Uncertainty-Aware Sensor Data Management and Early Warning for Monitoring Industrial Infrastructures

George Tzagkarakis, Institute of Computer Science, Foundation for Research and Technology – Hellas, Greece & EONOS Investment Technologies, Paris, France

Aleka Seliniotaki, Institute of Computer Science, Foundation for Research & Technology – Hellas, Greece & Institute of Computer Science, University of Crete, Crete, Greece

Vassilis Christophides, Department of Computer Science, University of Crete, Crete, Greece

Panagiotis Tsakalides, Institute of Computer Science, Foundation for Research and Technology – Hellas, Greece & Institute of Computer Science, University of Crete, Crete, Greece

ABSTRACT

In several industrial applications, monitoring large-scale infrastructures in order to provide notifications for abnormal behavior is of high significance. For this purpose, the deployment of large-scale sensor networks is the current trend. However, this results in handling vast amounts of low-level, and often unreliable, data, while an efficient and real-time data manipulation is a strong demand. In this paper, the authors propose an uncertainty-aware data management system capable of monitoring interrelations between large and heterogeneous sensor data streams in real-time. To this end, an efficient similarity function is employed instead of the typical correlation coefficient to monitor dynamic phenomena for timely alerting notifications, and to guarantee the validity of detected extreme events. Experimental evaluation with a set of real data recorded by distinct sensors in an industrial water desalination plant reveals a superior performance of our proposed approach in terms of achieving significantly reduced execution times, along with increased accuracy in detecting extreme events and highly correlated pairs of sensor data streams, when compared with state-of-the-art data stream processing techniques.

Keywords: *Data Management Systems, Fast Correlation Monitoring, Industrial Surveillance, Uncertainty-Aware Alerting, Wireless Sensor Networks*

DOI: 10.4018/IJMSTR.2014100101

INTRODUCTION

Recent advances in information and communications technology (ICT) have led to a significant progress in the design of devices incorporating wireless communication, processing and storage capabilities, as well as diverse sensing and actuation functionalities in a single unit that is compact, economical, autonomous and destined to become ubiquitous. This revolution appears in the form of dense and distributed large-scale self-organized wireless sensor networks (WSN) for carrying out various tasks that are of great societal interest, such as environmental monitoring and surveillance or monitoring and management in large-scale industrial infrastructures.

The HYDROBIONETS project[1] is a characteristic example of such an infrastructure for water resource management. Specifically, it targets at developing a real-time microbiological wireless networked control system for water desalination and treatment plants, providing the fundamental design principles of a wireless BioMEM network (WBN) with distributed multi-sensing and multi-actuation capabilities.

The HYDROBIONETS infrastructure focuses on monitoring the complete water cycle in large-scale water treatment and desalination plants via the deployment of a WBN. Distinct sensors of the WBN measure critical microbiological and electrochemical parameters in the water at different stages of the water treatment. The associated distributed, autonomous sensing is further exploited to produce *intelligent reasoning* over the data by supporting advanced operations, such as, querying, high-level analysis, and alerting.

At the core of the HYDROBIONETS system, which curries out those operations, is an efficient *data management* and processing module. This module comprises of distinct collaborating computational nodes, which monitor and control several physical entities and dynamic phenomena. The sensor data and metadata, which are produced in streams by the sensors, can be either processed in real time or stored for further exploitation. Those data can be raw (as produced by the sensors) or aggregated, which are produced based on calculations at the node level. To accommodate the requirements of our industrial paradigm we focus on the design and development of a set of tools to deal with high-level analysis of the collected data. These tools will work on the available data, use a set of policies that would govern the "normal" operation of the sensors and the data values they report and employ in a coherent manner an appropriate statistical analysis in order to: (i) account for the underlying *uncertainty* of the recorded data, (ii) detect *extreme events* (*e.g.,* presence of highly contaminant substances) and provide specific *alerts* depending on the level of severity of the event, and (iii) guarantee the validity of the detected extreme events by computing and observing pairs of distinct sensors which are *highly correlated*.

Concerning the first issue, we are motivated by the fact that typical WSN nodes deployed for monitoring industrial infrastructures do not handle any quality aspect of physical device data. Instead, they interface with a *high-level representation* and reconstruction of the sensed physical world. As a result, data processing has to additionally cope with the inherent data uncertainty, where stream data may be incomplete, imprecise or even misleading, thus hindering an accurate and reliable decision making.

Uncertainty-aware data management (Aggarwal, 2009) presents numerous challenges in terms of collecting, modeling, representing, querying, indexing and mining the data. Given that uncertainty has been recently recognized as an additional source of valuable information for data analysis which should be preserved, in contrast to existing data management systems, our approach incorporates an appropriate submodule to handle the inherent data uncertainty. More specifically, a *spreadsheet-based* approach is employed to identify, quantify, and combine the underlying uncertainty from the most dominant potential sources of uncertainty.

Another major functionality assigned to our *uncertainty-aware data management system* (UADM) is to perform high-level operations, and specifically to provide notifications of

extreme events by employing raw sensor data (Tran *et al.*, 2012). Two widely-used methods for notifying a system operator whether the data has unexpected values are: (i) *compliance with operating limits* (COL), and (ii) the method of *peaks over a threshold* (POT) (McNeil & Saladin, 1997). Since the detection of abnormal behavior is affected by the underlying uncertainty, the above two extreme event detectors are modified accordingly so as to account for the imprecise nature of the raw sensor data.

Rather than computing single stream statistics, such as average and standard deviation, in an industrial monitoring setting it is also of significance to identify sensor pairs characterized by high correlations. More specifically, a system operator may rely on pairwise sensor stream correlations to reveal interrelations between seemingly independent physical quantities monitored by distinct sensors. This can be further exploited to guarantee the validity of a detected extreme event and provide the necessary alerting notifications.

Whereas traditional statistical machine learning provides well-established mathematical tools for data analysis (Tran *et al.*, 2010; Tran *et al.*, 2012; Yeh *et al.*, 2009; Canton, 2002), their performance is limited when processing high-dimensional data streams. Specifically, existing techniques for monitoring pairwise stream correlations exhibit several drawbacks. In a recent work (Datar *et al.*, 2002), the problem of maintaining data stream statistics over sliding windows is studied, with the focus being only on single stream statistics. On the other hand, (Gehrke *et al.*, 2001) introduced an extension for monitoring the statistics of multiple data streams, but the computation of correlated aggregates is limited to a small number of monitored streams. In addition, StatStream (Zhu & Shasha, 2002) has been proven a successful data stream monitoring system, which enables the computation of single- and multiple-stream statistics. How-ever, the main drawback of this technique is the difficulty to define an appropriate "similarity" function for data streams describing dynamic phenomena with unknown prior distributions, which is normally the case in an industrial environment.

To overcome the limitations of previous approaches, our UADM system is equipped by a computationally efficient "*similarity extraction*" module, which enables the monitoring of pairwise correlations between high-dimensional and heterogeneous sensor data streams in a fast online fashion. To this end, instead of computing all pairwise correlations between the original full-dimensional data streams, we exploit the compressibility property of the discrete Fourier transform (DFT) to concentrate the inherent energy content of a given sensor stream in the first few high-amplitude coefficients, as in (Zhu & Shasha, 2002). Then, an appropriate similarity measure, which incorporates the estimated underlying uncertainty, is defined and applied on the associated pairs of truncated DFTs as a proxy of the corresponding correlation coefficients.

The performance of our proposed UADM system is evaluated on a set of real data provided by ACCIONA Agua, recorded by a set of distinct electromechanical sensors in La Tordera's desalination plant[2]. The results reveal a significant improvement of our approach, in terms of achieving highly reduced execution times in conjunction with accurate estimation of the highly-"similar" pairs of sensor streams, as well as a timely alerting performance, when compared with existing widely used data analysis techniques.

To summarize, the main contribution of this paper is threefold: (i) in contrast to common data management, which relies on the raw measurements, we verify that the underlying data uncertainty is a valuable source of information, which should be preserved, towards providing more ubiquitous data descriptions; (ii) the performance of two widely used extreme event detection methods is enhanced by incorporating the inherent data uncertainty component; and (iii) a fast and robust method is proposed for uncertainty-aware monitoring of pairwise interrelations ("similarities") between distinct sensors, which outperforms state-of-the-art pairwise correlation extraction methods. Our utmost goal is to provide a valuable insight

into the design and implementation principles of efficient and robust data management system integrating these the above three functionalities for industrial monitoring and surveillance applications, while emphasizing the importance of accounting for the underlying data uncertainty as an additional source of information, which should be preserved across all stages of the data processing chain.

PROBLEM DESCRIPTION

Industrial Monitoring Setting

Our developed data management system is at the core of the HYDROBIONETS WSN infrastructure, which enables multi-sensing and multi-actuation in water treatment and desalination plants. Nowadays, membrane cleaning in these plants, which is performed by chemical shock and backwashing, does not take into account any bacteria sensor information, thus it is usually performed more aggressively than needed. Moreover, the amount of energy spent in this process is more than necessary, since the operation of current systems does not exploit any precise characterization of the amount of *fouling* on the membranes. Fouling refers to the accumulation of unwanted material on solid surfaces, most often in an aquatic environment. The fouling material may consist of either living organisms (biofouling) or a non-living substance (inorganic or organic). However, in practice, when the fouling takes place includes all types of material, that is, organic, inorganic and bacterial fouling (biofouling).

