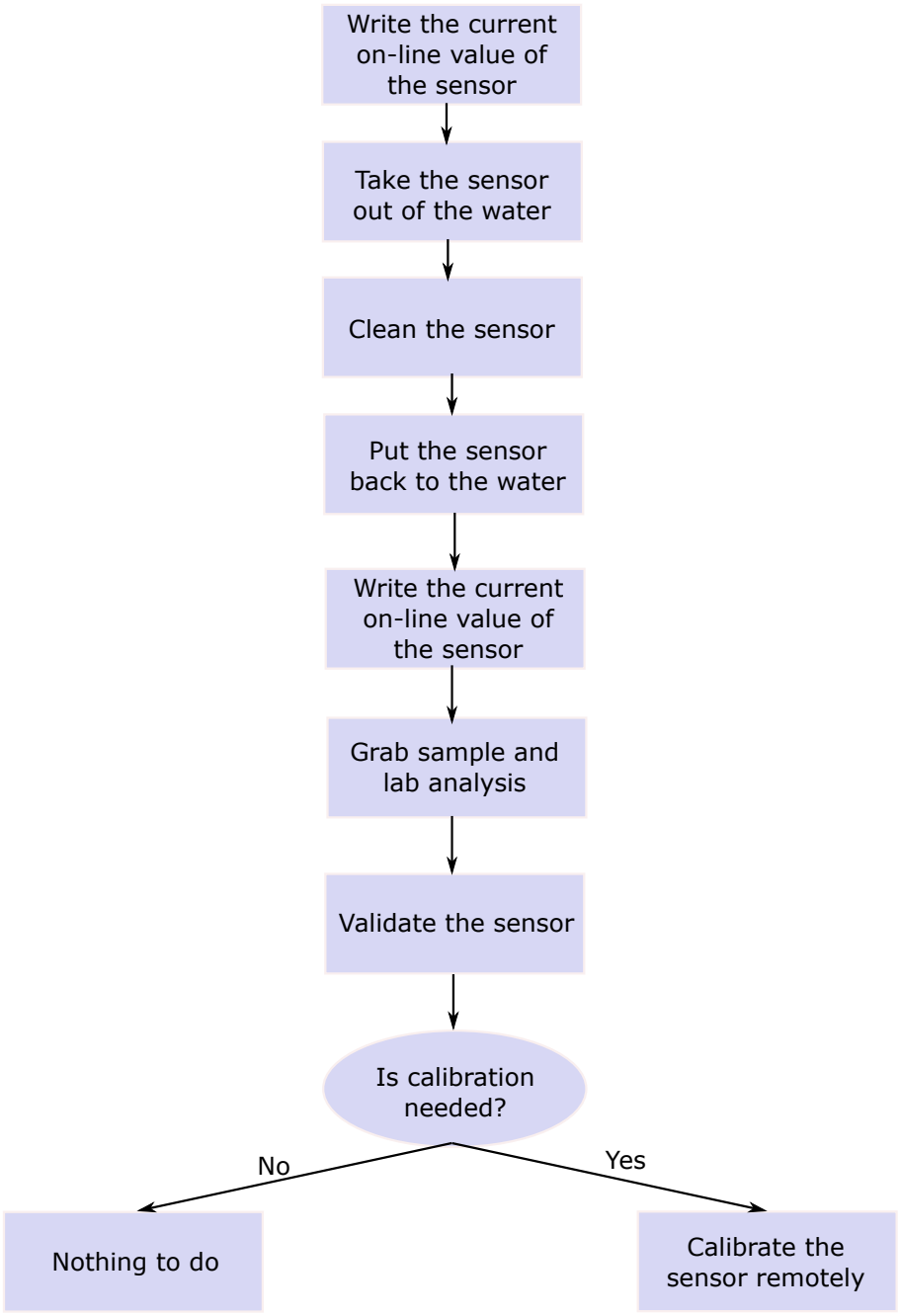


Figure 4.16: Schema of the maintenance protocol for spectro::lyser sensor.

When the difference is too large, i.e. it is outside the acceptability limits, it has to be concluded that cleaning should be done more frequently. Meanwhile, in the opposite case, the cleaning frequency can be reduced.

For example, in figure 4.19, where the cleaning activities are carried out monthly or every two months, it is very easy to observe the effect of cleaning on the DO measurements. It

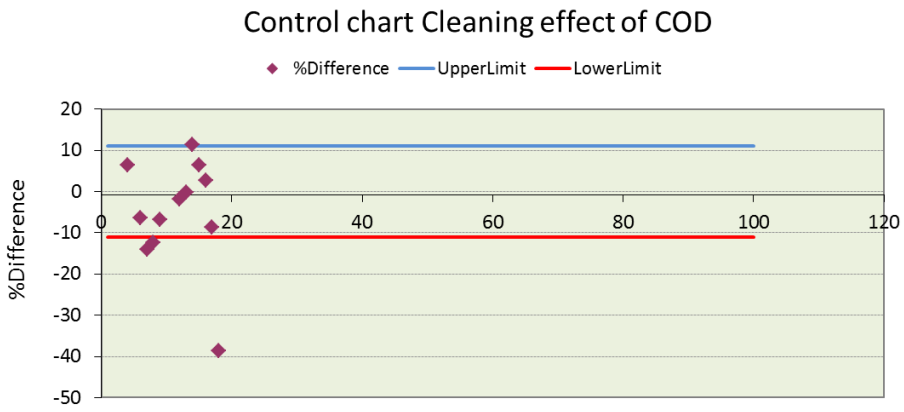


Figure 4.17: Control chart to evaluate the cleaning effect for the COD at the outlet in Grandes-Piles.

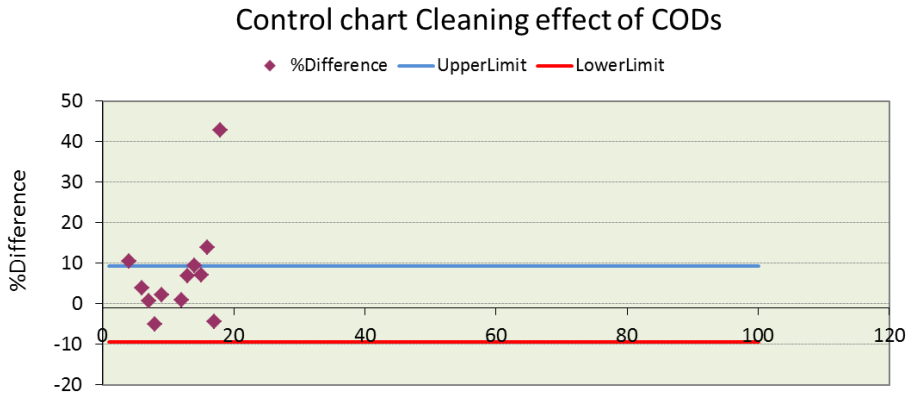


Figure 4.18: Control chart to evaluate the cleaning effect for soluble COD measured with the spetro::lyser installed at the outlet of the Grandes-Piles F/AL.

means that the cleaning frequency is insufficient. However, for the temperature parameter, the cleaning effect is insignificant. Thus, here a monthly cleaning interval is correct.

For the final validation and calibration schedule, the determined frequency has been estimated according to the developed control charts (see section 4.3.1) and depending on the fouling of the sensor.

For example, in figure 4.20, the impact of the calibration on the temperature parameter is highly significant. So a complete calibration should have been done earlier. Conversely, for the other three parameters, the difference is smaller, so the calibration could be done later.

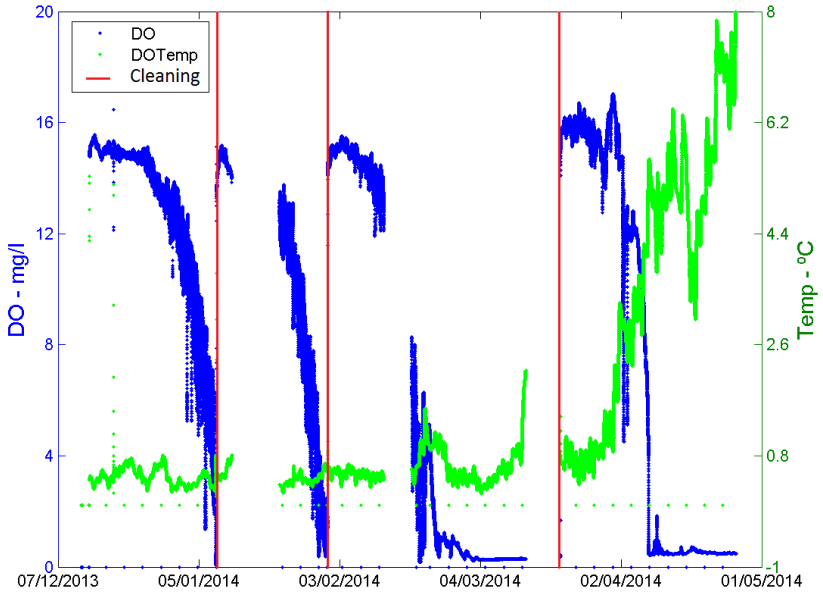


Figure 4.19: Six months of DO data from the outlet in Wemotaci indicating cleaning events with a vertical red lines.

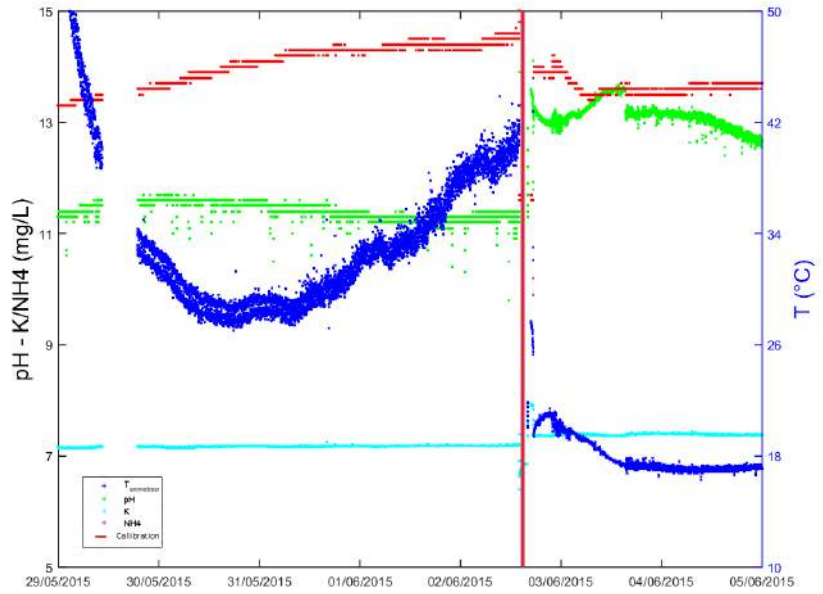


Figure 4.20: One week of filtered ammo:lyser data from the outlet in Grandes-Piles indicating a calibration activity with a vertical red line.

4.3 Validation of time series

According section 2.3, a good installation and a good maintenance are not enough to obtain good quality data series. Validation of the sensors and of these time series is needed to assure their quality. In this section, the application of the univariate method is presented.

4.3.1 Off-line validation

According the methodology presented in section 3.3.1, control charts have been developed to validate the sensor and to determine systematic sensor errors. Depending on the type of sensor, the control chart is based on standard solutions, reference values measured with a portable sensor or on grab samples analyzed with a reference laboratory method. In the three cases, the procedure to build the control chart is the same, the only difference being the procedure to obtain the reference value.

To build any type of control chart and determine the control limits, at least 20 points, dismissing poor values [2], should be used to have a representative distribution and assuring the reliability of the control limits to detect out-of-control situations. For some sensors, in this application the control chart is still under construction due to a lack of information, such as the value of the lab measurement or when the sample was taken. In the section below, some examples are presented.

Control charts based on standard solutions

Control charts based on standard solutions were constructed for pH, conductivity and Solitax sensors. In this case, in equation 3.8, μ is replaced by the reference value of the standard solution.

An example of a control chart based on standard solutions is presented in figure 4.21. In this case, there are not yet 20 points when the system is thought to be in control. Only 11 good points have been selected to determine the control limits as an example even if the control chart is not completed and not used as it should. Every time that the difference between the reference value and the measured value is larger than the control limits (outside the boundaries), a calibration has to be performed.

Despite the fact that this control chart has already been built and used, more "good" measurements should be included to more reliably set the limits.

[2] A value is considered poor when the difference between the reference value and the measured value with respect to the reference value represents more than the 5%.

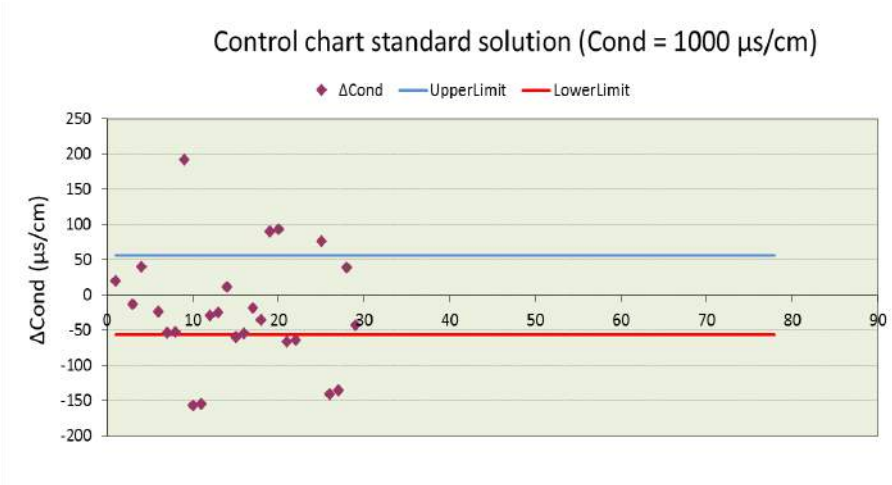


Figure 4.21: Control chart based on standard solutions to determine out-of-control situations for the conducivity sensor installed at the inlet of the Grandes-Piles F/AL.

Control charts based on reference values measured with portable sensors

To detect out-of-control situations for the LDO sensor, the control chart was based on reference values measured with a portable sensor. Also, for the pH and conductivity sensors, this type of control charts can be built.

In this case, the reference value in equation 3.8 is replaced by the value measured with the well-calibrated portable sensor. For example, in figure 4.22, the control chart built for a pH sensor is shown. All values outside the limit indicate that a calibration is needed. Unfortunately, the calibrations done were not successfully completed. Thus, a poorly performing pH sensor was the conclusion and a new probe had to be installed. Afterwards, a new control chart for the new probe had to be developed (Due to insufficient data, the new incomplete control chart is not shown).

Control charts based on grab samples

The last type of control charts is based on grab samples measured with a reference method. In equation 3.8, the difference is now calculated with respect to lab results. Some analyses can be done on site whereas for others an external lab is involved, depending on the type of analysis.

Due to a lack of lab measurements and a lack of metadata reported in the log file, these control charts are still under construction. However, some graphics have been made to visually compare how much the lab measurements differ from the on-line measurements, e.g. figures 4.23 and 4.24.

