Characterizing Long-term Wear and Tear of Ion-Selective pH Sensors

Kito Ohmura[a,b], Christian M. Thürlimann[a,c], Marco Kipf[a], Juan Pablo Carbajal[a], Kris Villez[a,c,*]

[a] *Eawag, Überlandstrasse 133, CH-8600 Dübendorf, Switzerland*
[b] *Toshiba Infrastructure Systems & Solutions Corporation, Tokyo, Japan*
[c] *ETH Zürich, Institute of Environmental Engineering, 8093 Zürich, Switzerland*

Abstract

The development and validation of methods for fault detection and identification in wastewater treatment research today relies on two important assumptions: *(i)* that sensor faults appear at distinct times in different sensors and *(ii)* that any given sensor will function near-perfectly for a significant amount of time following installation. In this work, we show that such assumptions are unrealistic, at least for sensors built around an ion-selective measurement principle. Indeed, long-term exposure of sensors to treated wastewater shows that sensors exhibit important fault symptoms that appear simultaneously and with similar intensity. Consequently, our work suggests that focus of research on methods for fault detection and identification should be reoriented towards methods that do not rely on the assumptions mentioned above. This study also provides the very first empirically validated sensor fault model for wastewater treatment simulation and we recommend its use for effective benchmarking of both fault detection and identification

*Corresponding author, Email *kris.villez@eawag.ch*

methods and advanced control strategies. Finally, we evaluate the value of redundancy for the purpose of remote sensor validation in decentralized wastewater treatment systems.

Keywords: data quality, drift, fault detection and identification, ion-selective electrodes, predictive maintenance, wastewater

1. Introduction

By several accounts, the lack of online sensor data quality poses a long-standing challenge for both the advancement of environmental science and engineering practice (Rieger et al., 2005, 2006; Rosén et al., 2008; Rieger et al., 2010; Haimi et al., 2013; Corominas et al., 2018). It is therefore not surprising that considerable time and energy has been invested in methods for automated quality assessment and quality control of online measurement devices (e.g., Thomann et al., 2002; Thomann, 2008; Corominas et al., 2011; Spindler and Vanrolleghem, 2012; Alferes et al., 2013; Spindler, 2014; Villez and Habermacher, 2016; Le et al., 2018).

Methods that are finding their way into practice today mainly consist of sanity checks. In the authors' experience, these work rather well to detect and classify a subset of commonly recognized fault symptoms, including outliers, spikes, stuck, and out-of-range values. For sensor faults that lead to more subtle symptoms, current practice relies primarily on regular on-site sensor maintenance, e.g. once every one or two weeks, to counter such subtle faults. For unstaffed wastewater treatment plants, on-site maintenance may be feasible economically only if this is limited to once per year. This practical constraint to the adoption of quality assessment and control practices forms the primary motivation for this study.

The literature suggests that data-analytical techniques can enable automated and remote detection of sensor faults. Without exception, such techniques rely on redundant relationships and can therefore be categorized by the type of redundancy that is used. A first category consists of techniques relying on reference measurements and computing a deviation between online sensor signal and the reference signal. A second category relies on hardware redundancy by placing multiple online sensors, possibly built around a distinct measurement principle, in the same location and then computing deviations between them. A third category relies on temporal redundancy, essentially assuming that meaningful changes in the sensor signal can only be smooth when measured with a sufficiently high frequency. Finally, the fourth category relies on spatial redundancy, relating signals produced at distinct locations or for different measured variables. Examples of this last category include both methods based on first principles, e.g. balance equations, as well as methods rooted in statistical practice, e.g. principal component analysis. Importantly, each of these advanced methods require tuning to maximize the number of true alarms and to ensure suitable quality control efforts while simultaneously minimizing the number of false alarms and futile maintenance actions. Invariably, such tuning is obtained by means of a historical, fault-free data set from which acceptable limits for computed residuals are derived. Consequently, this means that these methods rely on the availability of representative data of an acceptable quality. In addition, the use of most techniques implies that sensor fault symptoms can be assumed to appear independently from each other, i.e. the probability that two faults start at the same time is assumed to equal zero.

The prevalence of faults in actuators, sensors, and processes as well as the complexity of the fault detection and identification (FDI) task, has led to a plethora of methods that exploit one or more of the types of redundancy discussed above. In fact, the wealth of literature as well as the number of reviews on this or related topics (Venkatasubramanian et al., 2003c,a,b; Haimi et al., 2013; Corominas et al., 2018) suggest that the science and practice of FDI is all but settled, an observation also supported by no free lunch theorems (Wolpert, 1996).

Despite the tremendous amount of research on FDI methods, little is actually known about the cause-and-effect relationships between sensor ageing, the occurrence of sensor faults and failures, and the production of faulty data. This is explained by the fact that the availability of information describing the exact circumstances under which faults occur or faulty data is produced, i.e. meta-data, is usually severely limited. This is the secondary motivation of this study.