In the adopted industrial application scenario, a desalination pilot plant is equipped with a number of various distinct electrochemical sensors, scattered at different locations for monitoring several physical and mechanical variables in the plant related to the formation of biofouling. The major task of HYDROBIONETS is to monitor and control the fouling phenomena developed during the different stages of the water treatment. Fouling phenomena are common and diverse including fouling of natural surfaces in the aquatic environment, fouling of heat-transfer components through ingredients contained in the cooling water, and most importantly fouling attached on the desalination membranes.

The growth of a fouling layer due to the deposition of undesirable materials on the membrane is a persistent problem in water treatment and desalination plants. Specific types of fouling, such as, the deposition of suspended solids, colloids, and microbiological cells, onto or into the membranes, is a severe issue which impedes the normal operation of the membranes. Complete removal of fouling mass by intensive pretreatment of the feed water is not always feasible. Motivated by this, developing techniques for monitoring the fouling formation and providing early warning notifications in case of high concentration is a necessity in order to achieve the long-term and stable operation of the filtration membranes, while reducing the energy consumption and maintenance expenses.

To fulfill these requirements, an appropriately deployed WSN acquires measurements from distinct physical variables recorded by various electrochemical sensors, such as, temperature, turbidity, conductivity, oxygen content, pH, redox potential, nitrate and chlorine. Based on the monitoring process further operations take place including: (i) the detection of high fouling potential, (ii) the optimization of chemical cleaning of the ultrafiltration membranes, and (iii) the control of chlorine dosage during the reverse osmosis phase.

To conform to the above requirements, our data management system comprises of collaborating computational elements, which observe and control distinct physical entities and dynamic phenomena. Furthermore, in our industrial environment, measurements from heterogeneous sensors, distributed over a geographic area, need to be processed efficiently in order to extract the spatio-temporal behavior of the monitored physical variables or to detect, identify and localize sources and events of interest.

Timely actuation is a crucial issue, while providing guarantees for the validity of a detected extreme event is also of high significance.

To this end, we need to monitor continuously and in an online fashion the interrelations between a number of distinct data streams produced by sensors at different stages of water treatment (*e.g.,* pre-filtering, pre-treatment and reverse osmosis), while accounting for their inherent imprecision expressed in terms of uncertainty. Although this uncertainty component may be due to hardware defections or environmental variations, its effects can be only observed and quantified from the recorded sensor measurements.

Figure 1 presents a generic structure of our proposed uncertainty-aware data management system, which consists of the three building blocks: (i) *uncertainty estimation*, (ii) *correlations extraction*, and (iii) *detection of extreme events*. Appropriate data services are provided to manipulate the sensor measurements, as well as to characterize the generated data quality. Computationally efficient extraction of correlations from uncertain data streams is then coupled with modified uncertainty-aware extreme event detectors to enable higher-level analysis, which form the basis for the development of an integrated UADM system for monitoring dynamic sensor networks and alerting in case of abnormal events.

Monitoring Stream Interrelations

Depending on the monitored phenomenon and the environmental conditions, the behavior of the recorded data streams may evolve significantly over time. Changes in data characteristics (*e.g.,* statistical distribution) may indicate anomalies in the "normal" behavior of the monitored streams, or alterations in the data acquisition or transmission process. Quantification of the degree of interrelation between pairs of seemingly different sensors, in conjunction with the detection of behavior variations, is crucial for a meaningful and reliable decision making in an industrial infrastructure as is our case.

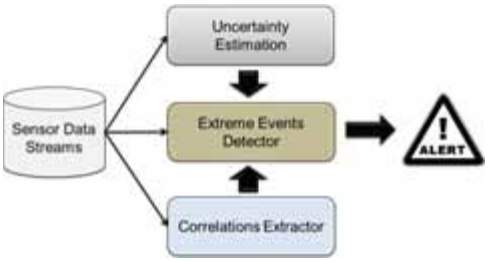
Since the detection of abnormal behavior is affected strongly by the underlying data uncertainty, its integration in the decision-making process is a prerequisite. To this end, the designed method for computing and monitoring interrelations between distinct sensors should have the following characteristics:

- **Computational Efficiency:** Fast and robust computation for detecting any abnormal system behavior in an almost time-continuous fashion.
- **Agility:** The memory space requirements must be linear on the stream length *n*.
- **Accuracy:** Given that the exact computations require increased space and time resources, we rely on suitable approximations with minimal approximation error.

To summarize, the problem we address in this paper is defined as follows:

Problem: "Given S co-evolving uncertain data streams of equal length n, detect at any time instant the occurrence of an extreme event, along with the pairs of streams whose correlation is above a predetermined threshold."

Figure 1. Building blocks of our proposed uncertainty-aware data management system



Data Processing Over Sliding Windows

Depending on their type, the available electrochemical sensors report a measurement value within a predefined period of time, usually at a scale of a few seconds or minutes. In our proposed data management system, processing of raw data streams is performed on a basis of *sliding windows*.

Doing so, we focus on the latest acquired data according to an appropriate ordering of tuples, by means of timestamp values attached to every transmitted data packet. Specifically, at any time instant, a *window* determines a finite set of recent tuples from a given sensor stream. This finite portion of the stream will be subsequently used to produce results corresponding to that time instant. As time evolves, new data tuples are included in the current window replacing the older tuples, which are stored in a database. From another point of view, a window is generally considered as a mechanism for adjusting flexible bounds on an unbounded stream in order to fetch a finite, yet ever-changing set of tuples, which may be regarded as a temporal relation.

In practice, data windows are mainly built in two different ways, namely, (i) *attribute-based* windows, and (ii) *count-based* windows. In the former case, an attribute is first assigned as the *windowing attribute* (*e.g.,* time), and consecutive tuples for which this attribute is within a certain interval constitute a window. Here, tuples are assumed to arrive in increasing order of their windowing attributes. In the later case, a certain number of consecutive tuples constitute a window (*e.g.* the last *K* readings from a sensor).A formal definition of the windowing operator is the following:

Definition (Windowing Operator): Let WE be a window with conjunctive condition E applied at time instant τ0 T over the elements of a data stream S, that is, over its current values $S(\tau_0)$. Then:

$$\forall \tau_\iota \in \mathrm{T}, \tau_\iota \geq \tau_0,\ W_E(S(\tau_\iota)) = \{s \in S(\tau_\iota) : \mathrm{E}(s, \mathrm{T}) \ holds\}$$

Figure 2 illustrates the mechanism of sliding windows over a data stream *S*, which consists of a sequence of continuously-arriving elements with distinct timestamps, that is, $S = \{s[1], s[2], ..., s[t], ...\}$, where $s[i]$ is an element with timestamp *i*, and *t* is the current timestamp. When a new element $s[t + 1]$ arrives at the next time instant $(t + 1)$, it is appended to *S*. Doing so, the oldest element $s[t - w + 1]$ expires and is stored in a database for future use. Thus, the data corresponding to the timestamp $(t + 1)$ is included in a new sliding window $\{s[t - w + 2], ..., s[t + 1]\}$ of size *w*.

MANAGING UNCERTAINTY IN SENSOR DATA STREAMS

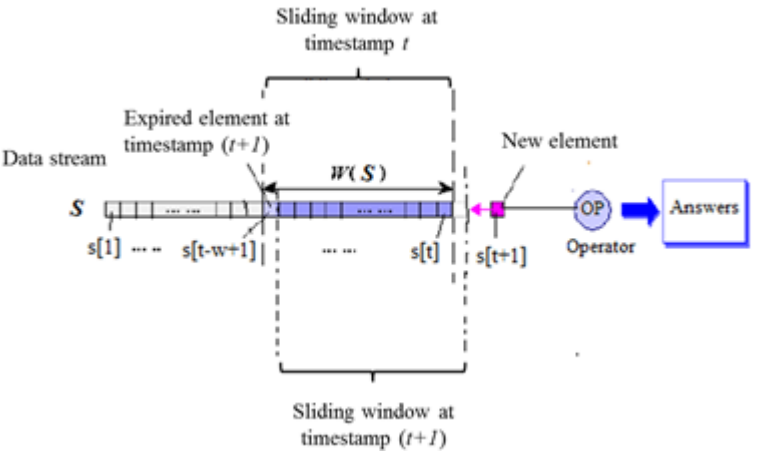
Having acquired the raw sensor data from distinct electrochemical sensors distributed across the plant, our data management system first estimates their corresponding inherent uncertainty. The estimation is carried out in two consecutive steps, namely, *identification* of all the potential sources of uncertainty, followed by their *quantification* and *propagation*. In the following, each one of these steps is described in detail.

Step 1: Identification of Uncertainty Sources

Identification of uncertainty sources comprises the first step towards the design of an integrated uncertainty-aware data management system. In practice, the underlying uncertainty may arise due to several distinct sources, such as, an incomplete definition of the observed quantities, sampling effects and interferences, varying environmental conditions, and inherent uncertainties of the equipment.