The first figure shows a comparison of the on-line ammonia and nitrate measurements with

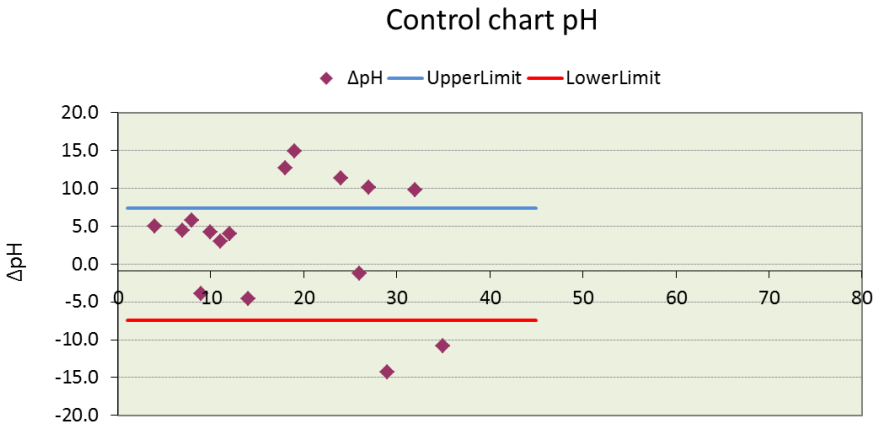


Figure 4.22: Control chart based on reference values measured with a portable sensor to determine out-of-control situations for the pH sensor installed at the inlet of the Grandes-Piles F/AL.

the lab measurements. Visually, it is possible to note a small difference and a calibration was not done that week. However, given the status of the control charts, it is not possible to determine whether the difference is statistically significant or not.

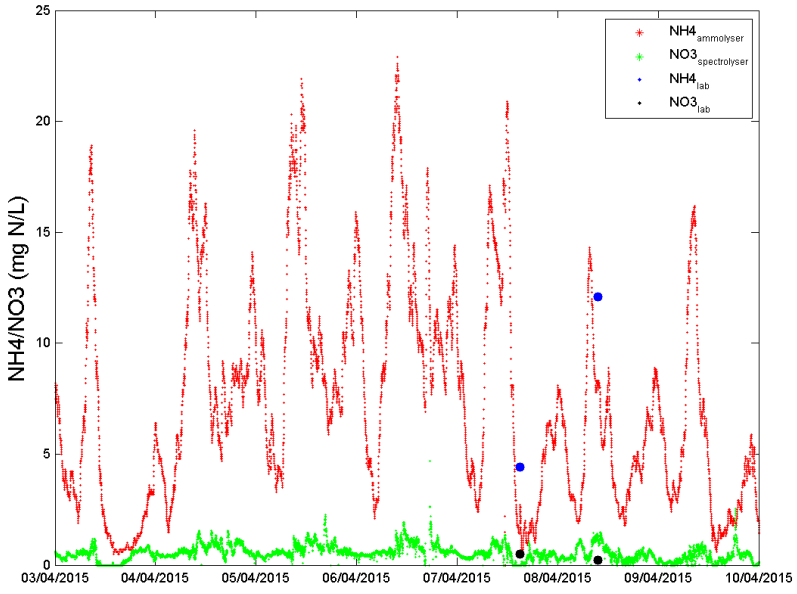


Figure 4.23: A week of raw and lab data of nitrogen parameters at the inlet of the Grandes-Piles F/AL.

In the second figure, a comparison with the on-line COD and COD_s measurements and lab measurements is presented. It is possible to observe considerable differences, so two calibrations were done for each parameter that week. Similar to the first figure, it cannot be stated whether the difference is significant or not.

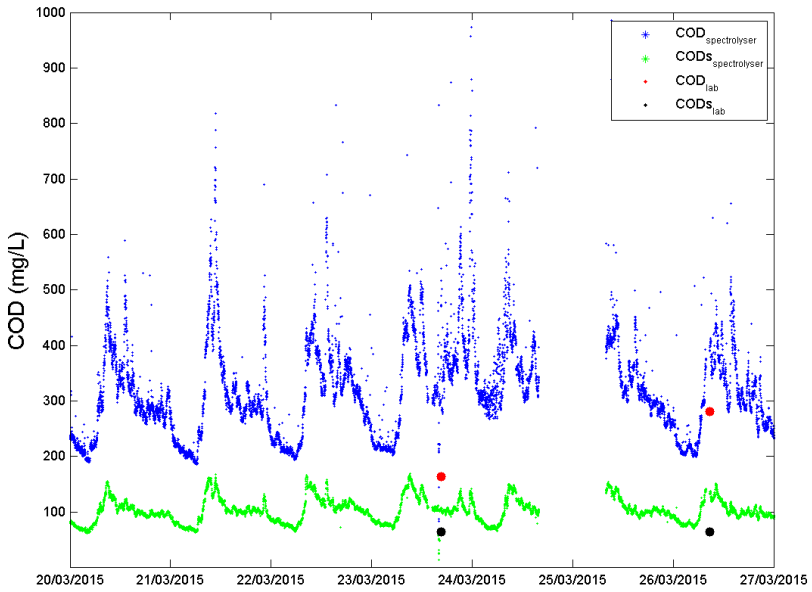


Figure 4.24: A week of raw and lab data of COD parameters at the inlet of the Grandes-Piles F/AL.

4.3.2 On-line validation

A recently developed automatic data quality evaluation method (Alferes et al., 2013) has been used to detect outliers and faults on the on-line data. According the methodology presented in section 3.3.2, an univariate model developed in a numerical computing environment has been applied. For every measured parameter, the following steps have been used to filter the data:

1. Select a good time series of data to calibrate the forecasting model.
2. Calibrate the model to determine α and α_{std}. If these values are similar for different calibrations, it does not matter what time series range is selected and which quantity of values was used for the calibration.
3. Select the data series to be validated.
4. Adjust the following parameters for the outlier detection method:

- The proportional constant K that drives the calculation of the prediction interval (Equation 3.23).
- The initial mean absolute deviation, Δ, used to start and reinitialize the outlier detection method.

As mentioned in section 3.3.2, the outlier detection method consists of an algorithm to detect out-of-control system situations. For that, it is necessary to fix:

- the number of consecutive rejected values needed to reinitialize the outlier detection method. When this number of consecutive rejected values is reached, an out-of-control situation is determined;
- the number of data before the last rejected value where the outlier detection method is reinitialized.

5. Apply the outlier detection method to generate the *accepted data* series.
6. Set the bandwidth for the Kernel smoother, h, used in equation 3.25. The Kernel function algorithm is designed to calculate the optimal bandwidth smoother automatically. However, in this case, it has been fixed manually according the experience of the researchers.
7. Apply the Kernel smoother method to obtain *smoothed data* series.
8. Estimate the data features for fault detection presented in table 3.2 and their acceptability limits.
9. Validate the data series.

These steps are performed until a good tuning has been found. Generally, the parameters are adjusted according the experience of the researcher. Because such expert-based approach leads to subjectivity, an effort should be made to come up with a tuning algorithm.

After having found a good tuning and adjustment for each parameter, data validation has been performed routinely and remotely on a weekly basis.

In the following sections, some examples are presented of successful and unsuccessful on-line data validation.

Examples of successful data validation

A validation is considered successful when after the model application, the outliers and the noise are removed, and faults are detected. For example, in figures 4.25 and 4.26, it is possible to observe the difference between the raw and filtered data for one week of the spectro::lyser parameters. On the first figure (Figure 4.25), the data is noisy and some outliers are observed.

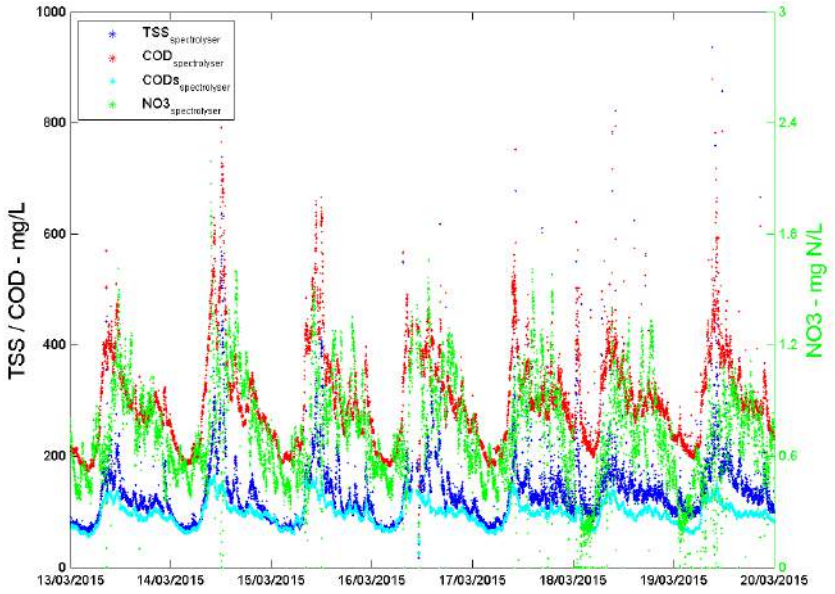


Figure 4.25: A week of raw data from the spectro::lyser installed at the inlet of the Grandes-Piles F/AL.

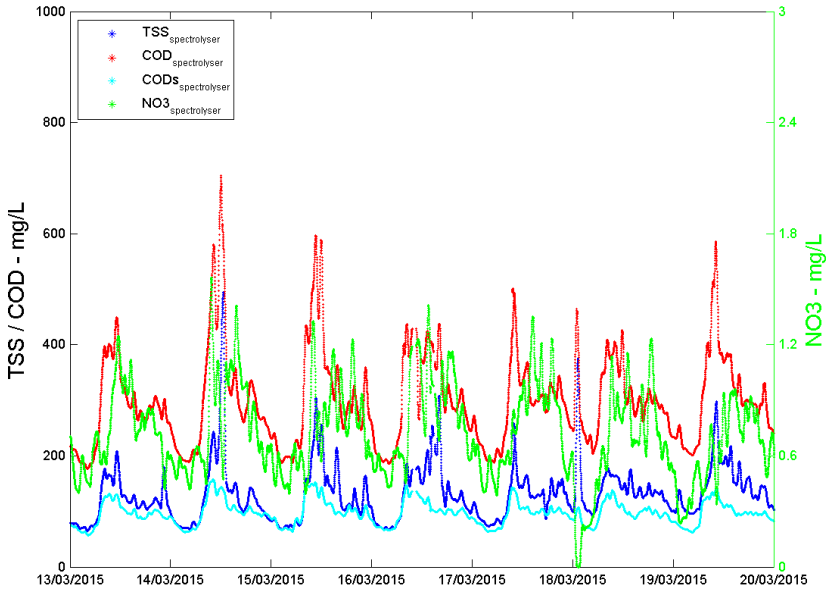


Figure 4.26: A week of filtered data from the spectro::lyser installed at the inlet of the Grandes-Piles F/AL.

On the second figure (Figure 4.26), the outliers and the noise are removed, and the performance of the system can be analyzed deeply.

Specifically, to determine whether a validation is successful, the results of the filtration and fault detection are studied. Below, the particular case of the successful on-line data validation for the COD_s at the inlet in Grandes-Piles for the time series presented above is presented.

Firstly, a graphic with the outliers, accepted and smoothed data is made, also indicating the forecast limits (see figure 4.27a). Also, out-of-control situations are determined (see yellow lines in figure 4.27b). When an out-of-control situation is detected, the model is reinitialized. This reinitialization permits a better description of the system dynamics, e.g. in the inlet where higher variations are observed.

On the 16^{th} of March, several outliers have been clearly observed. This abnormal behavior is due to a cleaning activity while the sensor was outside the water. So its detection and replacement have been done properly.

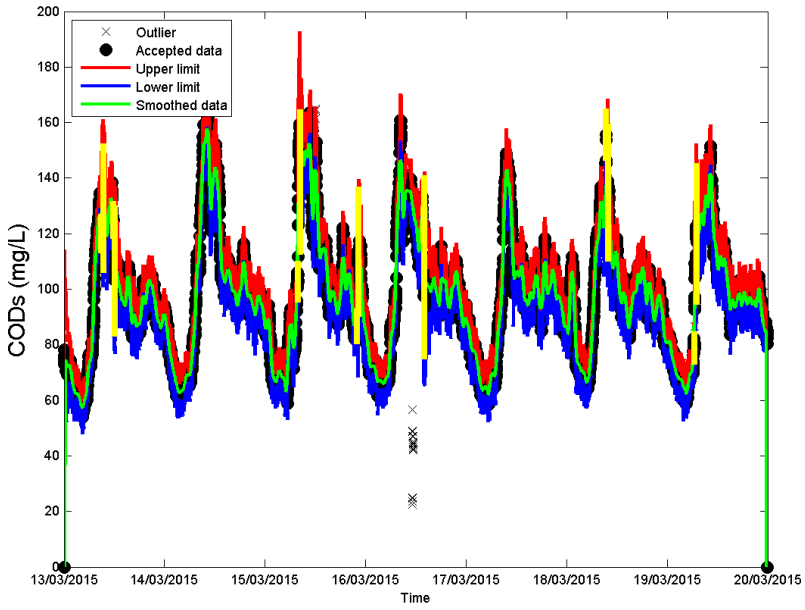
Secondly, after the application of the outlier detection methods and the smoother, some data features are calculated to detect faults. In figure 4.28, the studied features are presented. The first test is the number of the outliers detected and replaced (See figure 4.28 (a)). It permits to study the quality of the smoothed data and its features. The biggest peak of the % of replaced values also corresponds to the cleaning activity.

The runs test allows to determine the auto-correlation of the residuals (See figure 4.28 (b)). Most of the values pass this test, thus the forecasting model is adequately representing the raw data. The low slope during this period indicates that the dynamics of the variable are less important and acceptable (See figure 4.28 (c)). The last feature studied is the standard deviation of the residuals (See figure 4.28 (b)). Most of the data also pass the test. A peak is also observed around the 16^{th}, corresponding to a peak in the raw data with a large variation.

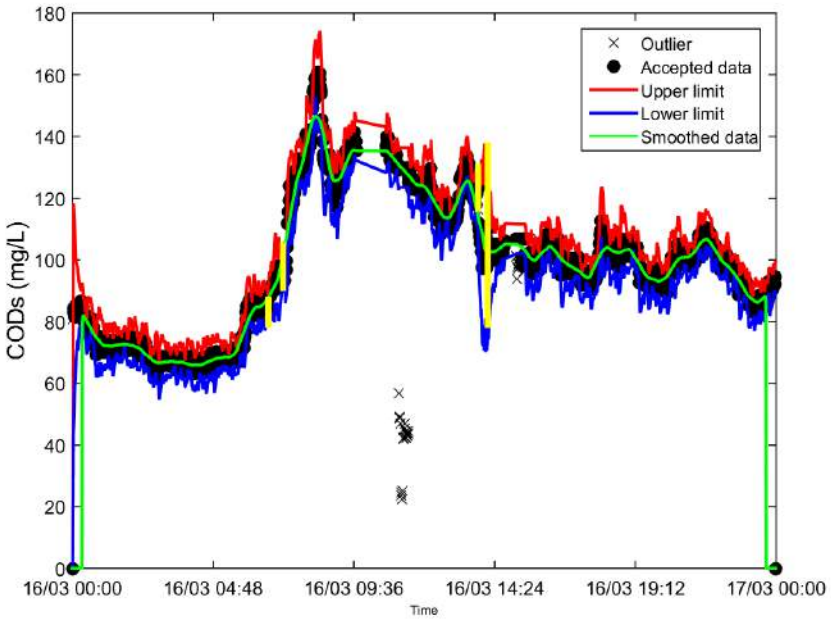
In conclusion, it is considered that the time series for this period is well validated.