To facilitate performance evaluation of FDI tools, the formulation of simulation benchmarks has been an accepted practice in engineering sciences (Barty et al., 2006; Downs and Vogel, 1993). Similarly, the Benchmark Simulation Model No. 1 was conceived as a way to test and compare innovative FDI and control strategies (Jeppsson et al., 2007). Today, it is primarily used as a starting point for a family of plant-wide models of water resource recovery facilities (Nopens et al., 2009; Volcke et al., 2006). Actual benchmarking of FDI methods has been limited to one study so far (Corominas et al., 2011). The BSM family includes a set of sensor models which include sensor faults and this allows the user to add realism to the sensor signals. The simulated sensor faults always start at a time that is substantially later than the start of the simulated time. This provides ideal conditions for FDI method tuning as high-quality sensor data are always present in the first sections of the simulated data set. Moreover, a simulated fault always appears independently of any other sensor fault, i.e. no two sensor faults are simulated to start at the same time or with the same direction or magnitude. We expect that the situation in real-world conditions is very different. We thus hypothesize that typical fault symptoms will appear at the same time and with similar directions and magnitudes when exposed to the same harsh medium, especially when the same measurement principle is applied. Evaluating the merit of this hypothesis is the tertiary motivation of this study.

The following paragraphs are focused on the results and conclusions drawn directly from experimental data obtained during a long-term sensor exposure experiment. Additional insight is however obtained by studying a variety of dynamic models to describe our measurements.

2. Materials & Methods

2.1. Theoretical and real-world behavior of the ion-selective electrodes for pH measurement

The ion-selective measurement principle for pH measurement is understood rather well. According to the Nernst equation (Westcott, 2012) one measures an electric potential E (in mV), which is related to the activity of the protons, $[H^+]$, in the measured medium in steady state:

$$E = E^0 + \frac{RT}{F} \ln\left(\left[H^+\right]\right) \tag{1}$$

where E^0 is the reference potential, F is the Faraday constant ($96485.33289\, C\, mol^{-1}$, Taylor et al., 2007), $[H^+]$ is the proton activity in the reference cell, R is the molar gas constant ($8.3144598\, J\, mol^{-1}\, K^{-1}$, Taylor et al., 2007), and T is the temperature measured in Kelvin. The pH is defined as $-\log{[H^+]}$ (Buck et al., 2002) so that $S(T)$ is the temperature-specific sensitivity, which can be computed as:

$$S(T) = \frac{RT}{F\, \log{(e)}} \tag{2}$$

Most typically, pH sensors are designed to deliver 0 mV at pH 7 so that E^0 is theoretically 0 mV. Similarly, the theoretical sensitivity at standard ambient temperature and pressure (SATP) thus is $S(298.15) = 59.1593$ mV per pH unit. Because the actual values of these parameters tend to deviate from their theoretical values, it is common to identify their values through a 2-point calibration procedure. At the engineering department at Eawag, the most common practice is to use buffered calibration media with pH 4.01 and 7.00 for validation, followed by calibration when the absolute deviations between the produced pH measurements and the known pH values exceed a predetermined threshold. The data end user sets this threshold. Depending on the application, this ranges from 0.1 to 0.4 pH units. The theoretical potential at pH 4.01 and SATP is 177.0 mV.

2.2. Studied sensors

A total of 12 pH sensors are produced by Endress+Hauser (Reinach, Switzerland). These sensors consist of 5 sensor types (T1-T5) whose exact type cannot be revealed due to a confidentiality agreement. The first eight sensors consist of pairs of four commercially available sensor types (T1-T4) which are typically sold with a one-year warranty agreement. The first (second) sensor in each pair is designated with an *a* (*b*), e.g. T1a, T1b. The last 4 pH sensors are replicates of a recently developed sensor prototype (T5) and are referred to as T5a, T5b, T5c, and T5d.

The first three sensor pairs (T1-T3) have been in use throughout a long-term exposure experiment which lasted for 731 days (Oct. 4th, 2016 – Oct. 4th, 2018). An overview of this experiment is given in Fig. 1. The 4th pair

(T4) has been in use during the first half year and was replaced with the 5th pair (T5) on April 3rd, 2017 (day 182) as *(i)* the T4 sensors exhibit a long response time (not shown) and *(ii)* the opportunity arose to test the T5 prototypes. The T5a sensor stopped producing a meaningful signal on June 30th, 2017 (day 270) while T5b became faulty (details below) on August 31st, 2017 (day 332). These sensors were replaced with another sensor of the same prototype (T5) on Oct. 2nd, 2017 (day 364). In this last pair, one sensor (T5d) failed within 1 day (day 365) while the other (T5c) has been fully functional until the end of the experiment.