A very convenient way to determine the most dominant uncertainty sources, along with

Figure 2. Sliding window applied on a data stream



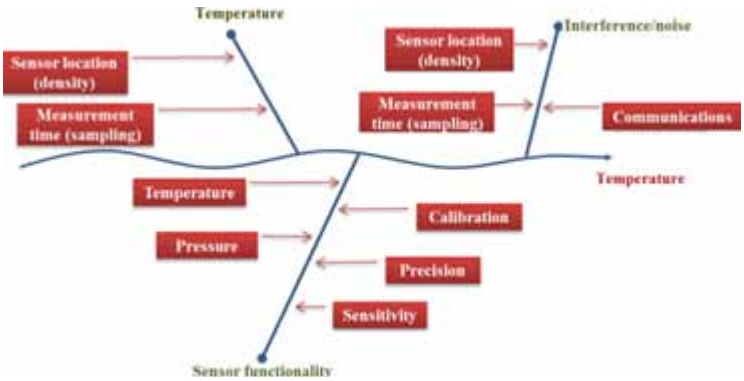
their potential interdependencies, is to exploit the so-called *cause and effect (or Ishikawa) diagram*. This diagram also ensures comprehensive coverage, while helping to avoid double counting of sources. Once the set of most significant uncertainty sources is formed, their effects can be usually represented in terms of a measurement model.

As a typical example, Figure 3 shows a cause and effect diagram for a temperature sensor. The first source of uncertainty is the sensor's functionality by itself. However, its performance is affected by several distinct factors, such as, its sensitivity and precision, the calibration, the operating temperature, and the water flow-rate and pressure. On the other hand, the accuracy of the recorded values depends also on the sensors' deployment density and location, as well as on the sampling process we use. Possible misplacement or a very sparse time-sampling is expected to increase the uncertainty, especially when the monitored variable varies rapidly across time.

Step 2: Quantification of Uncertainty

The identification of uncertainty sources is followed by a quantification process. This is done by estimating the uncertainty of each individual

Figure 3. Cause and effect diagram for a temperature sensor



source and then combining them appropriately to obtain a single overall uncertainty.

Towards assessing the underlying uncertainty component in a given raw data stream, we recall here its distinction into two separate categories, namely, *type A* (aleatoric, or statistical) and *type B* (epistemic, or systematic) uncertainty.

Uncertainties of type A are characterized by the estimated variances σ_i^2 (or the standard deviations σ_i), which are obtained by statistical analysis of the observations in the raw data streams. Following a *sliding window* approach, as it was described in a previous section, the variance σ_i^2 of the i-th sensor is estimated from its measurements in the current window. This is equivalent to obtaining a standard uncertainty from a probability density function (pdf) derived from an observed frequency (empirical) distribution. Let $\boldsymbol{y}$ be a data stream with N values $\{y_1, ..., y_N\}$, which corresponds to a specific observed variable. Then, the standard uncertainty of $\boldsymbol{y}$, which is denoted by $u(\boldsymbol{y})$, is expressed in terms of the corresponding standard deviation σ_y, estimated directly from the observations y_i, as follows,

$$u(\boldsymbol{y}) = \frac{\sigma_y}{\sqrt{N}} \tag{1}$$

On the other hand, for uncertainties of type B, the estimated "variance" s_j^2 is obtained from an assumed probability density function based on our prior knowledge for the corresponding source of uncertainty, which may include a) data from previous measurements, b) experience or knowledge of the properties of instrumentation and materials used, c) manufacturer's specifications, and d) calibration data. In general, concerning type B uncertainties, the quantification is performed either by means of an external information source, or from an assumed distribution. Typical assumptions for the prior distributions include the Gaussian (*e.g.,* when an estimate is made from repeated observations of a randomly varying process, or when the uncertainty is given as a standard deviation or a confidence interval), the uniform (*e.g.,* when a manufacturer's specification, or some other certificate, give limits without specifying a confidence level and without any further knowledge of the distribution's shape), and the triangular distribution (*e.g.,* when the measured values are more likely to be close to a value a than near the bounds of an interval with mean equal to a).

Having estimated the individual uncertainties, expressed as standard uncertainties, the next step is to combine them in the form of a *combined standard uncertainty*. Although in practice there may exist correlations between the individual uncertainty sources, however, it is usually impossible to compute those correlations accurately. For this purpose, it is more convenient to rely on an assumption of independence between the individual uncertainty sources.

In the following, let y denote the observed variable associated with the acquired data stream $\boldsymbol{y}$. Furthermore, let $y = f\{x_1, ..., x_L\}$ be an observed variable, which depends on L input variables x_l through a functional relation $f(\cdot)$. Then, the *combined standard uncertainty* of y, for independent input variables x_l, $l = 1, ..., L$, is given by:

$$u_c(y) = \sqrt{\sum_{l=1}^{L} \left(\frac{\partial f}{\partial x_l} \right)^2 u^2(x_l)} \tag{2}$$

where $u(x_l)$ denotes the standard uncertainty of the input variable x_l (either of type A, or of type B), while the partial derivatives $\partial f / \partial x_l$, the so-called *sensitivity coefficients*, quantify how much the output y varies with changes in the values of the input variables x_l, $l=1, ... L$. Finally, the combined standard uncertainty, which may be thought of as equivalent to one standard deviation, is transformed into an *overall expanded uncertainty*, U, via multiplication with a coverage factor k, that is,

Table 1. Coverage factor as a function of confidence level for the Gaussian distribution

Coverage Factor (k)	Confidence Level (%)
$k = 1$	67%
$k = 1.96$	95%
$k = 2.576$	99%
$k = 3$	99.7%

$$U(y) = k \cdot u_c(y) \quad (3)$$

where the value of k is determined in terms of the desired confidence level as shown in Table1.

The most convenient way to summarize all this information and compute the overall uncertainty is by means of *spreadsheet tables*. A spreadsheet table lists the dominant sources of uncertainty and categorizes them according to their type. Based on that, the individual standard uncertainties are stated explicitly, along with the overall combined uncertainty. An example of such a table for a temperature sensor is shown in Table 2.

The final output of the above spreadsheet-based approach is the assignment of the combined and expanded uncertainty values to the current windows of all the sensors. This completes the first building block of our uncertainty-aware data management system. In the following section we describe the sec-

Table 2. Example of a spreadsheet table for a temperature sensor

Source of Uncertainty			Value (±)	Probability Distribution	Divisor	Standard Uncertainty $u(x)$
Type B	Sensor	Calibration	C_1	Normal	2	$C_1/2$
		Precision (Resolution)	C_2	Rectangular	$\sqrt{3}$	$C_2/\sqrt{3}\ \sqrt{3}$
		Sensitivity	C_3	Rectangular	$\sqrt{3}$	$C_3/\sqrt{3}\ \sqrt{3}$
	Sensor density		C_4	Rectangular	$\sqrt{3}$	$C_4/\sqrt{3}\ \sqrt{3}$
	Sampling		C_5	Rectangular	$\sqrt{3}$	$C_5/\sqrt{3}\ \sqrt{3}$
Type A	Temperature		C_T	-		σ_T
	Pressure		C_P	-		σ_P
Combined standard uncertainty $u_{c,b}(y)$						
Coverage factor k_b						
Expanded uncertainty U_b						

ond building block, namely, the detection of extreme events by employing two modified extreme event detectors in order to account for the inherent data uncertainty.

UNCERTAINTY-AWARE DETECTION OF EXTREME EVENTS

Concerning the design of mechanisms notifying for extreme events, the estimated uncertainty, in conjunction with appropriate assumptions for the prior probabilistic models, can be exploited in a statistical framework for the detection of extreme values. Extreme value theory allows, under specific conditions, to predict *rare events*, which diverge from a "normal" pattern because of their rareness. For instance, in the HYDROBIONETS framework, a typical extreme event is the detection of high chlorine concentration in the water, or a high concentration of biofilms on the desalination membranes. As mentioned before, early warning for abnormal behavior is crucial when working in large-scale industrial environments.

In our developed UADM system, the identification of critical events is performed by means of two robust and computationally efficient methods. More specifically, we enhance the performance of two widely used techniques for extreme events detection by incorporating the underlying estimated data uncertainty. The first one, namely, the *compliance with operating limits*, performs simple comparisons of predetermined user-specified operating limits with the recorded measurements augmented by their estimated uncertainty. This modification maintains the computational efficiency of the original version, while improving its adaptivity to imprecise measurements. In a similar way, the second approach, the so-called *Peaks-Over-Threshold* (POT) method, is modified accordingly so as to identify the time instants when the measurements (also augmented by the estimated combined or expanded uncertainty) exceed an estimated threshold.

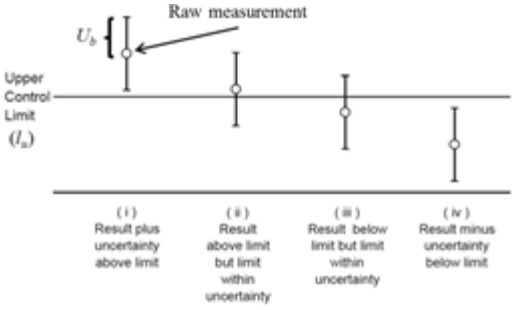
Compliance with Operating Limits

The simplest way to exploit the estimated combined or expanded uncertainty to design an alerting mechanism is as shown in Figure 4. More specifically, let l_u denote an upper operating limit dictated by a manufacturer or a specification standard. Although, for convenience, we restrict ourselves in the case of an upper limit, however, the same remarks are straightforward when compliance with a lower operating limit l_m is required.