Examples of unsuccessful data validation

An unsuccessful tuning of a data series is observed when after smoothing, the data is still noisy. For example, for the conductivity parameter, the tuning established under winter conditions cannot be used for the summer conditions. The on-line data is more noisy, and the smoother parameters have to be adjusted. In figure 4.29, a difference can be noticed before and after the snow melting around the middle of March. Before snow melting, the smoother parameters were adjusted with a good data series from winter period (e.g. figure 4.30). Afterwards, for the summer period, the data is more noisy and this tuning is not good and thus, the model parameters have to be readjusted (e.g. figure 4.31).



(a) A week of raw and filtered data with the red and blue forecast limits.



(b) Zoom in of one day of the raw and filtered data presented in 4.27b.

Figure 4.27: Raw and filtered data with the red and blue forecast limits of COD_s from the spectro::lyser installed at the inlet of the Grandes-Piles F/AL. (x) indicate outliers; (.) are the accepted values and yellow dots indicate out-of-control situations.

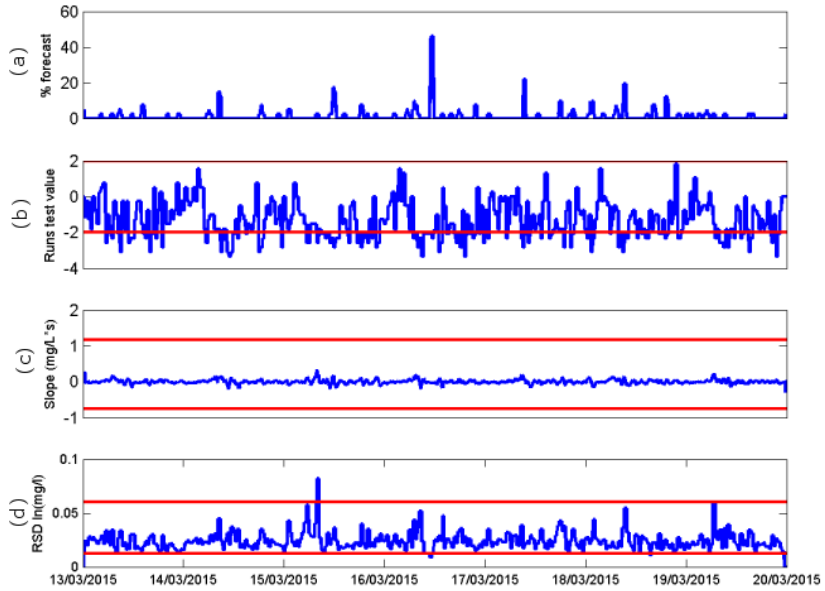


Figure 4.28: A week of fault detection data features of COD_s from the spectro::lyser installed at the inlet of the Grandes-Piles F/AL. (a) % of replaced data. (b) Residuals' runs test. (c) Slope. (d) Residuals' standard deviation.

Even if the tuning is unsuccessful, a comparison between raw and filtered data can be made (see figure 4.31). However, for the outliers detection and smoothing filtration plot, out-of-control situations are always present because the signal is very noisy (see figure 4.32). Also, at every out-of-control situation, the prediction limits are changing drastically. It is the reason why they are noisy too.

For the feature graphs to detect faults (Figure 4.33, a large % of data has been replaced by the forecast values (See figure 4.28 (a)). Moreover, only half of the data passed the runs test (See figure 4.28 (b)). It means that the forecasting model does not describe the raw data adequately. Furthermore, due to the noisy data, any slope can be detected and the result obtained is 0 (See figure 4.28 (c)). And for the standard deviation of the residuals, also have of the data pass the test (See figure 4.28 (d)).

In summary, when the tuning of the validation methods is not performed well, the filtered data may still be noisy, faults are detected continuously, and the forecast model cannot be considered representative. Thus, an exhaustive evaluation of these filtered data cannot be done to study system performance and the tuning parameters have to be readjusted until to obtain a successful data validation is achieved. However, once the tuning parameters are defined properly, these situations are rarely occurring.

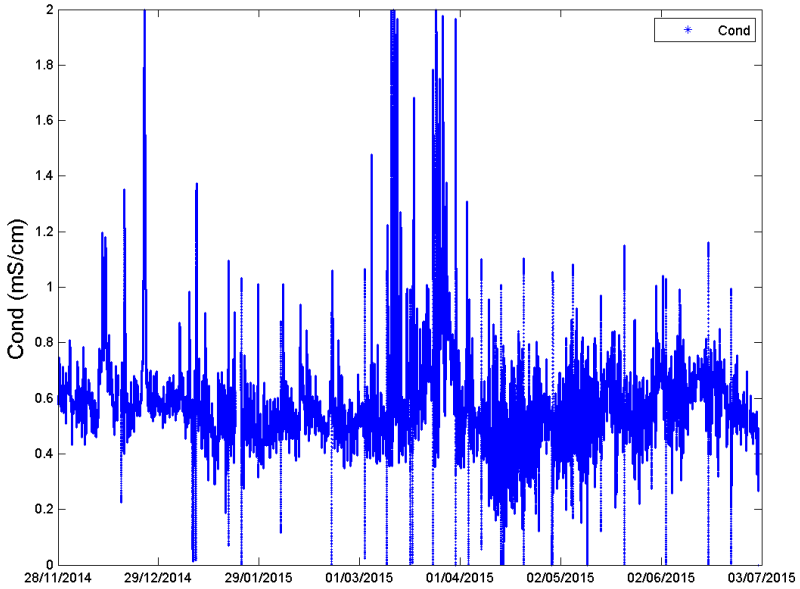


Figure 4.29: Seven months of filtered data from the conductivity sensor installed at the inlet of the Grandes-Piles F/AL.

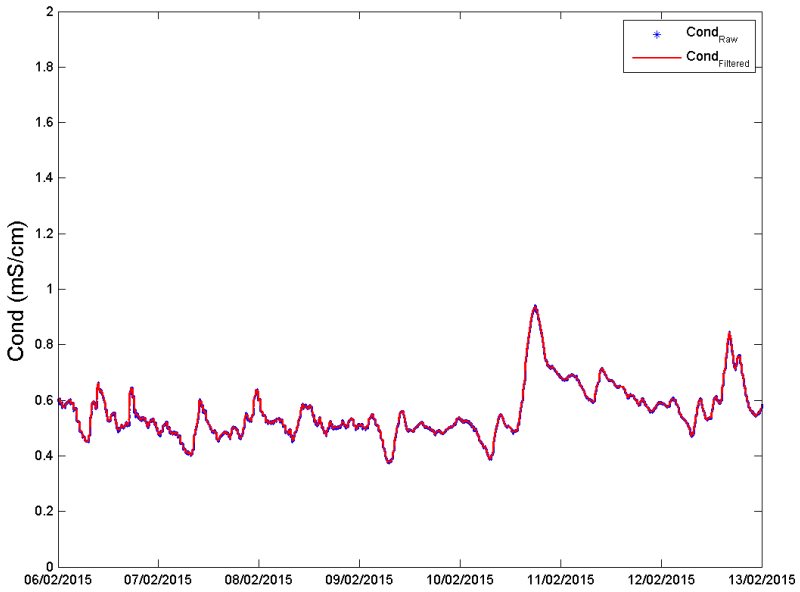


Figure 4.30: A week of raw and well filtered data from the conductivity sensor installed at the inlet of the Grandes-Piles F/AL.

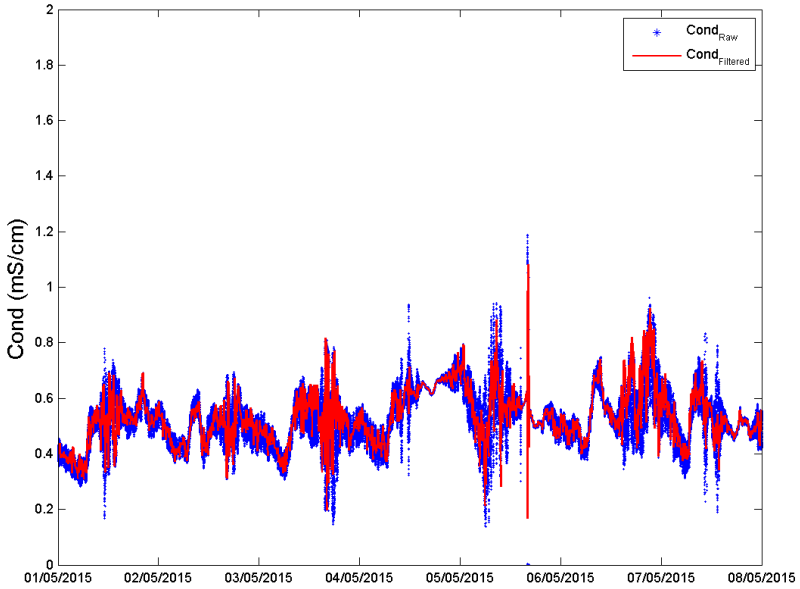


Figure 4.31: A week of raw and poorly filtered data from the conductivity sensor installed at the inlet of the Grandes-Piles F/AL.

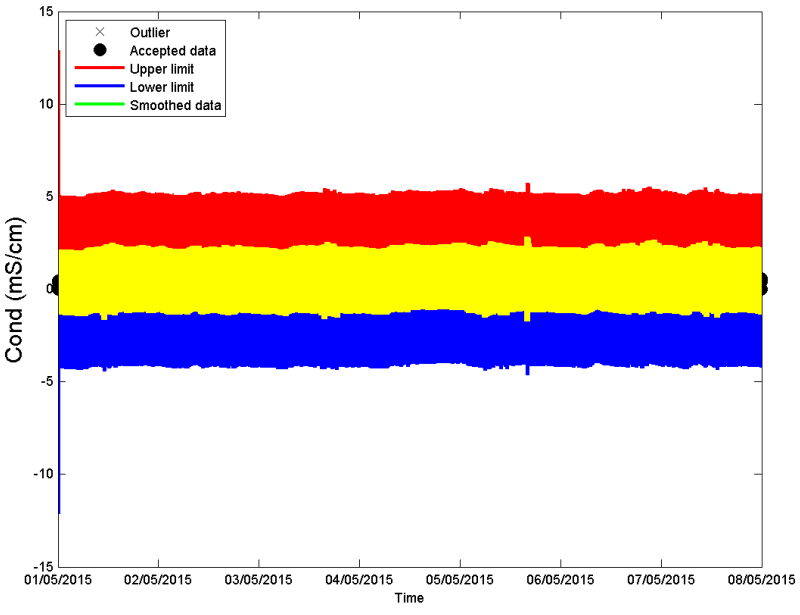


Figure 4.32: A week of raw and filtered data from the conductivity sensor installed at the inlet of the Grandes-Piles F/AL.

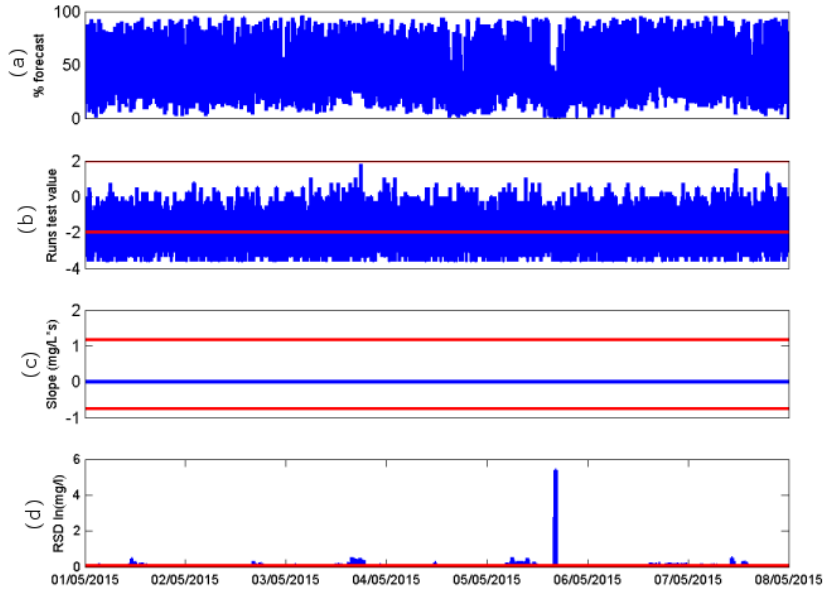


Figure 4.33: A week of feature data graphs from the conductivity sensor installed at the inlet of the Grandes-Piles F/AL. (a) % of replaced data. (b) Residuals' runs test. (c) Slope. (d) Residuals' standard deviation.

4.3.3 Validation by redundant measurements

Redundant measurements can also be used to detect errors as a multivariate method. When a sensor is not working properly, a deterioration can be detected by visually comparing the redundant measurements. For example, the redundant measurements available in this case study are the temperature and the pH. Both parameters are measured by different sensors because they are used to compensate the corresponding measurement of the sensor.

Figure 4.34 shows temperature measurements from four different sensors at the outlet of the Wemotaci F/AL. Even though some differences are observed, they are insignificant because these differences are 5 % or less. The average of all of them is similar. Thus, it is possible to consider that all temperature measurements are correct.

With redundant measurements it is also possible to compare the behavior of the sensors. If the dynamics of a sensor is not corresponding with the others, a bad or abnormal operation of one of the sensors may be deduced. For example, in figure 4.34, the temperature of the ammo::lyser, represented with a light blue line, is squashed compared with the others. This is due to the slow response time of the sensor.

The other parameter, which is measured by two different sensors, is pH. In figure 4.35, an example is given of a week of data from the inlet at Grandes-Piles. In this case, the difference

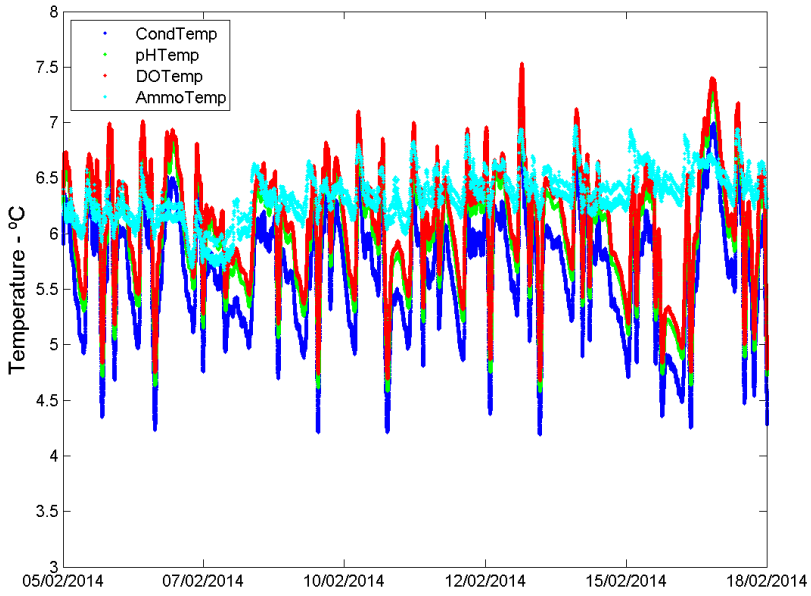


Figure 4.34: Two weeks of temperature measurements from four different sensors (pH, conductivity, LDO, ammo::lyser sensors) installed at the outlet of the Wemotaci F/AL.