As shown in Figure 4, there are four possible cases for a measurement and its associated expanded uncertainty interval, $y \pm U_b$ when compared with an upper limit l_u, namely,

1. Both the measurement and the expanded uncertainty interval are above the upper limit l_u,

Figure 4. Compliance with a predetermined upper operating limit



2. The measurement is larger than l_u and the expanded uncertainty interval contains l_u,
3. The measurement is lower than l_u and the expanded uncertainty interval contains l_u, and
4. Both the measurement and the expanded uncertainty interval are below l_u.

Case (i) clearly triggers an alerting notification for the occurrence of an extreme event, while (iv) is the only one which is in compliance with the specifications. On the other hand, in cases (ii) and (iii) we could not infer with absolute certainty whether an alert should appear or not. However, in a socially "sensitive" application, such as the water treatment, a system operator should classify cases (ii) and (iii) as possible divergences from normal operation, and thus draw more attention on the associated monitored variables. Notice also that, in contrast to the original version of this method, which supports only two cases (above or below l_u), the modified one exploits two additional ones due to the presence of uncertainty.

Despite its simplicity, the main drawback of this method is that it can be very sensitive to an under- or over-estimate of the expanded uncertainty, as well as of the measurement value, increasing the probability of false alerts. However, with appropriate setup of the hardware (sensors) and continuous monitoring of the environmental conditions, we could increase our trust to this method.

Peaks-Over-Threshold

Similarly to the previous method, we also extend the original POT method in order to account for the underlying data uncertainty. More specifically, we consider $\tilde{\boldsymbol{y}} = \{\tilde{y}_1, \ldots, \tilde{y}_N\}$ to be a data stream with N measurement values, which are spread out by the corresponding estimated expanded uncertainty, that is, $\tilde{y}_i = y_i + U_i$ or $\tilde{y}_i = y_i - U_i$, where we assume for the cumulative distribution function (CDF) F that, for $z > 0$, $F(z) = \Pr(\tilde{y} \le z) < 1$. Given a user-defined threshold ρ we study the statistical properties of the exceedances of $\tilde{y}_i$ over the threshold level ρ by fitting them with an appropriate distribution. In the following, we mainly rely on a threshold-dependent complementary CDF (or, equivalently, exceedance probability), which is given by

$$\bar{F}_\rho(z) \equiv \Pr(\tilde{y} - \rho > z \mid \tilde{y} > \rho) = \frac{\bar{F}(\rho + z)}{\bar{F}(\rho)} \tag{3}$$

where $\bar{F}(z) = 1 - F(z)$, for $z \ge 0$, denotes the tail of F. From (4) we also obtain

$$\bar{F}(\rho + z) = \bar{F}(\rho)\bar{F}_\rho(z), \quad z \ge 0 \tag{4}$$

The above identities can now be used to estimate tails and quantiles, to be used as refinements of the threshold ρ, adapted to the measurements statistics, within a predetermined level of confidence. To this end, let N_ρ be the subset of indices $j \in \{1,\ldots,N\}$ for which $\tilde{y}_j > \rho$, that is, $N_u = \{j \in \{1,\ldots,N\} : \tilde{y}_j > \rho\}$.

Then we denote by $E_1, \ldots, E_{N_\rho}$ the excesses of, $\tilde{y}_1$, $\tilde{y}_2$, ..., $\tilde{y}_N$, that is, the heights of the exceedances over ρ, as shown in Figure 5. $\bar{F}(\rho)$ is estimated simply as the relative frequency,

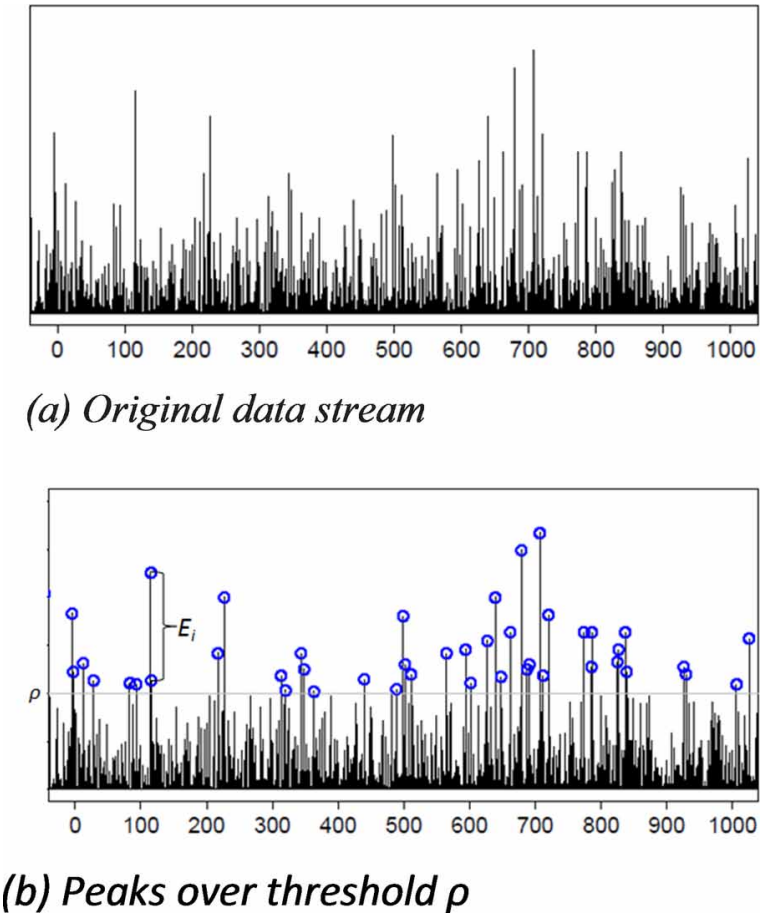
$$\bar{F}(\rho) \approx \frac{N_\rho}{N} \tag{5}$$

while $\bar{F}_\rho(z)$ is approximated by the *generalized Pareto (GP)* distribution as follows,

$$\bar{F}_\rho(z) \approx \left(1 + \gamma \frac{z}{\sigma(\rho)}\right)^{-\frac{1}{\gamma}}, \quad z \ge 0 \tag{6}$$

where the parameters γ, $\sigma(\rho)$ are obtained via maximum likelihood (ML) estimation from

Figure 5. Original data stream and its peaks over threshold ρ



the acquired sensor measurements directly. By combining (5)-(7) we obtain the overall tail estimator as follows,

$$\bar{F}(\rho+z) \approx \frac{N_\rho}{N}\left(1+\gamma\frac{z}{\sigma(\rho)}\right)^{-\frac{1}{\gamma}} \quad , \quad z \geq 0 \tag{7}$$

Finally, for a given $p \in (0,1)$ we obtain an estimator for the p-th quantile, z_p, as follows,

$$\hat{z}_p = \rho + \frac{\sigma(\rho)}{\gamma}\left(\left(\frac{N}{N_\rho}(1-p)\right)^{-\gamma} - 1\right) \tag{8}$$

This quantile can be further employed as a refinement of the initial threshold ρ in subsequent time windows of the data streams. Another benefit of using a probabilistic framework, as is the case of POT, instead of the simple compliance with operating limits, is that we can also estimate the average time interval between successive extreme events of similar intensity. This elapsed time is called *return period*, and is defined as the inverse of the exceedance probability as follows

$$T_R = \frac{1}{\bar{F}_\rho(z)} \tag{9}$$

We notice here that this does not mean that if an extreme event with a return period T_R occurs, then the next will occur in about T_R time units (*e.g.,* days, months, years). Instead, it means that, in any given time unit, there is a $1/T_R$ chance that it will happen, regardless of when the last similar event was.

Another main advantage of the POT method is that it allows us to work with larger sample populations, which ensures better fits to a distribution function. However, this comes at the cost of assuming that the data are considered to be identically distributed, which may not be always the case in practice. For the sensor data recorded in ACCIONA's plant our preliminary analysis showed that, for each individual sensor, the assumption of identically distributed data is a good approximation, thus allowing the use of the uncertainty-aware extension of the POT method.

EFFICIENT MONITORING OF DATA STREAM CORRELATIONS

Distinguishing efficiently between occasional and extreme events constitutes a major issue in the design of data management systems. It is of great importance to ensure in real or almost real time, especially when we deal with massive data sets, that a true extreme event occurs and not some coincidence or system/network failure. On the other hand, the degree of correlation between two or more sensor data streams characterizes their interrelations and dependencies. For this, the identification of highly correlated streams can be exploited as a further guarantee to verify the existence of a detected extreme event.

For instance, consider the case of two data streams recorded by a pressure and a temperature sensor, respectively. When the two sensors are placed nearby, we expect that a high pressure is associated with an increased temperature, which means that the correlation of these two streams should be relatively high. Thus, we assume that a potential notification for an extreme temperature should be related with a high measured pressure. If this is not the case, this information can be further exploited by a system operator to focus more on that part of the industrial infrastructure and perform a more thorough examination. The only ambiguous point here is related to the determination of "high correlation". The degree of "high correlation" is related to the specific application and the end-user, who has the flexibility to define how much strict this degree will be.

Doing so, the set of available sensors is divided into subsets of highly correlated sensors. This clustering enables a more convenient and meaningful monitoring of the overall infrastructure by a system operator, who focuses only on a subset of sensors, where an abnormal behavior has been detected for at least one of its members.