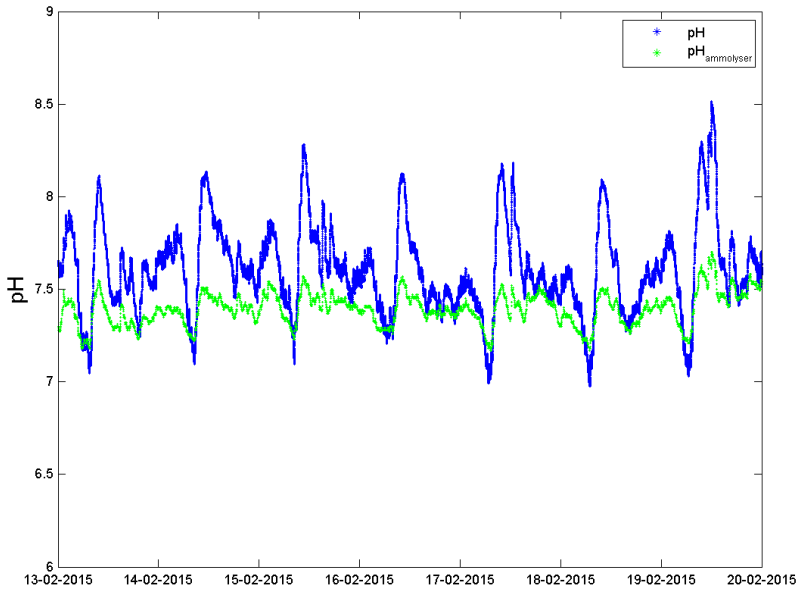


Figure 4.35: A week of pH measurements from the pH and ammo::lyser sensors installed at the inlet of the Grandes-Piles F/AL.

can be considered significant when the highest values are observed.

As mentioned above, the pH measurements of the ammo::lyser are squashed compared to the measurements from the Hach pH sensor. It is due to the the slow response time of the sensor. However, the average of the pH from the ammo::lyser is lower than the pH of the Hach pH sensor. In this case, since only two measurements can be compared, another validation of both probes should be done with portable sensors or standard solutions.

Even if bad or abnormal operation of one of the sensors may be deduced, further tests have to be done to verify which is the sensor working out of control.

4.4 Examples of data quality problems

Data quality problems can be observed, which can be either under our responsibility, e.g. due to a lack of maintenance activities, or out of our control, e.g. electricity failure.

To detect these problems, a general overview and interpretation was done on a weekly basis. This visual analysis allows detecting failures of the system and also helps interpreting the data later. Below some examples are presented.

Firstly, the most frequent problem is the lack of cleaning. A good example was already presented in figure 4.19. The LDO sensor installed in Wemotaci, was cleaned only monthly due to the far location of the installation. The difference of the DO measurement before and after the cleaning was really large, which allowed concluding that there was a lack of cleaning. However, in case of the temperature measurements, the difference before and after cleaning was negligible.

Another common problem is the lack of calibration. As shown in figure 4.20, this occurred with the ammo::lyser. The difference of the temperature parameter, before and after the calibration, is very clear. Thus, a calibration of the sensor should have been done earlier.

Both problems are due to a lack of maintenance of the installation and they impact directly the reliability of the measured data in a negative way. To avoid them in the future, following the maintenance schedule is basic. Also, a deep data evaluation, as well as the use of control charts, is highly recommended.

Other problems that are not under our control have also been observed. These problems are caused by system failures, e.g. electricity or compressor interruptions. In these cases, some data gaps of data have been observed.

When an electricity failure is produced, no sensor is working. So, during the failure period, no measurement is done and a gap on all data series is observed. For example, in figures 4.36 and 4.37, a period with no data is noticed. Since this gap appears in all data series from both

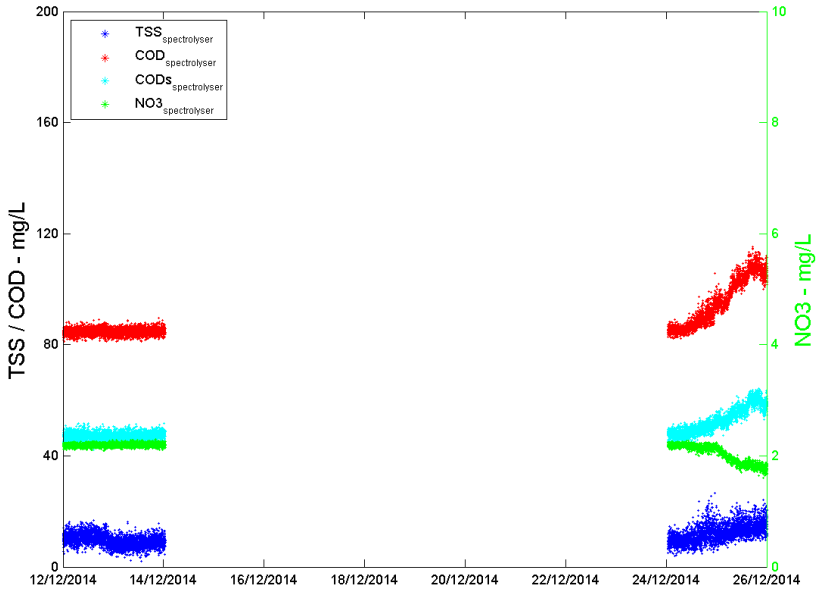


Figure 4.36: Raw data from the spectro::lyser installed at the outlet of the Grandes-Piles F/AL showing a data gap.

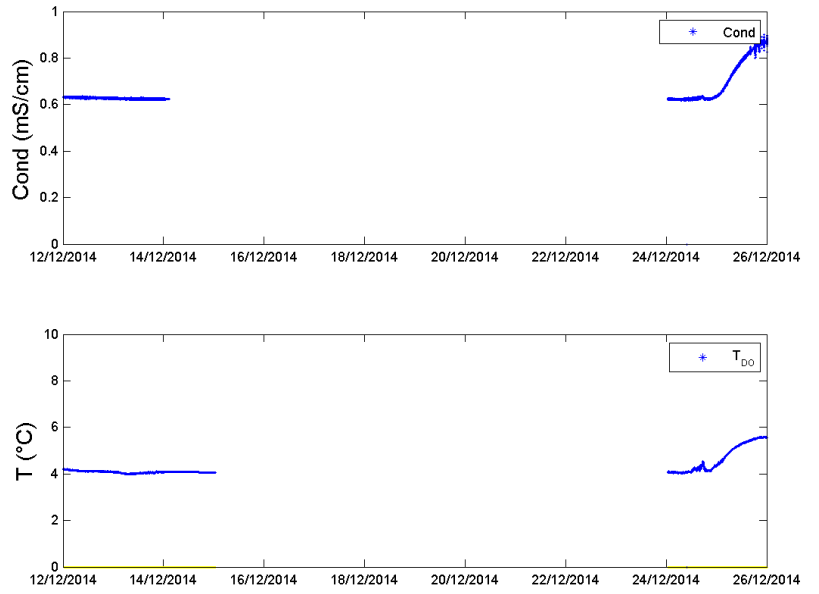


Figure 4.37: Raw data from the conductivity from the outlet of the Grandes-Piles F/AL showing a data gap.

softwares on that period, it could be concluded that an electricity failure occurred.

To prevent the interruption of system operation due to an electricity failure, an UPS can be installed to keep the RSM30 working for some hours. If the failure goes on for days, the UPS power will not be enough. Thus, a lack of data during these days will be observed too.

Finally, when the compressor which provides the air for the sensors' autocleaning fails, it can be detected by evaluating the data from the ana::pro software. The spectro::lyser fouls very quickly without air clean and stops measuring while the ammo:lyser is still working. For example, by comparing figures 4.38 and 4.39, a gap of data is appreciated in the spectro::lyser data series, while for the ammo:lyser, no gap occurs.

Even if a failure of a compressor is out of the control of the technicians and researchers, a regular maintenance of the compressor can be done to keep it safe and working properly. This includes draining the water in the compressor's tank and cleaning the air filters on a weekly basis.

During this study, any of these problems can be detected every two months. However, following the proper maintenance of the sensors and the stations, this frequency can be improved.

4.5 Observation and interpretation of lagoon system dynamics

By assuring that the AMS are working properly, and a good data validation is done, the performance of the system can be studied on the basis of the high quality validated data. In this section, some examples of the interpretation of observed system dynamics are presented and discussed.

4.5.1 Examples of daily profiles

A short term observation based on the AMS data consists in the study of a daily profile. For example, in figure 4.40, two days of COD_s data are presented to compare the inlet and the outlet of the Grandes-Piles F/AL. For the inlet, a high peak is observed between 9 a.m. and 10 a.m., corresponding with the activities of the inhabitants of Grandes-Piles. The rest of the day, generally the concentration of COD_s decreases with a small rise in the evening.

Regarding the outlet data, the concentration is lower than the inlet and hardly any significant variation is observed during these two days. This means that the lagoon is able to slightly reduce the COD (only 40 %, temperature in March is still very low, 6 °C) and also it softens the inlet concentration variations given the retention time of 6 days.

Other examples of daily profiles at the inlet during the same period are presented in figures 4.41, 4.42 and 4.43. In the first figure (Figure 4.41), the daily variations in NH_4^+ and K^+ present two different peaks. The first peak occurs again during the morning when discharges

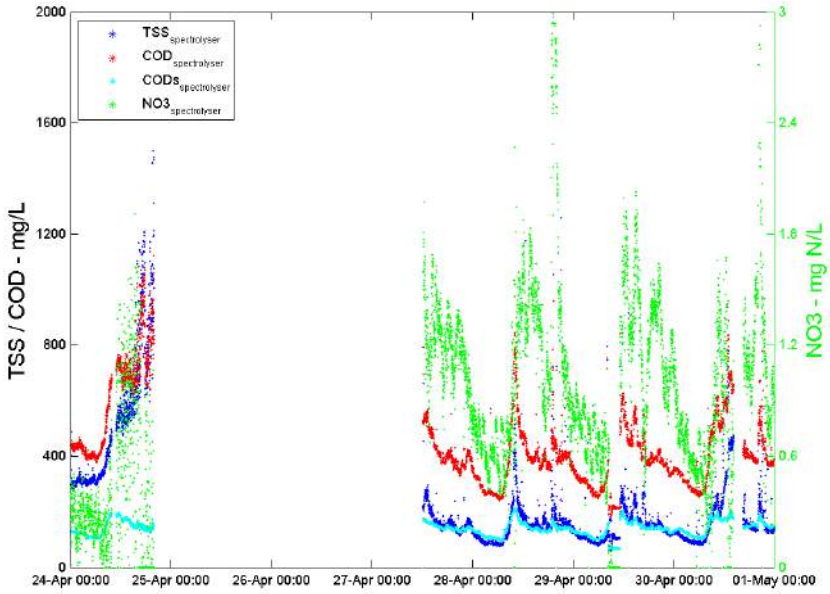


Figure 4.38: Raw data from the spectro::lyser installed at the inlet of the Grandes-Piles F/AL.

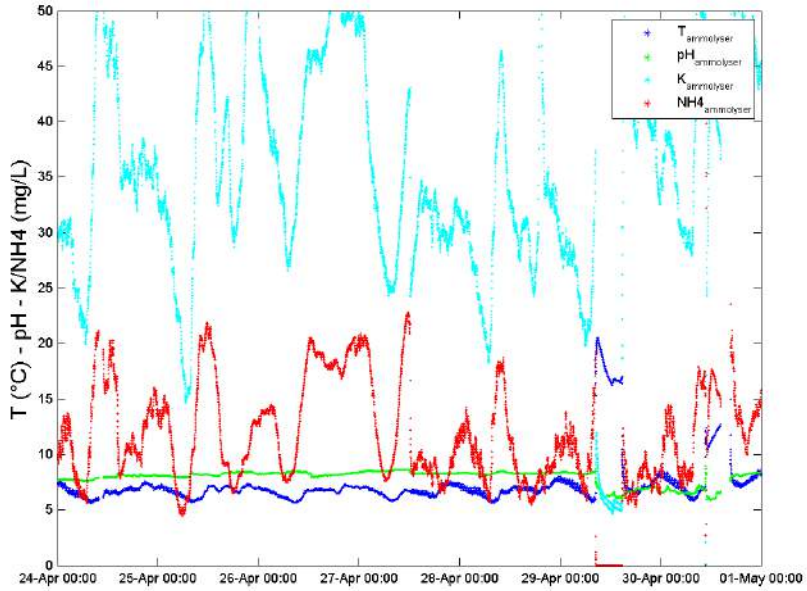


Figure 4.39: Raw data from the ammo::lyser installed at the inlet of the Grandes-Piles F/AL.

from toilet use occurs. The second peak occurs late in the evening when people go to the toilet before going to sleep.

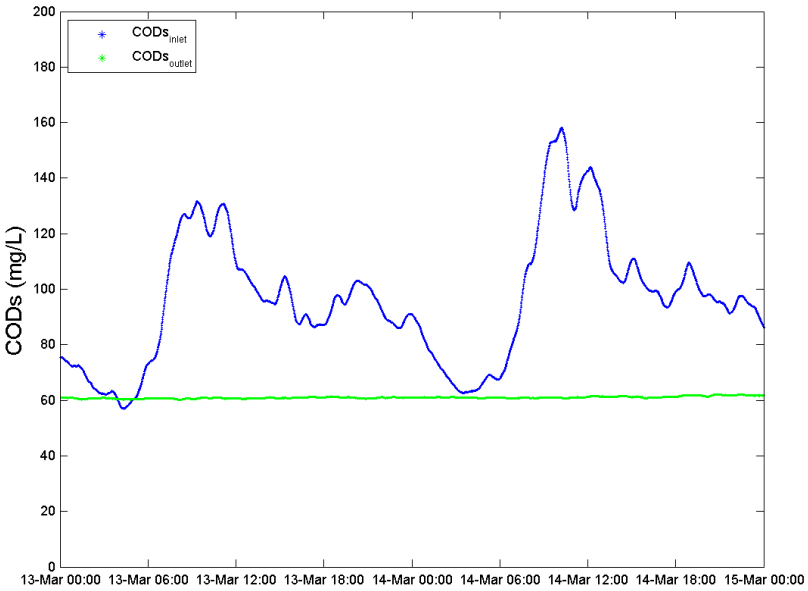


Figure 4.40: Two days of filtered COD_s data from the inlet and the outlet of the Grandes-Piles F/AL.