The typical approach for extracting pairwise sensor stream correlations is by means of the Pearson's correlation coefficient. For two given streams $\boldsymbol{x}$, $\boldsymbol{y}$ of equal length N, let $\boldsymbol{x}_w = (x_1,...,x_w)$, $\boldsymbol{y}_w = (y_1,...,y_w)$ be two time-synchronized windows of size w. Then, the correlation coefficient is given by

$$corr(\boldsymbol{x}_w, \boldsymbol{y}_w) = \frac{\sum_{i=1}^{w} x_i y_i - w\overline{x}\overline{y}}{(w-1)\sigma_x \sigma_y} \tag{10}$$

where $\overline{x}$ and $\overline{y}$ are the means of $\boldsymbol{x}_w$ and $\boldsymbol{y}_w$, respectively, and σ_x, σ_y denote their corresponding standard deviations. However, for each newly acquired measurement value, the correlation coefficient has to be recalculated, which yields an increased computational complexity, especially for high-dimensional data streams or for a large number of sensors. In particular, the major cost comes from the summation of inner products of the form (ref. (11))

$$\left\langle \boldsymbol{x}_w, \boldsymbol{y}_w \right\rangle = \sum_{i=1}^{w} x_i y_i$$

Motivated by this limitation, in our proposed UADM system we implement a compu-

tationally efficient method for nearly real-time extraction of highly correlated data streams by combining discrete Fourier transforms (DFT) over sliding windows with a proper *stream similarity measure*. In order to account for the underlying uncertainty or other data ambiguities, we restrict ourselves on the detection of pairs of streams whose correlation is above a specific threshold.

DFT-Based Correlation Computation

Let $\boldsymbol{x}_w$ and $\boldsymbol{y}_w$ denote two time windows of length w corresponding to the same time-interval. Working in a DFT framework, each sample x_i (similarly, y_i) can be expressed in terms of a linear combination of exponential functions

$$x_k \approx \frac{1}{\sqrt{w}} \sum_{f=0}^{N-1} X_f e^{i2\pi fk/w}, \quad k = 1,\ldots,w \tag{11}$$

where X_f is the set of N DFT coefficients, with $N < w$. In this way, the computational cost for computing the inner product between the two time windows (and subsequently the correlation coefficient) is reduced from w to N. The fast and efficient computation of the DFT guarantees that it can be used to compute inner products and, thus, correlations over sliding windows of any size.

The above DFT-based approach enables the fast monitoring of synchronized streams over a given time window, whose correlation exceeds a predefined threshold. This is dictated by the following lemma, which gives a correspondence between the correlation coefficient and the Euclidean distance between two data streams.

Lemma 1 (Rafie & Mendelzon, 1997***):*** The correlation coefficient of two data streams $\boldsymbol{x}$, and $\boldsymbol{y}$, of length w is expressed in terms of a Euclidean distance as follows

$$corr(\boldsymbol{x}, \boldsymbol{y}) = 1 - \frac{1}{2w} d^2(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}})$$

where $d(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}})$ is the Euclidean distance between $\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}$, that is, the original data streams normalized to mean zero and variance equal to one.

By reducing the correlation coefficient to Euclidean distance, we can apply the techniques described in (Zhu & Shasha, 2002) to report data streams with correlation coefficients higher than a specific threshold:

Lemma 2 (Zhu & Shasha, 2002***):*** Let the DFTs of the normalized data streams $\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}$ be $\hat{\boldsymbol{X}}$ and $\hat{\boldsymbol{Y}}$, respectively. Then,

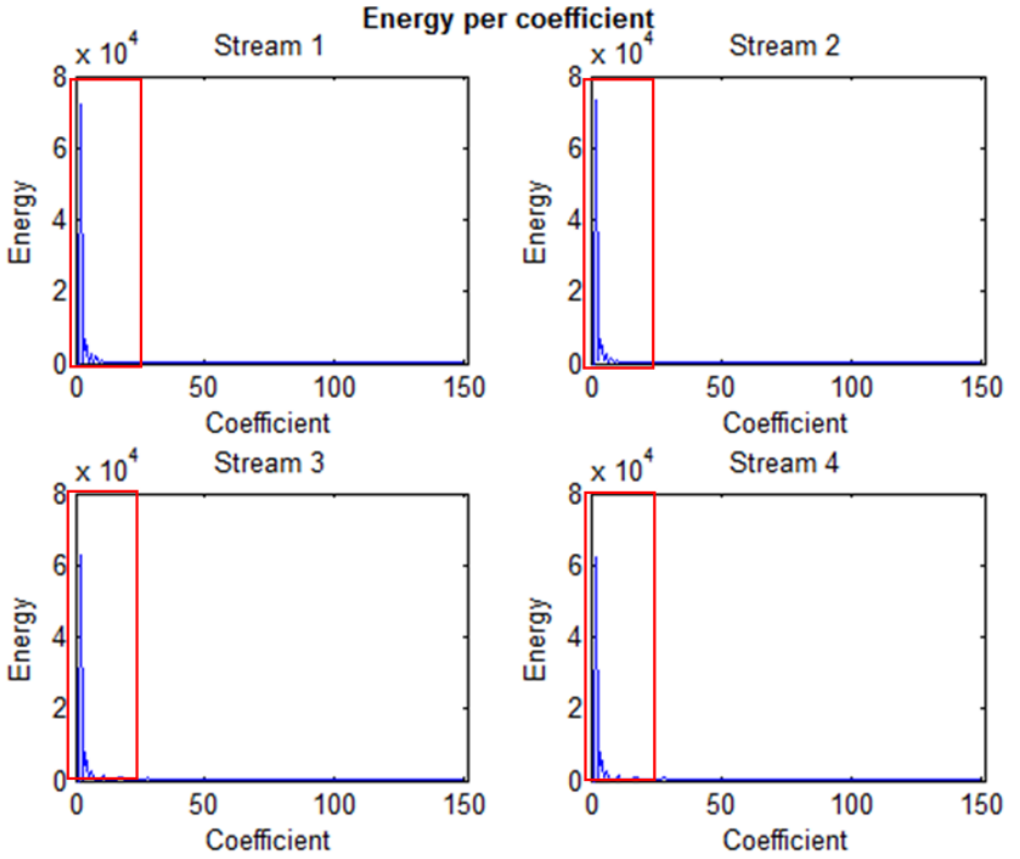
$$corr(\boldsymbol{x}, \boldsymbol{y}) \geq \varepsilon \Rightarrow d_M(\hat{\boldsymbol{X}}, \hat{\boldsymbol{Y}}) \leq \sqrt{2w(1-\varepsilon)}$$

where ε is a given threshold, and $d_M(\hat{\boldsymbol{X}}, \hat{\boldsymbol{Y}})$ is the Euclidean distance between the corresponding truncated DFTs, which are derived by keeping the first $M \leq w/2$ DFT coefficients with the largest amplitudes.

Lemma 2 implies that pairs of windowed sensor streams for which $d_M(\hat{\boldsymbol{X}}, \hat{\boldsymbol{Y}}) > \sqrt{2w(1-\varepsilon)}$ cannot have correlation coefficients above threshold ε. By ignoring those pairs, we get a set of *likely correlated* stream pairs. Notice that this is a superset of the correlated stream pairs, but there will be no false negatives.

The validity of Lemma 2 is based on the compactness of the DFT representations, that is, the concentration of the main portion of the energy for a given stream in the first few significant (high-amplitude) DFT coefficients. Figure 6 illustrates this property for four data streams recorded in ACCIONA's plant, from which it is apparent that the main energy content of the streams is concentrated in the first few low-frequency DFT coefficients.

Figure 6. Amplitudes of DFT coefficients for four real data streams acquired in ACCIONA's plant



Fast Pairwise Similarity Computation

In our data management system, we search for sensor pairs whose correlation is above a predefined threshold ε_{th}, in a fixed-sized sliding window. More specifically, let $\boldsymbol{s}$ be the reference stream, and $(\boldsymbol{y}_1,\boldsymbol{y}_2,..,\boldsymbol{y}_C)$ be the set of streams with which we compute the pairwise correlations in the current time interval. For a predetermined correlation threshold ε_{th}, the output of the process will be a subset of streams $\boldsymbol{y}_c$, for which the correlation with $\boldsymbol{s}$ is above ε_{th}.

In our proposed system, the problem of extracting highly correlated pairs of sensors is translated into a problem of identifying highly "*similar*" sensors, where the "similarity" is measured by an appropriately designed function. As in the previous DFT-based approach, the first step for each data stream values in the current window of length w, x_1, x_2,... x_w, is to normalize to mean zero and variance one, that is,

$$\hat{x}_i = \frac{x_i - \overline{x}}{\sigma_x}$$

where

$$\sigma_x = \sqrt{\sum_{i=1}^{w} (x_i - \overline{x})^2}$$

As a second step, the corresponding DFT of the normalized windowed data is computed. The data compression capability of the DFT is exploited to reduce the computational cost by approximating the original data by a highly reduced set of coefficients.

The final step towards our fast and robust extraction of highly "similar" sensor data streams is to identify those pairs $(\boldsymbol{s}, \boldsymbol{y}_c)$ with similarity above a given threshold ε_{th}. In order to avoid computing the similarity between all pairs of streams $(\boldsymbol{s}, \boldsymbol{y}_c)$, we reduce the set of candidate streams only to those streams that will be highly similar with $\boldsymbol{s}$ with high probability.