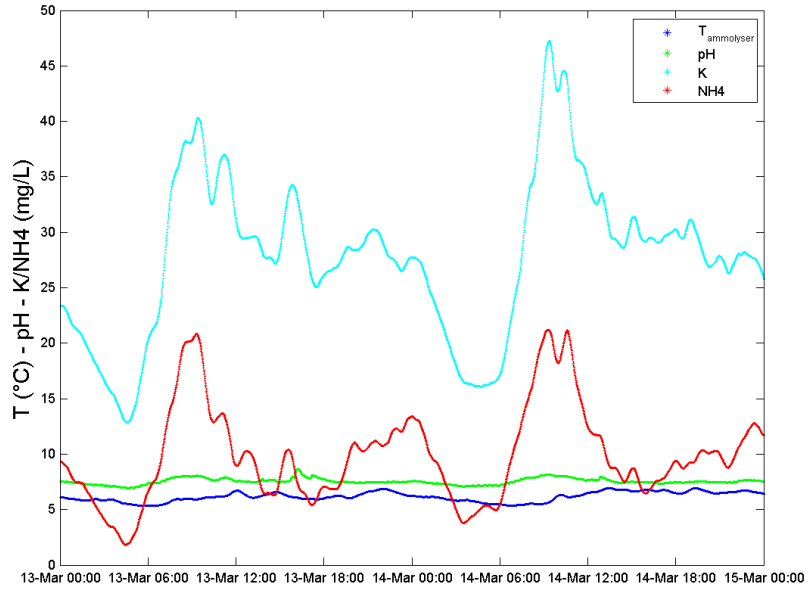


Figure 4.41: Two days of filtered ammo::lyser data at the inlet of the Grandes-Piles F/AL.

Regarding the temperature and the pH collected by the same sensor and presented on the same graph (Figure 4.41), some variations are observed but a daily profile cannot really be studied due to the scale. Hence, figure 4.42 shows a zoom in of both parameters and those from different sensors have been added. Again, a big rise is observed in temperature and pH during the morning and a smaller one during the evening. So, a similar behavior can be deduced as all parameters presented above.

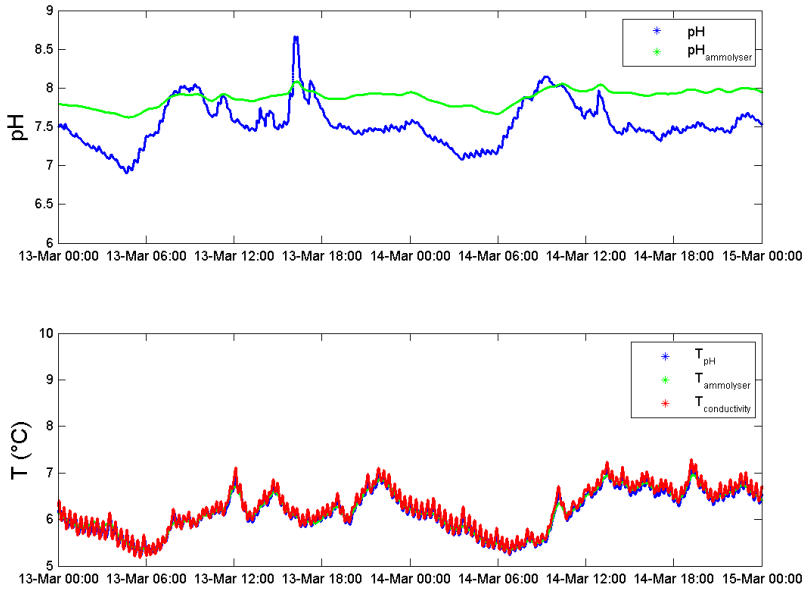


Figure 4.42: Two days of filtered pH and temperature data at the inlet of the Grandes-Piles F/AL.

However, not all parameters present the same daily behavior. For example, the conductivity does not vary the same way (see figure 4.43). A peak is observed during the afternoon, which is due to the temperature rise and snow melt that induced run-off of salt into the sewer system. A further explanation about these observations are presented in section 4.5.3.

4.5.2 Examples of weekly profiles

After understanding the daily profile, the weekly variations have been studied to determine whether there is a profile. Figures 4.44 and 4.45 show that the daily variations of COD are repeated continuously and following the same pattern, in Grandes-Piles and Wemotaci.

In figure 4.44, a small difference can be observed between weekdays and weekends. On the weekends the concentration is slightly higher than during the weekdays. This can be due to the lower inflow that provides the same load to the wastewater treatment plant, thus making

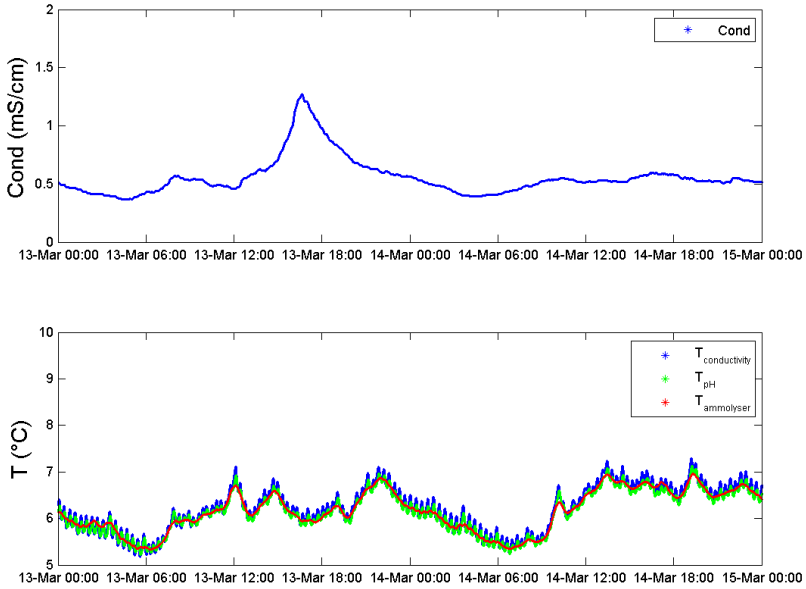


Figure 4.43: Two days of filtered conductivity and temperatures data at the inlet of the Grandes-Piles F/AL.

the COD concentration higher. This hypothesis should of course be confirmed with inflow data which were not available for this period of data.

In figure 4.45, this small difference in the COD concentrations between weekdays and weekends cannot be confirmed since the filtered data were not available. The presented graph with raw data and some remaining outliers can lead to an erroneous conclusion. However, the daily profile is clearly observed.

4.5.3 Examples of seasonal profiles

Long term on-line measurements allow studying some variations along the different seasons. Below, some examples are presented.

The first example, figure 4.46, shows the temperatures variation at the outlet of the Wemotaci F/AL for the whole period where the sensors were installed. Under winter conditions, the temperature at the outlet is around 0 °C. The temperature rise starts at the beginning of April. This is logical for Wemotaci because it is in the north of Québec and the winter is longer.

Also, in figure 4.46 shows that there is a bias on the temperature from the ammo::lyser. This temperature cannot be considered for further studies because it differs from the others and is below zero.

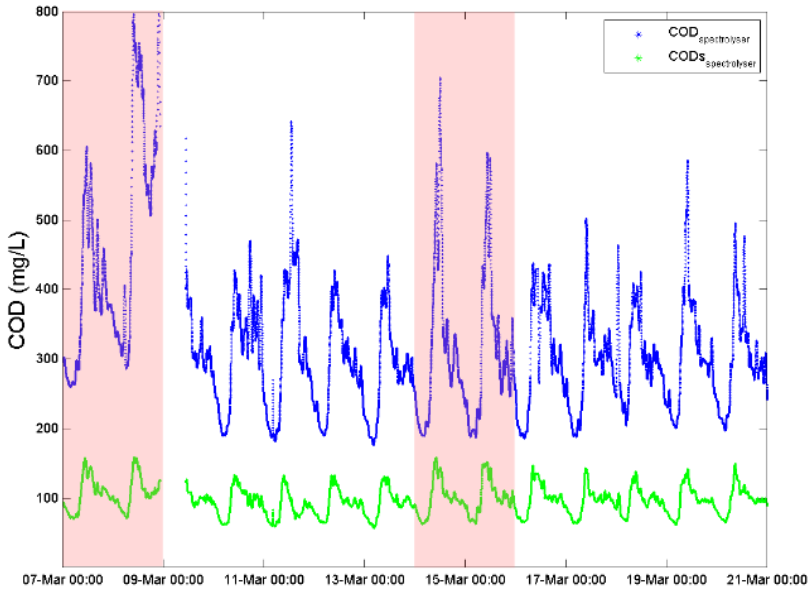


Figure 4.44: Two weeks of filtered COD and COD_s data from the inlet of the Grandes-Piles F/AL showing daily variations. The two red bands indicate the weekends.

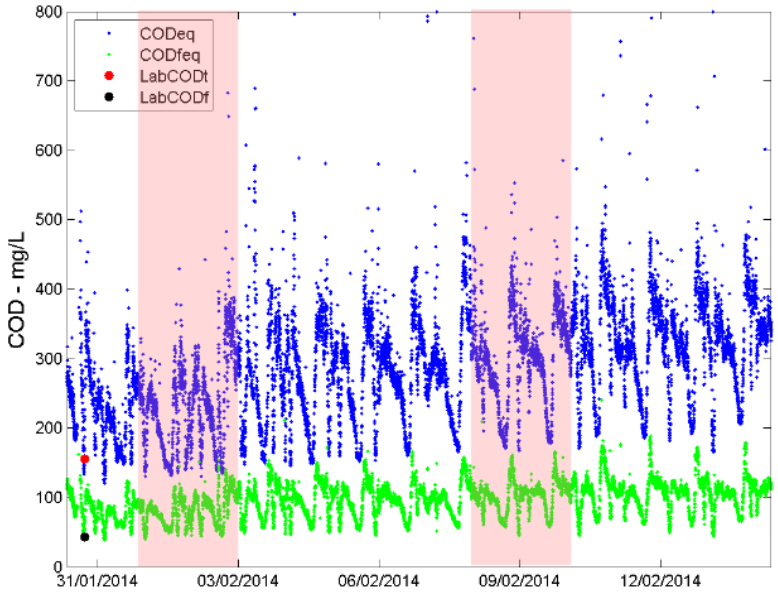


Figure 4.45: Two weeks of raw COD and COD_s data from the inlet of the Wemotaci F/AL showing daily variations. The two red bands indicate the weekends.

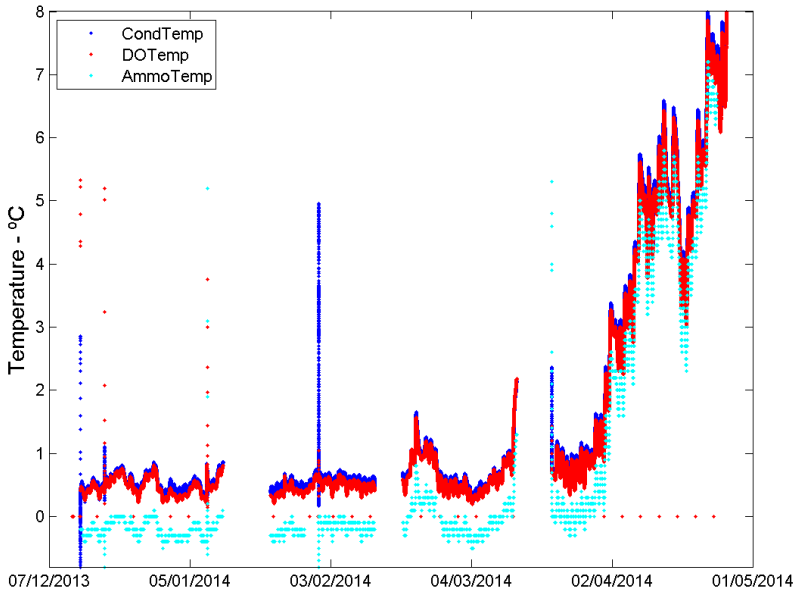


Figure 4.46: Five months of temperatures raw data from the outlet of the Wemotaci F/AL.

The second example, figure 4.47, depicts conductivity measurements from the inlet and outlet in Grandes-Piles during two weeks in March 2015. Considerable peaks are observed in both time series. On one side, at the inlet (blue line), the biggest peak is produced during the afternoon of March 11, 2015. On the other side, at the outlet (green line), the peak is observed during the early hours of March 12, 2015. And this peak is lower than the peak of the inlet. This is due to the mixing effect of the lagoon and the hydraulic retention time.

These peaks can be explained by the air temperature increase above 0 °C, between March 9, 2015 and March 11, 2015 (see figure 4.48). Due to the rise air temperatures, high snow melting rates are observed and salt that had accumulated on the streets during the winter arrives at the WWTP.

Finally, the last example represents the nitrogen parameters at the outlet of the Grandes-Piles F/AL (see figure 4.49). The concentration of NO_3 (green line) is affected by the temperature changes. At the end of autumn, it starts to decrease, and it is not until the end of spring, when it start to rise again. This is due to the lower nitrification activity at low temperatures and also the flow increases during the spring because of the snow melting and the nitrogen concentration is lower.

Regarding the concentration of NH_4^+ (red line), on-line measurements are only available from the beginning of May when a new ammo::lyser could be installed. The slope of these measurements is positive, it can be due to an increasing of the hydrolysis sludge caused by the air

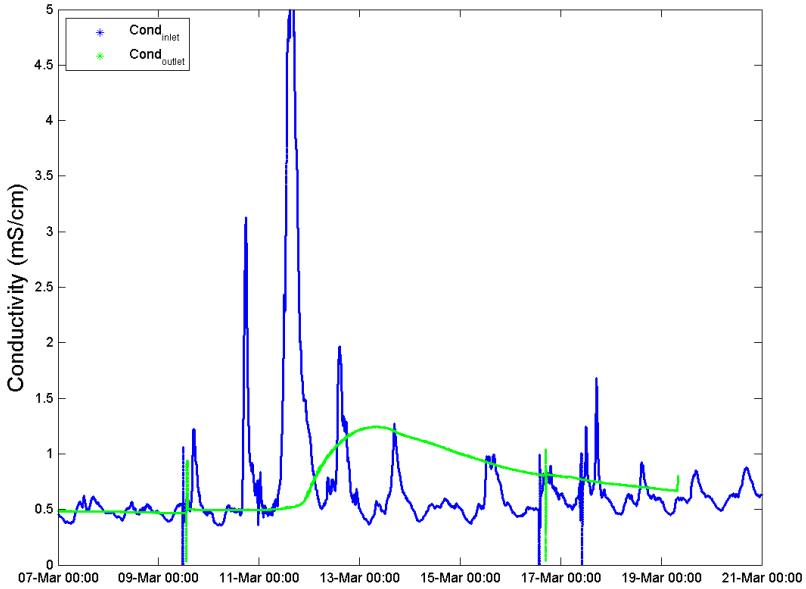


Figure 4.47: Two weeks of filtered conductivity data from the inlet and the outlet of the Grandes-Piles F/AL during the snow melting period.