For this purpose, we introduce *peak similarity*, p_{sim}, as an appropriate similarity measure. More specifically, the similarity between two windowed data streams $\boldsymbol{s}$, $\boldsymbol{y}$ is computed by employing a truncated set of the first M high-amplitude DFT coefficients, where $M << w/2$, and the peak similarity measure is defined as follows:

$$p_{sim}(\boldsymbol{s}, \boldsymbol{y}) = \frac{1}{M}\sum_{i=1}^{M}\left[1 - \frac{\left|S_i - Y_i\right|}{2 \cdot \max\left(\left|S_i\right|, \left|Y_i\right|\right)}\right] \quad (12)$$

In order to account for the potential loss of information caused by the truncation of the set of DFT coefficients, the peak similarity measure does not employ the same threshold ε_{th} for finding the similar streams. Instead, we determine a new threshold $\varepsilon_{th,new}$, with our proposed method reporting as "highly-correlated" pairs those streams $\boldsymbol{s}$ and $\boldsymbol{y}_c$ for which $p_{sim}(\boldsymbol{s},\boldsymbol{y}_c) > \varepsilon_{th,new}$. However, special attention should be given on the selection of the threshold value $\varepsilon_{th,new}$. From our experimental evaluation, employing data from a set of various distinct sensors, we observed that if we choose an "elastic" enough threshold $\varepsilon_{th,new}$, then the subset of streams $\boldsymbol{y}_c$ with the highest peak similarity with $\boldsymbol{s}$ will also contain the highly correlated streams with $\boldsymbol{s}$ (that is, those with correlation coefficient above ε_{th}). In our implementation we set $\varepsilon_{th,new} = \varepsilon_{th} + e$, where e is a small positive number (in our experimental evaluation described below we set $e < 0.05$).

Uncertainty-Aware Fast Pairwise Similarity Computation

Towards the design of an integrated uncertainty-aware data management system, we extend the above peak similarity measure in order to monitor similarities between uncertain data streams. For this, (13) is not applied directly on the raw data streams, but on the original recordings by also accounting for their estimated expanded uncertainty. We note that this also affects the choice of the thresholds used to decide whether two streams are highly similar or not. Specifically, the threshold $\varepsilon_{th,new} = \varepsilon_{th} + e$ is set based on the streams $\boldsymbol{s}_1 \pm U_1$ and $\boldsymbol{s}_2 \pm U_2$, where U_1 and U_2 are the corresponding estimated expanded uncertainties of the two streams.

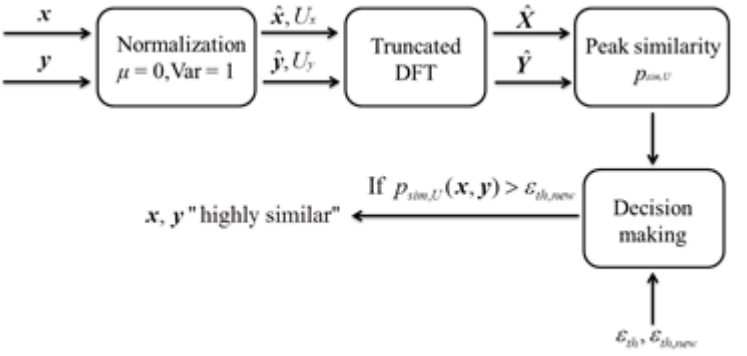
From the above, we derive an uncertainty-aware extension of p_{sim} which is given by

$$p_{sim,U}(\boldsymbol{s}, \boldsymbol{y}) = \frac{1}{M}\sum_{i=1}^{M}\left[1 - \frac{\left|\tilde{S}_i - \tilde{Y}_i\right|}{2 \cdot \max\left(\left|\tilde{S}_i\right|, \left|\tilde{Y}_i\right|\right)}\right] \quad (13)$$

where $\tilde{S}$, $\tilde{Y}$ are the truncated DFTs of the uncertain streams $\tilde{s} = s + U_s$ (or $\tilde{s} = s - U_s$) and $\tilde{y} = y + U_y$ (or $\tilde{y} = y - U_y$), respectively, with U_s and U_y denoting the uncertainties estimated in the current window of $\boldsymbol{s}$ and $\boldsymbol{y}$, respectively. The steps implementing our proposed fast and robust uncertainty-aware similarity measure are shown in Figure 7.

Having implemented all the necessary building blocks of our UADM system, as shown in Figure 1, in the following section we evaluate its performance on a set of real data streams recorded by distinct electrochemical sensors in ACCIONA's pilot plant.

Figure 7. Flow diagram for fast computation of uncertainty-aware pairwise sensor stream similarity



EXPERIMENTAL EVALUATION

The performance of our proposed UADM system, in terms of managing the underlying data uncertainty and providing early warnings, is evaluated on a real dataset provided by ACCIONA Agua. In particular, the dataset consists of 22 sensors of several types (pressure, temperature, conductivity, turbidity, pH, flow, and redox), while the corresponding measurements cover a period of 1 month at a sampling rate of one measurement every three minutes. Full sensor specifications (such as, sensor precision, sensitivity, and resolution), along with the corresponding measurements were provided for each individual sensor.

The inherent overall uncertainty of the recorded sensor data is quantified over sliding windows. If not stated explicitly otherwise, in the subsequent results, the window size is set equal to 80 samples, which corresponds to a time interval of approximately 4 hours, while the step size is fixed at 1 sample corresponding to a time-step of about 3 minutes. The expanded uncertainty is computed by fixing the coverage factor at $k = 1.96$, which is equivalent to a 95% confidence level.

Uncertainty Estimation

First, we illustrate the performance of the spreadsheet-based approach for estimating the underlying uncertainty in several distinct sensor streams. Figure 8 shows the overall estimated uncertainty for three randomly chosen sensors in our dataset, namely, conductivity, pH, and temperature sensors. The figure also reveals an additional potential use of the estimated uncertainty as an alerter of abnormal behavior. Indeed, the time instants where the uncertainty presents a peak coincides with the time instant where the corresponding sensor measurements deviate a lot from the previously recorded values. This observation concerning the use of expanded uncertainties as alerting indicators is left for a separate thorough study.

We remark two points concerning the estimation of the combined, and subsequently the expanded, uncertainty. The first is that, as it can be seen from Table 2, in case of a sensor with advanced specifications (*e.g.,* high precision and resolution, increased sensitivity, accurate calibration) the most dominant source of uncertainty is the standard deviation of the measurements by themselves. The second is that, in practice, it is often very difficult to estimate the values of the sensitivity coefficients in (2), which is also the case in our application. For this purpose, we relied on an assumption of independent sources of uncertainty, while setting all the individual sensitivity coefficients equal to 1, since we lack any prior knowledge about the relative degree of dominance between the uncertainty sources.

Pairwise Stream Similarity Monitoring

Concerning the performance of our proposed fast pairwise stream similarity monitoring approach, as a first step we examine its relation with the typical correlation coefficient. To do so, Figure 9 shows the values of the correlation coefficient (ref. (11)) and the uncertainty-aware peak similarity measure (ref. (14)) between a reference stream of pressure measurements and several distinct types of sensor streams. The labels Prx, Tx, FFx, FLx, PHx, Cx, BFx, TRx denote pressure, temperature, feed flow, filtrate flow, pH, conductivity, backwash flow and turbidity, respectively.

As it can be seen in Figure 9(a), the values of $p_{sim,U}$ are greater than or equal to the correlation coefficient values for the same pair of streams. For both the $p_{sim,U}$ and *corr*, the plotted values are the corresponding averages over all the sliding windows of each stream. This observation also motivated the heuristic rule for setting the new threshold $\varepsilon_{th,new} = \varepsilon_{th} + e$, as mentioned in the previous section. Furthermore, in Figure 9(b) we can see that for both the $p_{sim,U}$ and the correlation coefficient, the average values, when accounting for the underlying data uncertainty (with $k = 1.96$), are very close to each other as well. This verifies the validity of our proposed approach to extract highly correlated sensor streams.

As mentioned before, the computational complexity, and subsequently the execution time, is an important factor, which affects the overall performance of our UADM system. To this end, we compare the performance of our proposed approach, in terms of execution times for increasing stream lengths, against the typical correlation coefficient and two other state-of-the-art methods, namely, BRAID (Sakurai *et al.*, 2005) and StatStream (Zhu & Shasha, 2002). BRAID can handle data streams of semi-finite length, incrementally, quickly, and can estimate lag correlations with little error. On the other hand, as mentioned before, StatStream resembles more our approach, by finding high correlations among sensor pairs based on DFTs and a three-level time interval hierarchy.

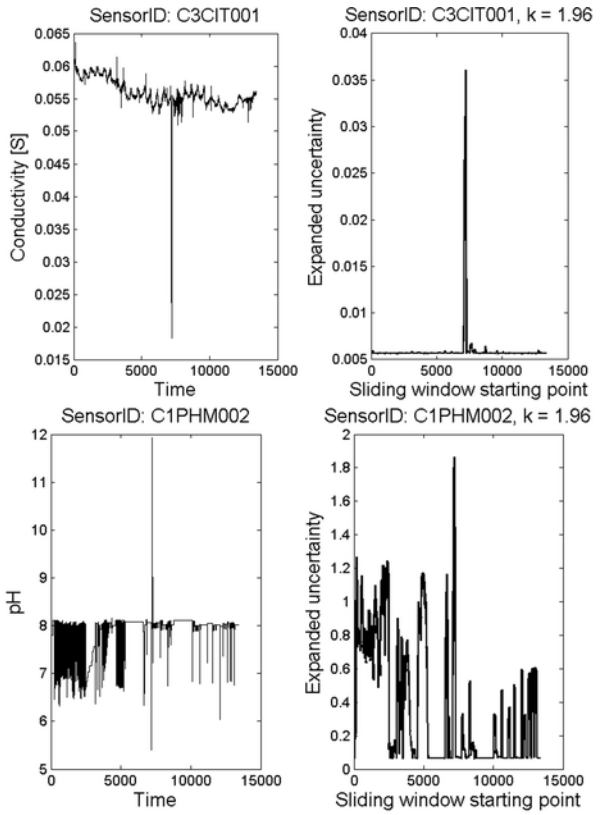
Figure 10 compares the execution times of our proposed method with the above three alternatives, as a function of the stream length. The results reveal a significant improvement in execution time achieved by our method, which is more prominent for higher stream lengths. Most importantly, we observe that the execution time of our method remains almost constant over the whole selected range of stream lengths, in contrast to the naïve and BRAID methods, whose execution times increase rapidly as the stream length increases.

For the BRAID algorithm we set the correlation lag to be equal to zero. This algorithm is characterized by gradual increase for increasing stream length, since it employs all the values of the recorded stream. For the StatStream algorithm, a simple hash function is used based on the mean value of each stream. Keeping the integer part of the mean values, the streams are mapped to appropriate cells in a grid structure. Doing so, only the correlations between neighboring cells are computed. The increased execution time of StatStream, compared to our approach, is due to the hash function, which involves more computations for the stream mapping. We expect though that the performance of StatStream could be enhanced, by designing a more efficient hash function.

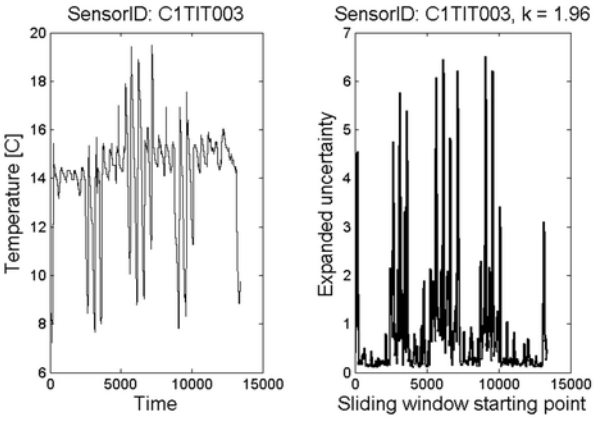
Detection of Extreme Events

In this section, we evaluate the performance of the uncertainty-aware extreme event detectors, which complete the design of our UADM system. As a first illustration, we evaluate the performance of the modified method which checks the compliance of the uncertainty-augmented raw data with given operating limits. Figure 11 compares the event detection capability of the original method of compliance with operating limits against its modified version for a pH sensor whose measurements are shown in Figure 8. The window size is set to 80 measurements, with a step size equal to 1, while the upper operating limit is set to 8.5.

Figure 8. Original measurements and estimated expanded uncertainties for three electrochemical sensors: a) conductivity, b) pH, c) temperature (window size = 80, step size = 1, k = 1.96)

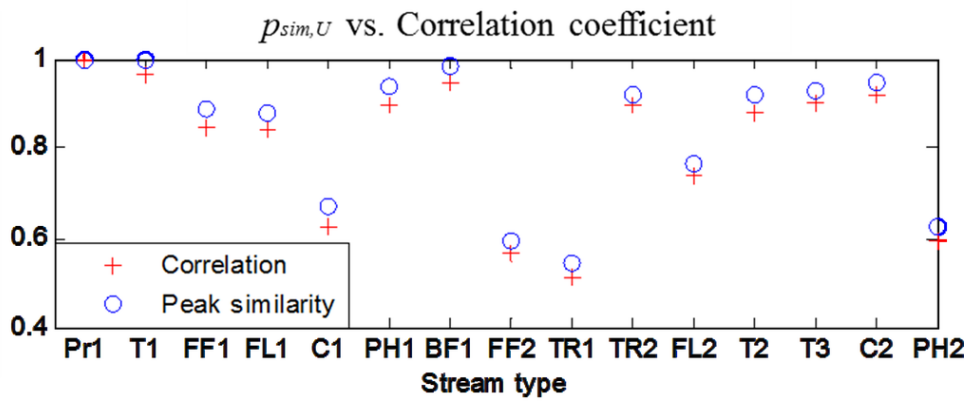


a) Conductivity sensor *b) pH sensor*

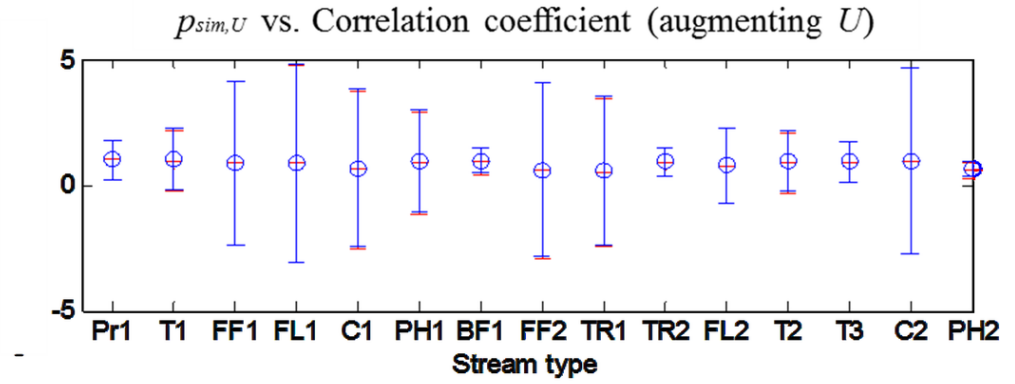


c) Temperature sensor

Figure 9. Comparison between uncertainty-aware peak similarity ($p_{sim,U}$) and correlation coefficient (corr) values, averaged over all sliding windows, between a reference pressure sensor and 15 distinct sensors: (a) original measurements without incorporating uncertainty, (b) results with confidence intervals by incorporating the estimated expanded uncertainty



(a) Averaged $p_{sim,U}$ vs. corr without expanded uncertainty



(b) Averaged $p_{sim,U}$ vs. corr with expanded uncertainty

As it can be seen, the original COL method (upper plot) is only capable of identifying as extreme events only those measurements which are strictly higher than 8.5. On the other hand, the uncertainty-aware version (bottom plot) enables the detection of additional points (orange dots) as possible extreme events. Doing so, a system operator is now triggered to focus more on these time instants for which the corresponding measurements are close to be identified as potential extreme events.

Next, we evaluate the performance of the original and modified POT method. For this, we apply both approaches on the temperature sensor shown in Figure 8, by setting the upper operating limit to 17.5 ^{o}C.

Figure 12 shows the output of both the COL (ref. Figure 12(a)) and POT ((ref. Figure

Figure 10. Comparison of execution times, as a function of the stream length, for four methods: a) Peak Similarity (our proposed), b) StatStream, c) BRAID, and d) Naïve Method (correlation coefficient)

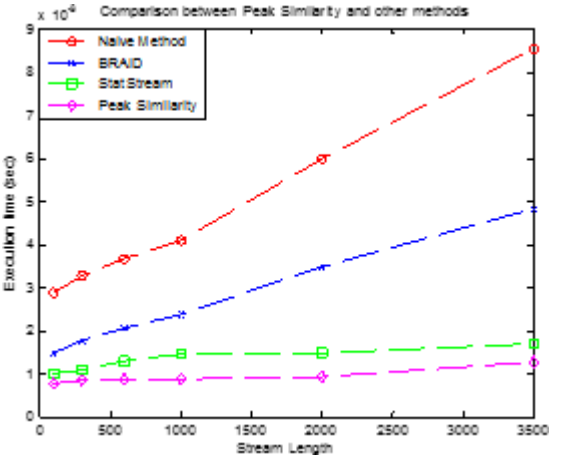
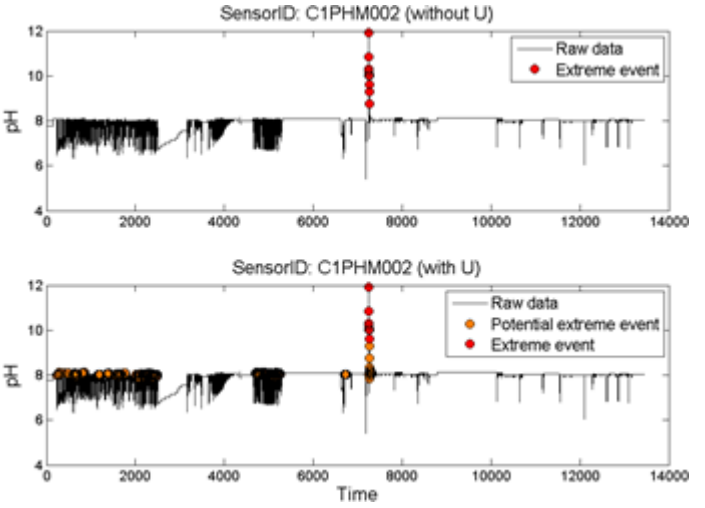