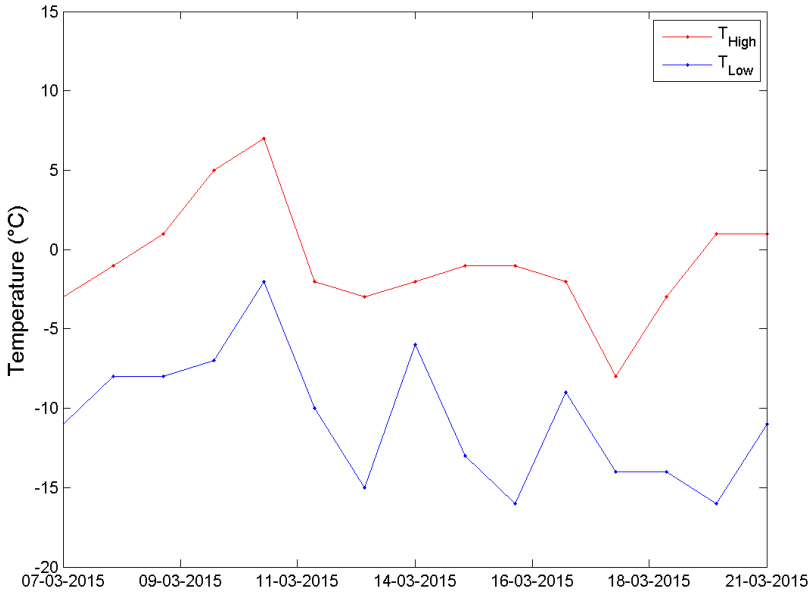


Figure 4.48: Two weeks of high and low daily air temperatures of the Grandes-Piles area during March 2015 (source Environment Canada).

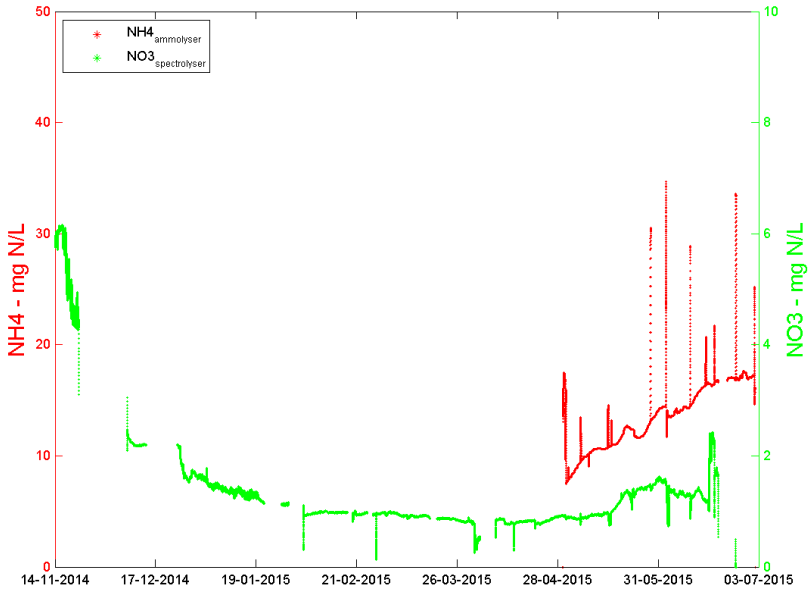


Figure 4.49: Seven months of filtered data of NH_4^+ and NO_3 from the outlet of the Grandes-Piles F/AL.

temperature rise. This makes the NH_4^+ concentration rise and it cannot be nitrified.

Chapter 5

Conclusions

The main conclusions of this thesis can be divided along the three objectives of this thesis: the database development and its application, maintaining the monitoring stations working properly, and the validation of the collected time series.

Firstly, an improved dat*EAU*base for environmental measured values has been developed. Also, it has been tested with some collected data. The following conclusions have been drawn:

1. A database is an essential tool to manage large amounts of data. Moreover, it allows storing the data in the same consistent format, giving quality to the content.
2. Any environmental parameter can be stored into the dat*EAU*base thanks to its design. Moreover, it offers flexibility and it can be modified and adapted for future studies.
3. The dat*EAU*base design is explicit to provide relevant metadata information to the measured values.
4. The metadata is fundamental to understand the measured values and use them for further studies. Moreover, a good storage of the metadata is important because:
   - all historical data should be documented
   - the personnel involved in data collection and data treatment can change
   - the person that is working in the field may not be the same as the person that is working with the data
5. The dat*EAU*base has been created for raw, filtered and lab data.
6. The MySQL software offers a large capacity of storage. In case the dat*EAU*base reaches the MySQL capacity limit, several databases may be linked.

7. Python is a powerful tool to develop a GUI, and it also allows basic statistic calculations and plotting time series.

   Secondly, the monitoring station have been kept working for a long period and under harsh winter conditions. The following conclusions have been made:

8. The installation and the supports of the sensors kept them working safely. Also, the installation of a heater during cold seasons kept them working inside the operation temperature ranges.

9. The installation of tents in Wemotaci, and the construction of small houses in Grandes-Piles, kept the technicians and researchers working reasonably comfortably during the whole study period.

10. A comprehensive maintenance protocol for F/ALs was proposed. It is essential to apply it imperiously to keep the sensors working properly.

11. An updated log file describing all maintenance events, problems, etc. is essential, especially, when several people are working with the stations. Without it, or not keeping it updated, complicates data understanding and a rigorous data evaluation. It is the basis of an important part of metadata.

    Finally, a validation of the time series has been done manually once a week. The following conclusions have been made:

12. Control charts help to detect systematic errors and assure the good operation of the sensors. Also, they increase the reliability of the sensors.

13. The lack of information on the log file and lab measurements prohibit the construction of control charts.

14. Good data validation treats can remove outliers and noise, and can detect faults. Thus, it assures the reliability of the mentioned smoothed data.

15. An automatic and continuous data validation could help to detect errors and problems earlier.

In summary, by having powerful tools to manage huge amounts of data, keeping the stations working and validating the data, reliable and representative water quality time series can be obtained. Afterwards, these time series are useful to study the performance and the dynamics of the system, in this case enhanced lagoon for wastewater treatment.

Chapter 6

Recommendations

Some recommendations for future work are proposed.

Firstly, the following advice for the datEAUbase is presented:

- Finish the GUI to facilitate the interaction between the user and the database, to protect the database according the user permissions, and to share the information between all project participants.

Secondly, recommendations for the installation and maintenance protocols for the continuation of the project, include:

- During the summer period, inside the housing, it was warm and humid. Installation of ventilation should be beneficial for the computer and avoid overheating failures.
- Keep the log file updated and completed. After a period of time, it is complicated to remember all events, failures, maintenance activities and it thus limits the data interpretation.

Thirdly, concerning the data validation procedures, the respective recommendations are:

- For any AMS application, the first months after the installation are essential to obtain enough data to build useful control charts, as well as establish an optimized maintenance schedule.
- The on-line validation is to be done routinely and manually once a week. Doing it more often and automatically could help to detect faults sooner, thus, allowing to fix faults earlier.

Finally, an overall recommendation for future studies is:

- to combine together the AMS, the datEAUbase, and the data validation tools. It will:
  - Store the raw data automatically into the datEAUbase
  - Facilitate the management of the large amounts of data
  - Be possible to validate the data automatically
  - Save time for the technicians and researchers

Bibliography

Alexis, P. and H. Bersini (2012). *Apprendre la programmation Web avec Python et Django : principes et bonnes pratiques pour les sites Web dynamiques*. Paris: Eyrolles.

Alferes, J., A. Lynggaard-Jensen, T. Munk-Nielsen, S. Tik, L. Vezzaro, A. K. Sharma, P. S. Mikkelsen, and P. A. Vanrolleghem (2013). Validating data quality during wet weather monitoring of wastewater treatment plant influents. *Proceedings of WEFTEC, Chicago, IL, October 3-9 2013*, 4507–4520.

Alferes, J., P. Poirier, C. Lamaire-Chad, A. K. Sharma, P. S. Mikkelsen, and P. A. Vanrolleghem (2013). Data quality assurance in monitoring of wastewater quality: Univariate on-line and off-line methods. In *Proceedings 11th IWA conference on instrumentation control and automation, Narbonne, France, September 18-20 2013*.

Alferes, J., P. Poirier, and P. A. Vanrolleghem (2012). Efficient data quality evaluation in automated water quality measurement stations. In R. Seppelt, A. Voinov, S. Lange, and D. Bankamp (Eds.), *6th International Congress on Environmental Modelling and Software, Leipzig, Germany, July 1 - 5 2012*, Leipzig, Germany.

Alferes, J., S. Tik, J. Copp, and P. A. Vanrolleghem (2013). Advanced monitoring of water systems using in situ measurement stations: data validation and fault detection. *Water Science and Technology 68*(5), 1022–1030.

APHA (1995). *Standard Methods for the Examination of Water and Wastewater*, Volume 9. Washington, DC, USA: American Public Health Association.

ASTM (1990). *ASTM Designation: D 3864 - 79 (reapproved 1990). Standard Guide for Continual On-Line Monitoring Systems for Water Analysis*. American Society of Testing and Materials (ASTM). Philadelphia, USA.

ATV-DVWK, Deutsche Vereinigung für Wasserwirtschaft, A. u. A. (2000). *Prozessanalysengeräte für N, P und C in Abwasseranlagen*. ATV-DVWK-Regelwerk. Hennef, Germany: GFA, Ges. zur Förderung der Abwassertechnik.

Aydin, D. (2007). A comparison of the nonparametric regression models using smoothing spline and kernel regression. *World Academy of Science, Engineering and Technology 36*, 253–257.

Bartram, J. and R. Ballance (1996). *Water Quality Monitoring - A Practical Guide to the Design and Implementation of Freshwater Quality Studies and Monitoring Programmes*. London, UK: United Nations Environment Programme and the World Health Organization.

BDSO (2014). Infrastructures municipales. Banque de données des statistiques officielles sur le Québec. http://www.bdso.gouv.qc.ca.

Beaupré, M. (2010). Characterization of on-line sensors for water quality monitoring and process control. Master's thesis, Université Laval, Québec, QC, Canada.

Berthouex, P. M. (1989). Constructing control charts for wastewater treatment plant operation. *Research Journal of the Water Pollution Control Federation 61*(9/10), 1534–1551.

Berthouex, P. M. and W. G. Hunter (1975). Treatment plant monitoring programs - preliminary analysis. *Journal Water Pollution Control Federation 47*(8), 2143–2156.

Bertrand-Krajewski, J.-L., S. Barraud, G. L. Kouyi, A. Torres, and M. Lepot (2007). Event and annual TSS and COD loads in combined sewer overflows estimated by continuous in situ turbidity measurements. *In: Proceedings of the 11th International Conference on Diffuse Pollution, Belo Horizonte, Brazil, August 26-31 2007*, 26–31.

Boudreau, V. (2011). *Standard Operating Procedure - Cleaning of the monEAU station sensors* (1 ed.). modelEAU: Université Laval. Québec, QC, Canada.

Bourgeois, W., J. E. Burgess, and R. M. Stuetz (2001). On-line monitoring of wastewater quality: a review. *Journal of Chemical Technology & Biotechnology 76*(4), 337–348.

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

Cai, Z. (2001). Weighted Nadaraya-Watson regression estimation. *Statistics & Probability Letters 51*(3), 307 – 318.

Camhy, D., V. Gamerith, D. Steffelbauer, D. Muschalla, and G. Gruber (2012). Scientific data management with open source tools–an urban drainage example. In *Proceedings IWA/IAHR 9th International Conference on Urban Drainage Modelling, Belgrade, Serbia, September 4-6 2012*.

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

Ceder, V. L., D. D. Harms, and K. McDonald (2010). *The quick Python book* (2nd ed ed.). Greenwich, Conn: Manning Publications.

Chapman, D. V. (1996). *Water Quality Assessment: A Guide to the Use of Biota, Sediments and Water in Environmental Monitoring* (2nd ed.). Abingdon ; New York : Taylor & Francis, cop. 1998.

Copp, J., E. Belia, C. Hubner, M. Thron, P. Vanrolleghem, and L. Rieger (2010). Towards the automation of water quality monitoring networks. In *Proceedings Automation Science and Engineering (CASE), IEEE Toronto, Ontario, Canada, August 21-24 2010*, pp. 491–496.

Corbitt, R. (1990). *Standard Handbook of Environmental Engineering* (2nd ed.). Washington, DC: McGraw-Hill.

Council of the European Union (1991). Council Directive 91/271/EEC of 21 May 1991 Concerning Urban Waste-water Treatment. Official Journal of the European Communities.

CUAHSI (2015). ODM Databases. https://www.cuahsi.org/. CUAHSI's Hydrologic Information System (HIS). Accessed: 2015-07-03.

Dandy, G. C. and S. F. Moore (1979). Water-quality sampling programs in rivers. *Journal of the Environmental Engineering Division-Asce 105*(4), 695–712.

Dippenaar, A., H. Wacheux, A. Mollon, H. Jansen, K. Edwards, T. Brueck, C. Cardone, H. Madiec, J. Schifini, W. Rogge, C. Charalambous, A. Adamou, L. Macek, R. Laukkanen, J. Schubert, W. Cheng, A. Lolli, G. Tavecchia, A. Wieczysty, P. dos Rios, M. Benoliel, A. Manescu, N. Stoica, A. Anton, G. Banatean, J. Vrabel, A. Dippenaar, N. Hu, and K. Edwards (2000). State of the art regarding on-line control and optimisation of water systems. In Nagle, P (Ed.), *22nd International Water Services Congress and Exhibition*, Volume 18 of *Water Supply: The Review Journal of the International Water Supply Association*, pp. 245–289. IWA Publishing.

Dochain, D. and P. Vanrolleghem (2001). *Dynamical Modelling and Estimation in Wastewater Treatment Processes*. London, UK: IWA Publishing.

DuBois, P. and S. T. B. Online (2013). *MySQL* (5 ed.). Developer's library. Addison-Wesley. Upper Saddle River, NJ, USA.

Duncan, A. (1967). *Quality control and industrial statistics*. D. Irwin. Washington, DC, USA.

EC (2003). Common implementation strategy for the water framework directive (2000/60/ec). planning process. European Communities.

EC (2012). Monitoring. http://www.ec.gc.ca/. Environment Canada. Accessed: 2015-06-05.

EPA (1977). *Aerobic Biological Wastewater Treatment Facilities. Process Control Manual.* U.S. Environmental Protection Agency. Washington, DC, USA.

EPA (2012a). An introduction to water quality monitoring. http://water.epa.gov. U.S. Environmental Protection Agency. Accessed: 2012-06-21.

EPA (2012b). Storet legacy data center. http://water.epa.gov. U.S. Environmental Protection Agency. Accessed: 2012-08-31.

EPA (2015). Clean Water Act (CWA) Compliance Monitoring. U.S. Environmental Protection Agency.

Gerardi, M. and B. Lytle (2015). *The Biology and Troubleshooting of Facultative Lagoons.* Wastewater Microbiology. Hoboken, NJ, USA: Wiley.

Government of Canada (2012). Wastewater Systems Effluent Regulations (SOR/2012-139). Accessed: 2015-04-06.

Grady, C., G. Daigger, N. Love, and C. Filipe (2011). *Biological Wastewater Treatment, Third Edition.* Boca Raton, FL, USA: Taylor & Francis.

Hach (2006a). *LDO Dissolved Oxygen Sensor. User Manual* (6 ed.). Loveland, CO, United States: Hach.

Hach (2006b). *pHD sc Digital. Differenctial pH/ORP Sensors* (4 ed.). Loveland, CO, United States: Hach.