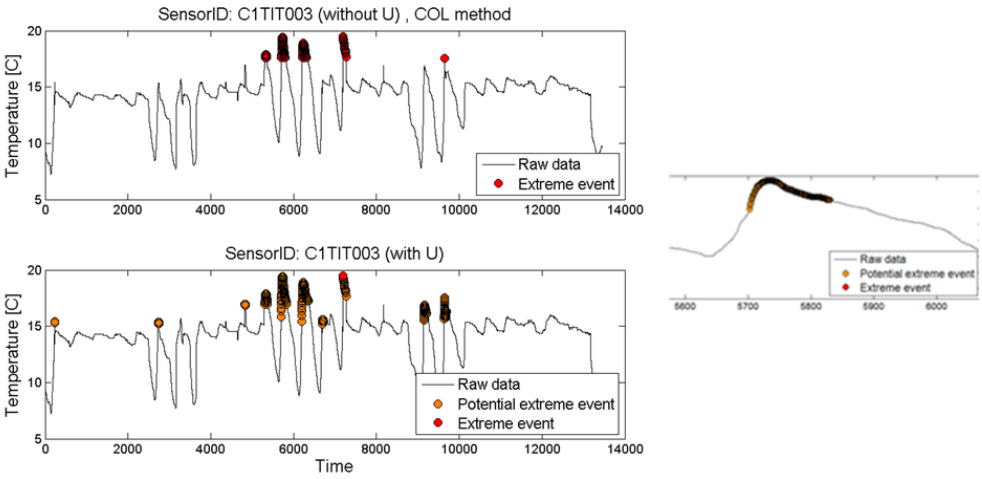
Figure 11. Extreme events detection performance of the original (upper plot) and the uncertainty-aware (bottom plot) COL method for a pH sensor (window size = 80, step size = 1). The red dots correspond to detected extreme events, while the orange dots correspond to possible extreme events



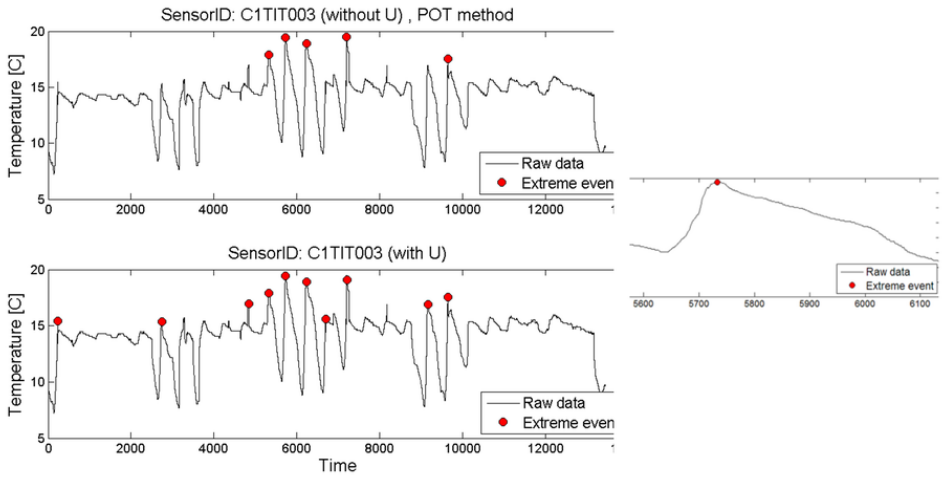
12(b))) extreme event detectors, with and without incorporating the estimated expanded uncertainty. The first remark is that, in contrast to the uncertainty-aware COL method which enables an intermediate state of "potential extreme events", the uncertainty-aware POT method identifies explicit extreme events. Despite of that, both methods are able to identify the extreme peaks in the recorded data stream.

The key difference between the two methods is that POT notifies for an extreme event only when we reach the peak of the curve. On the other hand, COL starts notifying for possible extreme events when the curve of the un-

Figure 12. Extreme events detection performance of: (a) the original (upper plot) and the uncertainty-aware (bottom plot) COL method; (b) the original (upper plot) and the uncertainty-aware (bottom plot) POT method for a temperature sensor (window size = 80, step size = 1). The red dots correspond to detected extreme events, while the orange dots correspond to possible extreme events



(a) Original and uncertainty-aware COL method



(b) Original and uncertainty-aware POT method

certainty-augmented measurements surpasses the defined threshold. Indeed, this can be seen clearly in the corresponding zoomed regions shown in Figure 12. Based on this observation, the uncertainty-aware COL can be considered to be more tolerant in detecting extreme events, when compared to the uncertainty-aware analogue of the POT method. A system operator may benefit from using the first method in the sense that the UADM system will start send-

ing notifications prior to the occurrence of the extreme event. As such, the uncertainty-aware COL method is mainly adopted in the integrated HYDROBIONETS platform.

CONCLUSION

Designing efficient data management systems capable of accounting for the inherent data uncertainty and providing early warning notifications is a challenging task in large-scale industrial infrastructures. In this paper we proposed an integrated uncertainty-aware data management system, which also supports timely detection of extreme events and fast online monitoring of pairwise sensor similarities in order to guarantee the validity of the detected extreme events. Comparison with state-of-the-art stream processing techniques revealed an improved performance of our proposed framework in terms of achieving accurate detection of extreme events, in conjunction with extraction of highly similar (correlated) pairs of possibly heterogeneous sensors, with significantly decreased execution times.

As a final outcome, we envisage to provide a set of data services to manipulate sensor measurements in large-scale industrial infrastructures, as well as to identify appropriate monitoring tools for the characterization of the generated data quality in real time. As a further extension, we will focus on the design of an automatic rule for the time-varying adaptation of the threshold $\varepsilon_{th,new}$, as well as the design of novel similarity measures of even lower computational complexity, while still approximating accurately the behavior of the correlation coefficient. Furthermore, a more thorough examination will be performed to design alternative extreme event detectors based on the variations of the estimated expanded uncertainty, as it was motivated by observing the results in Figure 8. Finally, an extension of the uncertainty-aware POT method so as to support an additional option for detecting "possible extreme events", similarly to the uncertainty-aware COL method, will be also studied.

ACKNOWLEDGMENT

This work has been supported by the HYDROBIONETS project (ICT-2011-7) funded by the European Commission under FP7 (GA-2011-287613).

REFERENCES

Aggarwal, C. (2009). *Managing and mining uncertain data*. Springer. doi:10.1007/978-0-387-09690-2

Canton, B., Erdman, W., Irvine, J., Lim, L., Mclaren, F., Meisel, R., & Speijer, J. etal. (2002). *Mathematics of Data Management*. McGraw-Hill.

Datar, M., Gionis, A., Indyk, P., & Motwani, R. (2002). Maintaining stream statistics over sliding windows. *SIAM Journal on Computing*, *31*(6), 1794–1813. doi:10.1137/S0097539701398363

Gehrke, J., Korn, F., & Srivastava, D. (2001). On computing correlated aggregates over continual data streams. In ACM SIGMOD Record, 30(2), 13-24, ACM. doi:10.1145/375663.375665

McNeil, A., & Saladin, T. (1997). The peaks over thresholds method for estimating high quantiles of loss distributions. In *Proceedings of 28th International ASTIN Colloquium*.

Rafiei, D., & Mendelzon, A. (1997). Similarity-based queries for time series data. In ACM SIGMOD Record, 26(2), 13-25, ACM. doi:10.1145/253260.253264

Sakurai, Y., Papadimitriou, S., & Faloutsos, C. (2005). Braid: Stream mining through group lag correlations. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data* (pp. 599-610). ACM. doi:10.1145/1066157.1066226

Tran, T. T., Peng, L., Diao, Y., McGregor, A., & Liu, A. (2012). CLARO: modeling and processing uncertain data streams. *The VLDB Journal—The International Journal on Very Large Data Bases, 21*(5), 651-676. doi: 10.1007/s00778-011-0261-7

Tran, T. T., Peng, L., Li, B., Diao, Y., & Liu, A. (2010). PODS: a new model and processing algorithms for uncertain data streams. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data* (pp. 159-170). ACM, Indianapolis, June 6 – 11. doi:10.1145/1807167.1807187

Yeh, M. Y., Wu, K. L., Yu, P. S., & Chen, M. S. (2009). PROUD: a probabilistic approach to processing similarity queries over uncertain data streams. In *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology* (pp. 684-695). ACM. doi:10.1145/1516360.1516439

Zhu, Y., & Shasha, D. (2002). Statstream: Statistical monitoring of thousands of data streams in real time. In *Proceedings of the 28th international conference on Very Large Data Bases* (pp. 358-369). VLDB Endowment, Hong Kong, China, Aug. 20 – 23. doi:10.1016/B978-155860869-6/50039-1

ENDNOTES

1 http://www.hydrobionets.eu

2 http://aca-web.gencat.cat/aca/documents/ca/sensibilitzacio/desal_Tordera/dessalinitzacio_en.pdf