Hach (2008). *3700sc Digital Conductivity Sensor. User Manual* (5 ed.). Loveland, CO, United States: Hach.

Hach (2012). *sc1000 controller. User Manual* (5 ed.). Loveland, CO, United States: Hach.

Hach, C. C. (2013, June). Principles of surface scatter turbidity measurement. Hach, Loveland, CO, United States.

Häck, M., D. Wedi, and W. Marx (1999). Analytische Qualitätssicherung für die Prozessmesstechnik. *Korrespondenz Abwasser 46*(9), 1–7.

Hardle, W. (1990). *Applied Nonparametric Regression*, Volume 27. Cambridge, United Kingdom: Cambridge University Press.

Harmancioglu, N. B., M. N. Alpaslan, and V. P. Singh (1998). Needs for environmental data management. *Environmental Data Management 27*, 1–12.

Harmancioglu, N. B., O. Fistikoglu, S. D. Ozkul, V. P. Singh, and M. N. Alpaslan (1999). *Water Quality Monitoring Network Design.* Water Science and Technology Library. Kluwer Academic Publishers. Dordrecht, The Netherlands.

Harremoës, P., A. G. Capodaglio, B. G. Hellstrom, M. Henze, K. N. Jensen, A. Lynggaard-jensen, R. Otterpohl, and H. Soeberg (1993). Waste-water treatment plants under transient loading - performance, modeling and control. *Water Science and Technology 27*(12), 71–115.

Holmes, M. E. and G. C. Poole (1998). *Management of a long-term water quality database: FlatDat for the Flathead lake biological station*, Volume 1, Chapter 18, pp. 111. Long Term Ecological Research Network Office, Department of Biology, University of New Mexico.

Holovaty, A. and J. Kaplan-Moss (2009). *The Definitive Guide to Django Web Gevelopment Gone Right* (2 ed.). The experts's voice in Web development. Berkeley, CA: Apress. New York, NY, USA.

Holovaty, A. and J. Kaplan-Moss (2015). The django book. Apress. Boston, MA, USA. Accessed: 2015-04-26.

ISA (1982). *ISA-37.1-1975 (R1982). Electrical Transducer Nomenclature and Terminology.* Durham, NC, United States: ISA (International Society of Automation).

ISO (1994). *ISO 5725-1 Accuracy (trueness and precision) of measurement methods and results - Part 1: General principles and definitions.* ISO (International Organizations for Standardization). Geneva, Switzerland.

ISO (2003). *ISO 15839 Water quality - On-line Sensors/Analysing Equipment for Water - Specifications and Performance Tests.* ISO (International Organizations for Standardization). Geneva, Switzerland.

Jeppsson, U., J. Alex, M. N. Pons, H. Spanjers, and P. A. Vanrolleghem (2002). Status and future trends of ICA in wastewater treatment - a European perspective. *Water Science and Technology 45*(4-5), 485–494.

Kaelin, D., L. Rieger, J. Eugster, K. Rottermann, C. Baenninger, and H. Siegrist (2008). Potential of in-situ sensors with ion-selective electrodes for aeration control at wastewater treatment plants. *Water Science and Technology 58*(3), 629–637.

Karpuzcu, M., S. Senes, and A. Akkoyunlu (1987). Design of monitoring systems for water quality by principal component analysis a case study. In *Proceeding, INT, Symp. On Environmental Management (Environment)*, Volume 87, pp. 673–690. Istanbul, Turkey, 1987.

Kofler, M. and D. Kramer (2005). *The Definitive Guide to MySQL* (3 ed.). The expert's voice in open source. Apress. Berkeley, CA, USA.

Lafond, R. (2009). A compact and efficient technology for upgrade of canadian municipal aerated lagoons. *Influents Summer*, 46–47.

Lovett, G. M., D. A. Burns, C. T. Driscoll, J. C. Jenkins, M. J. Mitchell, L. Rustad, J. B. Shanley, G. E. Likens, and R. Haeuber (2007). Who needs environmental monitoring? *Frontiers in Ecology and the Environment 5*(5), 253–260.

Lynggaard-Jensen, A. (1999). Trends in monitoring of waste water systems. *Talanta 50*(4), 707–716.

MDDELCC (2001). *Guide pour l'étude des technologies conventionnelles du traitement des eaux usées d'origine domestique*. Ministère du Développement durable, de l'Environnement et de la Lutte contre les changements climatiques. Québec, QC, Canada.

Metcalf and Eddy (2003). *Wastewater Engineering : Treatment and Reuse* (4 ed.). McGraw-Hill. McGraw-Hill. Boston, MA, USA.

Montgomery, D. (2008). *Introduction to Statistical Quality Control* (6 ed.). Wiley. Hoboken, NJ, USA.

Montgomery, D. C. (1980). The economic design of control charts - a review and literature survey. *Journal of Quality Technology 12*(2), 75–87.

Mourad, M. and 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.

Nelson, L. S. (1984). The shewhart control chart - tests for special causes. *Journal of Quality Technology 16*(4), 237–239.

Nelson, L. S. (1985). Interpreting shewhart xbar control charts. *Journal of Quality Technology 17*(2), 114–11.

NMKL (1990). *Quality Assurance Principles for Chemical Food Laboratories*. NMKL rapport. Nordic Council of Ministers. Oslo, Norway.

Olshausen, B. A. (2000). Psc–129 sensory process. *Aliasing 2*, 1–6.

Olsson, G., H. Aspegren, and M. K. Nielsen (1998). Operation and control of wastewater treatment - a scandinavian perspective over 20 years. *Water Science and Technology 37*(12), 1–13.

Peng, X., F. Chao-yang, C. Hong-wen, L. Zhi-gang, L. Bin, and F. Ming-lei (2011). Establishment of Water Quality Monitoring Database in Poyang Lake. In Wu, Y (Ed.), *3rd International Conference on Environmental Science and Information Application Technology (ESIAT), Xian, China, August 20-21, 2011*, Volume 10 of *Procedia Environmental Sciences*, pp. 2581–2586.

Plana, Q. (2013). Efficient on-line monitoring of river water quality using automated measuring stations. Master's thesis, Universitat Politècnica de Catalunya. Barcelona, Spain.

Poirier, P. (2015). Outiles automatiques d'Évaluation de la qualité des données. suivi en temps réel de la qualité de l'eau d'un ruisseau urbain. Master's thesis, Université Laval, Québec, QC, Canada.

Primodal (2015a). Base station. http://www.primodal.com/. Accessed: 2015-05-21.

Primodal (2015b). Precisionnow software. http://www.primodal.com/. Accessed: 2015-05-21.

Primodal (2015c). Primodal RSM30. http://www.primodal.com/. Accessed: 2015-05-21.

Quevauviller, P., O. Thomas, and A. Van Der Beken (2007). *Wastewater Quality Monitoring and Treatment*. Wiley. Hoboken, NJ, USA.

Rieger, L., M. Thomann, A. Joss, W. Gujer, and H. Siegrist (2004). Computer-aided monitoring and operation of continuous measuring devices. *Water Science and Technology 50*(11), 31–39.

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

Saberi, A. (2015). Automatic outlier detection in automated water quality measurement stations. Master's thesis, Université Laval. Québec, QC, Canada.

s::can (2006). *Manual ana::pro* (Version 5.3 ed.). Vienna, Austria: s::can.

s::can (2007a). *ammo::lyser V1 Manual* (1 ed.). Vienna, Austria: s::can.

s::can (2007b). *Manual s::can spectrometer probe* (1.0 ed.). Vienna, Austria: s::can.

s::can (2012). s::can. http://www.s-can.at/. Accessed: 2012-08-02.

Schimek, M. (2013). *Smoothing and Regression: Approaches, Computation, and Application*. Wiley. Hoboken, NJ, USA.

Sheldon, W. M., C. Laporte, T. Douce, and M. Alber (2011). A coastal water quality metadata database for the Southeast U.S.A. In *Proceedings of the 2011 Georgia Water Resources Conference*. Warnell School of Forestry and Natural Resources The University of Georgia, Athens, GA, USA.

Takahama, T. and S. Sakai (2009). A comparative study on kernel smoothers in differential evolution with estimated comparison method for reducing function evaluations. In *Proceedings of the IEEE Congress on Evolutionary Computation, 2009. CEC'09.*, pp. 1367–1374.

Thomann, M. (2008). Quality evaluation methods for wastewater treatment plant data. *Water Science and Technology 57*(10), 1601–1609.

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

Thomas, O. and M.-F. Pouet (2005). Wastewater quality monitoring: On-line/on-site measurement. In *Water Pollution*, Volume 2 of *The Handbook of Environmental Chemistry*, pp. 211–226. Berlin / Heidelberg, Germany: Springer.

UNEP (2015). Global environment monitoring system (gems) water programme. http://www.gemstat.org. United Nations Environment Programme. Accessed: 2015-07-03.

USGS (2012). National water information system (nwis). http://qwwebservices.usgs.gov/. U.S. Geologycal Survey. Accessed: 2012-08-31.

van Griensven, A., V. Vandenberghe, J. Bols, N. De Pauw, P. Goethals, J. Meirlaen, P. Vanrolleghem, L. Van Vooren, and W. Bauwens (2000). Experience and organistation of automated measuring stations for river water quality monitoring. In Proceedings 1st IWA World Water Congress, Paris, France. July 3-7 2000.

Vanrolleghem, P. (1994). Building blocks for wastewater treatment process control: a review. Advanced course on Environmental Biotechnology. Delft, The Netherlands, May 25 - June 3 1994.

Vanrolleghem, P. A. (2010). *Integrated Assessment for Water Framework Directive Implementation: Data, Economic and Human Dimension* (1 ed.), Volume 2. IWA Publishing. London, UK.

Vanrolleghem, P. A. and D. S. Lee (2003). On-line monitoring equipment for wastewater treatment processes: state of the art. *Water Science and Technology 47*(2), 1–34.

Vesilind, P. (2003). *Wastewater Treatment Plant Design*, Volume 1. IWA Publishing. London, UK.

Wagner, R. J., H. C. Mattraw, G. F. Ritz, and B. A. Smith (2000). *Guidelines and Standard Procedures for Continuous Water-Quality Monitors: Site Selection, Field Operation, Calibration, Record Computation, and Reporting*. US Department of the Interior, US Geological Survey. Reston, Virginia, USA.

Wand, P. and C. Jones (1994). *Kernel Smoothing*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis. Boca Raton, FL, USA.

WHO (1963). International standards for drinking water. *International standards for drinking water WHO/PEP/91.2*, 68 p. World Health Organization, Geneva, Switzerland, New York 27.

WSDOT (2008). *Water Quality Monitoring Databse User's Guide* (1 ed.). Washington State Department of Transportation. Washington, DC, USA.

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

Appendix A

dat*EAU*base

In this chapter, a detailed overview of the dat*EAU*base is given.

Firstly, some basic concepts about a database are summarized to understand the dat*EAU*base details described to the following sections.

A.1 Database basic concepts

Generally, a database is built by several tables. All these tables are related between them through keys. Every table contains a primary key column that uniquely identifies each row in a table.

To relate the different tables, the links can be:

- **1:1** - One row of the first table is related to only one row of the second table
- **1:n** - One row of the first table is related to multiple rows of the second table. In the opposite direction, each row of the second table is related to only one row of the first table.
- **m:n** - Each row of the first table can be related to multiple rows of the second table and also in the opposite direction.

In the cases of 1:1 and 1:n relations, the primary key of one table becomes a column in the other table and this new column is called foreign key because it is not a primary key in this table but it is for the other. In the case of a m:n relation between two tables, the two primary keys of both tables are added a new separate table. This table is connected to the those tables with a 1:n link.

Furthermore, each column of a table in a database a data type has to be defined. In table A.1, the different data types used for the design of the dat*EAU*base are described.

Table A.1: Explanation of different data types used in the dat*EAU*base.

Data type	Description	Example
INT	Short for integer. Only numbers without decimal points	'56234'
FLOAT	Only decimal numbers can be entered	'12345.1236584'
DOUBLE	Same as FLOAT but more accurate	'12348987.3222156545798946'
DATE	The data must be entered in the date format	YYYY-MM-DD: '2015-08-01'
TEXT	Only text can be entered	'This is an example of TEXT.'
VARCHAR(100)	A string which is 100 bit long can be entered	'ABcde%256.66FG?'
TINYTEXT	Same as TEXT but shorter	'This is a short text.'

A.2 dat*EAU*base tables

In this section, a detailed description of the datEAUbase tables described in section 4.1 is presented. Also, how they are composed, which data types are used and in which format the data must be entered. The characteristics of each column illustrates if the column is a primary key, foreign key, not null (must be filled out) or auto increment (this column is filled out by MySQL automatically).

All collected data is stored in a table called *Value*. The composition of the *Value* table is shown in table A.2. All data stored in that table is related to the table called *Metadata* (Table A.3). Those two table are the heart of the database and all other tables are linked to them giving more detailed background information to the metadata.

In the *Metadata* table all important information about collected data is stored (See table A.3). In that table, only IDs are stored. Each ID corresponds to a table with detailed information.

If there is a particular comment to be added on a value, it has to be added in the *Comments* table (See table A.4). Afterwards, in the *Value* table, the Comments_ID is added to the corresponding value row.

All different measured and analyzed parameters are stored in the *Parameter* table with more detailed information about them. The comprised information is explained in table A.5.

In the *Unit* table which is explained in table A.6, all different kinds of units are stored. This table is linked to the *Metadata* table and also to the *Parameter* table.

All equipment information is stored in the *Equipment* table. The considered information to be included is explained in table A.7.

Table A.2: Detail of the *Value* table in the dat*EAU*base.

Table columns	Data type	Characteristic	Description
Value_ID	INT	Primary key, not null, auto increment	A unique ID is generated automatically by MySQL
Date	DATE	Not null	Date of collected data: 'YYYY-MM-DD'
Time	TIME	Not null	Time in 24h of collected data: 'hh:mm:ss'
Value	DOUBLE	Not null	Value of collected data
Number_of_experiment	TINYINT	Not null	Number of replica of an experiment
Metadata_ID	INT	Foreign key, not null	Metadata related to collected value. Link to the *Metadata* table
Comment_ID	INT	Foreign key	Comment of value. Link to the *Comments* table

In the *Equipment* table, only an ID of the model is stored. For further information about the model of the equipment is detailed in the *Equipment_model* table. The information comprised in that table is explained in A.8.

Measurement, cleaning, calibration and other procedures are established to be consistent with same methodology for future work. All of these procedures are identified and stored. Their information, explained in table A.9, has been included in the *Procedures* table.

To identify which equipment can measure a parameter, the ID of an equipment model and the concerning parameter ID have to be added in the table *Equipment_model_has_Parameter*. This table is only a relational ID table (See table A.10).

The *Equipment_model_has_Procedures* table allows to identify the related procedures to an equipment model, or viceversa. In this table, only the IDs of the equipment model and the procedures are included and related (See table A.11.

To identify which procedure is used to measure a parameter, or viceversa, their IDs are linked in the table *Parameter_has_Procedures*. This relational IDs table is presented in table A.12.

The different purposes of the measurements are stored in *Purpose* table. The goal of this table is to differ if it is an on-line measurement, a measurement on the field, in the lab, a cleaning process, a calibration, etc. The information included in that table is explained in table A.13.

All weather conditions are stored in the Weather_condition table which is explained in table A.14. For example, dry weather, wet weather or storm event. Each condition is accompanied with the corresponding description.

All sampling point where measurements are taken are stored in the *Sampling_point* table. The

Table A.3: Detail of the *Metadata* table in the dat*EAU*base.

Table columns	**Data type**	**Characteristic**	**Description**
Metadata_ID	INT	Primary key, not null, auto increment	A unique ID is generated automatically by MySQL
Parameter_ID	INT	Foreign Key	Measured parameter. Link to the *Parameter* table
Unit_ID	INT	Foreign Key	Unit of the parameter. Link to the *Unit* table
Purpose_ID	INT	Foreign Key	Purpose of the data collection. For example: Measurement, lab analysis, calibration or cleaning. Link to a the *Purpose* table
Equipment_ID	INT	Foreign Key	Equipment which was used. Link to the *Equipment* table
Procedure_ID	INT	Foreign Key	Procedure corresponding to the purpose and/or the equipment. Link to the *Procedure* table
Condition_ID	INT	Foreign Key	Weather condition during the measurement. Link to the *Weather_ condition* table
Sampling_point_ID	INT	Foreign Key	Sampling point where the data was collected. Link to the *Sampling_ point* table
Contact_ID	INT	Foreign Key, not null	Person who is responsible of the measurement. Link to the *Contact* table
Project_ID	INT	Foreign Key	Name of the project for which the data was collected. Link to the *Project* table

Table A.4: Detail of the *Comments* table in the dat*EAU*base.

Table columns	**Data type**	**Characteristic**	**Description**
Comment_ID	INT	Primary key, not null, auto increment	A unique ID is generated automatically by MySQL
Comment	TEXT	Not null	Comment on the data in the *Value* table

Table A.5: Detail of the *Parameter* table in the dat*EAU*base.

Table columns	Data type	Characteristic	Description
Parameter_ID	INT	Primary key, not null, auto increment	A unique ID is generated automatically by MySQL
Parameter	VAR-CHAR(100)	Primary key, not null	Name of the parameter
Unit_ID	INT	Foreign key, not null	SI-unit of the parameter. Link to the *Unit* table
Description	TEXT	Not null	Description of the parameter

Table A.6: Detail of the *Unit* table in the dat*EAU*base.

Table columns	Data type	Characteristic	Description
Unit_ID	INT	Primary key, not null, auto increment	A unique ID is generated automatically by MySQL
Unit	VAR-CHAR(100)	Not null	SI-units only

Table A.7: Detail of the *Equipment* table in the dat*EAU*base.

Table columns	Data type	Characteristic	Description
Equipment_ID	INT	Primary key, not null, auto increment	A unique ID is generated automatically by MySQL
Equipment_identifier	Not null	Identification name of the equipments	
Serial_number	VAR-CHAR(100)	Not null	Serial number of the equipment
Owner	TEXT	Not null	Name of the owner of the equipment
Storage_location	VAR-CHAR(100)	Not null	Location where the equipment is stored
Purchase_date	DATE	Not null	Date when the equipment was bought: 'YYYY-MM-DD'
Equipment_model_ID	INT	Foreign key, not null	Name of the model of this equipment. Link to the *Equipment_model* table

information included in that table is explained in table A.15. To each sampling point, there is a Site_ID which tells in which site belongs to this sampling point providing site details.

Every site which is studied, it is stored in the *Site* table including the information explained in table A.16.

Each watershed, where a sample can be taken, the relative information is stored in the *Watershed* table. The information provided for each watershed is described in table A.17.

Table A.8: Detail of the *Equipment_ model* table in the dat*EAU*base.

Table columns	**Data type**	**Characteristic**	**Description**
Equipment_model_ID	INT	Primary key, not null, auto increment	A unique ID is generated automatically by MySQL
Equipment_model_name	VARCHAR(100)	Not null	Name of the equipment model. For example: ammo::lyser
Method	VARCHAR(100)	Not null	Method behind the equipment
Functions	TEXT	Not null	Description of the functions of the equipment
Manufacturer	VARCHAR(100)	Not null	Name of the manufacturer
Manual_location	VARCHAR(100)	Not null	Location where the manual is stored

Table A.9: Detail of the *Procedures* table in the dat*EAU*base.

Table columns	**Data type**	**Characteristic**	**Description**
Procedure_ID	INT	Primary key, not null, auto increment	A unique ID is generated automatically by MySQL
Procedure_name	VARCHAR(100)	Not null	Title name of the procedure
Procedure_type	TINYTEXT	Not null	Type of the procedure. For example, SOP
Description	TEXT	Not null	Description of the procedure
Storage_location	VARCHAR(100)	Not null	Where is the procedure stored

Table A.10: Detail of the *Equipment_ model_ has_ Parameter* table in the dat*EAU*base.

Table columns	**Data type**	**Characteristic**	**Description**
Equipment_model_ID	INT	Primary key, foreign key, not null	Link to the *Equipment_ model* table
Parameter_ID	INT	Primary key, foreign key, not null	Link to the *Parameter* table

Table A.11: Detail of the *Equipment_ model_ has_ Procedures* table in the dat*EAU*base.

Table columns	**Data type**	**Characteristic**	**Description**
Procedure_ID	INT	Primary key, foreign key, not null	Link to the *Procedures* table
Equipment_model_ID	INT	Primary key, foreign key, not null	Link to the *Equipment_ model* table

Table A.12: Detail of the *Parameter_has_Procedures* table in the dat*EAU*base.

Table columns	**Data type**	**Characteristic**	**Description**
Parameter_ID	INT	Primary key, foreign key, not null	Link to the *Parameter* table
Procedure_ID	INT	Primary key, foreign key, not null	Link to the *Procedures* table

Table A.13: Detail of the *Purpose* table in the dat*EAU*base.

Table columns	**Data type**	**Characteristic**	**Description**
Purpose_ID	INT	Primary key, not null, auto increment	A unique ID is generated automatically by MySQL
Purpose_name	VARCHAR(100)	Not null	Purpose of the data collection. For example, "Measurement", "Lab_analysis", "Calibration" and "Cleaning"
Description	TEXT	Not null	Description of the purpose

Table A.14: Detail of the *Weathe_condition* table in the dat*EAU*base.

Table columns	**Data type**	**Characteristic**	**Description**
Condition_ID	INT	Primary key, not null, auto increment	A unique ID is generated automatically by MySQL
Condition	VARCHAR(100)	Not null	Type of weather condition
Description	TEXT	Not null	Description of the condition

Also, for each watershed, urban and hydrological characteristics are included into the dat*EAU*base. The information about urban characteristics is stored in the *Urban_characteristics* table which is explained in table A.18.

Table A.15: Detail of the *Sampling_point* table in the dat*EAU*base.

Table columns	**Data type**	**Characteristic**	**Description**
Sampling_point_ID	INT	Primary key, not null, auto increment	A unique ID is generated automatically by MySQL
Sampling_point	VARCHAR(100)	Not null	Where the sample was taken. For example: "Inlet", "Outlet" or "Upstream"
Sampling_location	VARCHAR(100)		Where the sample was taken. For example: "Biofiltration", "Sewer_01" or "Retention_Tank"
Site_ID	INT	Foreign key, not null	The site where the sampling point is located. Link to the *Site* table
Latitude_GPS	VARCHAR(100)	Not null	GPS coordinates. For example: 47°54'25.103"
Longitude_ID	VARCHAR(100)	Not null	GPS coordinates. For example: 73°47'00.024"
Description	TEXT	Not null	Description of the sampling point
Picture	BLOB		Picture of the sampling point

Table A.16: Detail of the *Site* table in the dat*EAU*base.

Table columns	**Data type**	**Characteristic**	**Description**
Site_ID	INT	Primary key, not null, auto increment	A unique ID is generated automatically by MySQL
Site_name	VARCHAR(100)	Not null	Name of the site
Site_type	TINYTEXT	Not null	For example: "WWTP", "River" or "Sewer_system"
Watershed_ID	INT	Foreign key	Name of the watershed in which the site is located. Link to the *Watershed* table
Description	TEXT	Not null	Description of the site
Picture	BLOB		Picture of the site
Street_number	VARHAR(100)		Address: number of the street
Street_name	VARCHAR(100)		Address: name of the street
City	TINYTEXT		Address: name of the city
Zip_code	VARCHAR(100)		Address: zip code
Province	TINYTEXT	Not null	Address: name of the province
Country	TINYTEXT	Not null	Address: name of the country

Table A.17: Detail of the *Watershed* table in the dat*EAU*base.

Table columns	**Data type**	**Characteristic**	**Description**
Watershed_ID	INT	Primary key, not null, auto increment	A unique ID is generated automatically by MySQL
Watershed_name	VARCHAR(100)	Not null	Name of the watershed
Description	TEXT	Not null	Description of the watershed
Surface_area	FLOAT	Not null	Surface area of the watershed [ha]
Concentration_time	INT(100)	Not null	Concentration time in minutes [min]
Impervious_surface	FLOAT	Not null	Percentage of the impervious surface of the watershed in percentage [%]

Table A.18: Detail of the *Urban_characteristics* table in the dat*EAU*base.

Table columns	**Data type**	**Characteristic**	**Description**
Watershed_ID	INT	Primary key and foreign key, not null	Linked to the *Watershed* table
Commercial	FLOAT	Not null	Percentage [%] of commercial areas. For example stores or bank areas
Green_spaces	FLOAT	Not null	Percentage [%] of green spaces
Industrial	FLOAT	Not null	Percentage [%] of industrial areas. For example factories
Institutional	FLOAT	Not null	Percentage [%] of institutional areas. For example schools, police stations or city hall
Residential	FLOAT	Not null	Percentage [%] of residential areas. For example houses or apartment buildings
Agricultural	FLOAT	Not null	Percentage [%] of agricultural land use. For example farm land
Recreational	FLOAT	Not null	Percentage [%] of recreational areas. For example parks or sport fields

For the hydrological characteristics, the information about them is stored in the *Hydrological_characteristics* table which is explained in table A.19.

Table A.19: Detail of the *Hydrological_characteristics* table in the dat*EAU*base.

Table columns	**Data type**	**Characteristic**	**Description**
Watershed_ID	INT	Primary key and foreign key, not null	Linked to the *Watershed* table
Urban_area	FLOAT	Not null	Percentage [%] of urban areas
Forest	FLOAT	Not null	Percentage [%] of forest areas
Wetland	FLOAT	Not null	Percentage [%] of wetlands
Cropland	FLOAT	Not null	Percentage [%] of croplands
Meadow	FLOAT	Not null	Percentage [%] of meadow areas
Grassland	FLOAT	Not null	Percentage [%] of grasslands

All people who have access to the dat*EAU*base, them information is stored in the *Contact* table. The relative included is explained in table A.20.

All projects that can benefit of the dat*EAU*base, their information is stored in the *Project* table. The information included on that table is presented in the table A.21.

The sampling points for a determined project are related in *Project_has_Sampling_point* table. This table only relates the IDs of the *Sampling_point* table and the *Project* table (See table A.22).

All people involved in a project is specified in the *Project_has_Contact* table. This is only a relational IDs table (See table A.23).

Finally, to identify which equipment is used in which project, the IDs of both table are related belong the *Project_has_Equipment* table (See table A.24).

Table A.20: Detail of the *Contact* table in the dat*EAU*base.

Table columns	**Data type**	**Characteristic**	**Description**
Contact_ID	INT	Primary key, not null, auto increment	A unique ID is generated automatically by MySQL
Last_name	VARCHAR(100)	Not null	Last name of the contact
First_name	VARCHAR (100)	Not null	First name of the contact
Company	TEXT	Not null	Company name
Status	TINYTEXT	Not null	Status of the person. For example: "Master student", "Postdoc" or "Intern"
Function	TEXT	Not null	More detailed description about the functions
Officer_number	VARCHAR(100)	Not null	Number of the office
Email	VARCHAR(100)	Not null	E-mail address
Phone	VARCHAR(100)	Not null	Phone number
Skype_name	VARCHAR(100)		Skype name
LinkedIn	VARCHAR(100)		LinkedIn account
Street_number	VARHAR(100)	Not null	Address: number of the street
Street_name	VARCHAR(100)	Not null	Address: name of the street
City	TINYTEXT	Not null	Address: name of the city
Zip_code	VARCHAR(45)	Not null	Address: zip code
Province	TINYTEXT	Not null	Address: name of the province
Country	TINYTEXT	Not null	Address: name of the country

Table A.21: Detail of the *Project* table in the dat*EAU*base.

Table columns	**Data type**	**Characteristic**	**Description**
Project_ID	INT	Primary key, not null, auto increment	A unique ID is generated automatically by MySQL
Project_name	VARCHAR(100)	Not null	Name of the project
Description	TEXT	Not null	Description of the project

Table A.22: Detail of the *Project_has_Sampling_point* table in the dat*EAU*base.

Table columns	**Data type**	**Characteristic**	**Description**
Project_ID	INT	Primary key, foreign key, not null	Link to the *Project* table
Sampling_point_ID	INT	Primary key, foreign key, not null	Link to the *Sampling_point* table

Table A.23: Detail of the *Project_has_Contact* table in the dat*EAU*base.

Table columns	**Data type**	**Characteristic**	**Description**
Project_ID	INT	Primary key, foreign key, not null	Link to the *Project* table
Contact_ID	INT	Primary key, foreign key, not null	Link to the *Contact* table

Table A.24: Detail of the *Project_has_Equipment* table in the dat*EAU*base.

Table columns	**Data type**	**Characteristic**	**Description**
Project_ID	INT	Primary key, foreign key, not null	Link to the *Project* table
Equipment_ID	INT	Primary key, foreign key, not null	Link to the *Equipment* table

Appendix B

Evaluation of the cleaning effect

In this chapter, how to develop a control chart to evaluate the cleaning effect is presented.

Following the same methodology presented in section 3.3.1, the steps pursued are:

1. Analyze if the data (x) is normally distributed.
2. Calculate the % difference between the value before cleaning ($Value_{before,i}$) and the value after the cleaning ($Value_{after,i}$).

$$\%d_i = \frac{Value_{before,i} - Value_{after,i}}{Value_{after,i}} \cdot 100 \tag{B.1}$$

3. The center line is defined in 0 since it is the expected that there is no difference between the values before and after the cleaning.

$$Center\ line = 0 \tag{B.2}$$

4. Select 20 values which the difference is lower than 10 %.
5. With them, calculate the standard deviation ($\sigma_{\overline{\%d}}$)

$$\sigma_{\overline{\%d}} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} \%d_i^2} \tag{B.3}$$

6. Calculate the upper and lower control limits (UCL, LCL) using the formulas:

$$UCL = +L\sigma_{\overline{\%d}} \tag{B.4}$$

$$LCL = -L\sigma_{\overline{\%d}} \tag{B.5}$$

   The L parameter is taken as 2, corresponding to a 95 % of probability that the value can be accepted.

7. Graph the control chart, by drawing the different lines and plotting all the % differences between before cleaning and after cleaning values (y-axis) versus the cleaning time (x-axis).

In this case, it is established that when a point is out of the boundaries, the difference between the values before and after the cleaning activity is significant. Thus, the corresponding sensor should be clean earlier.