2.3. Long-term exposure experiment

The sensors are exposed to the contents of a reactor used primarily to study advanced control strategies for nitrite accumulation prevention in a urine nitrification process (Thürlimann et al., Submitted). To this end, the nitrified urine is pumped through a closed tube made from PVC with a flow rate of 43 L/h. The design of this tube equipped with sensor-holding locks is shown in the *Supplementary Information (Section B).*

The treated urine is from anthropogenic origin during the whole experimental period. The treated urine was collected from male lavatories in the Forum Chriesbach building at Eawag, with exception of the period from day April 30th, 2018 to June 21st, 2018 (day 574-625), when it was collected from female lavatories in the same building. From October 4th, 2017 to November 24th, 2017 (day 366 to 417), the reactor was additionally fed with a nitrite

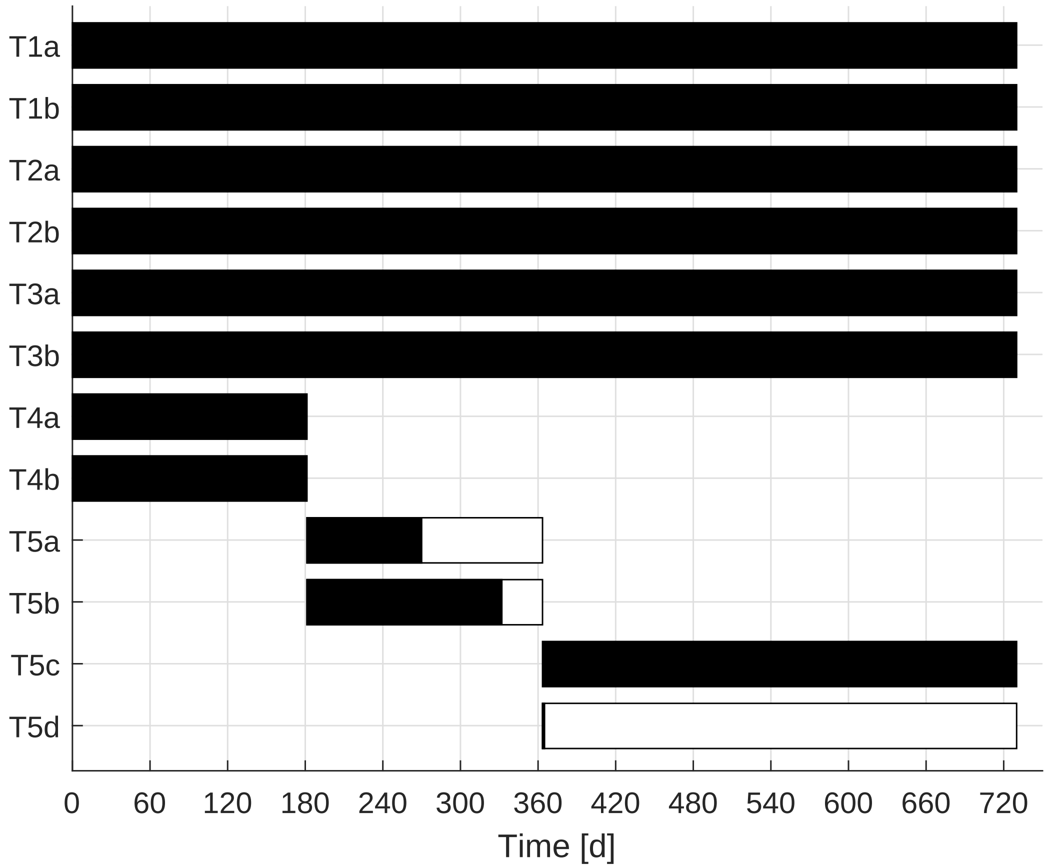

Figure 1: **Overview of the complete experimental campaign.** The periods of sensor exposure are indicated by rectangles. The periods during which the sensors produced meaningful data are marked black.

stock solution. During the experimental period, the measured concentrations of nitrogen species in the nitrified urine ranged between 1180 and 2730 mgN/L (mg atomic nitrogen per liter) for total ammonia, 0 and 82 mgN/L for nitrite, and 1290 and 2720 mgN/L for nitrate. These measurements are copied from Thürlimann et al. (Submitted) and are shown in the *Supplementary Information (Section C)*. The pH value of the nitrified urine, as measured by two independent and regularly calibrated pH sensors installed directly in the reactor, ranged between 5.7 and 7.3.

2.4. Sensor characterization tests

At regular intervals, the sensors were removed from their normal position and exposed to other media for sensor characterization. This was executed 47 times in total. The exact times of these sensor characterization tests are listed in the *Supplementary Information (Section G.1).* Two pairs of tests were executed on the same day to ensure acceptable experimental reproducibility (day 70: tests 11-12; day 351: tests 29-30). The selected media include (C4) pH 4.01 calibration solution (CPY20-C10A1, Endress+Hauser, Reinach, Switzerland); (C7) pH 7.00 calibration solution (CPY20-E10A1, Endress+Hauser, Reinach, Switzerland); (U4) nitrified urine at pH 4; (U7) nitrified urine at pH 7; and (W) tap water. For the present work, only the exposure to W, C4, and C7 is relevant. This occurs in five distinct phases (P0-P4), each lasting at least 5 minutes and exposing the sensors to W, C4, C7, C4, and W in this order. Exemplary results are shown in Fig. 2 and discussed in detail below.

Raw potential measurements recorded during P1, P2, and P3 are used to compute the offset ($\tilde{E}^0$) and two measurements of the sensitivity ($\tilde{S}_D$ and $\tilde{S}_R$). In line with (Carr, 1993), the following steps are applied for every sensor and every sensor characterization test:

1. Compute the median value among the potential measurements collected

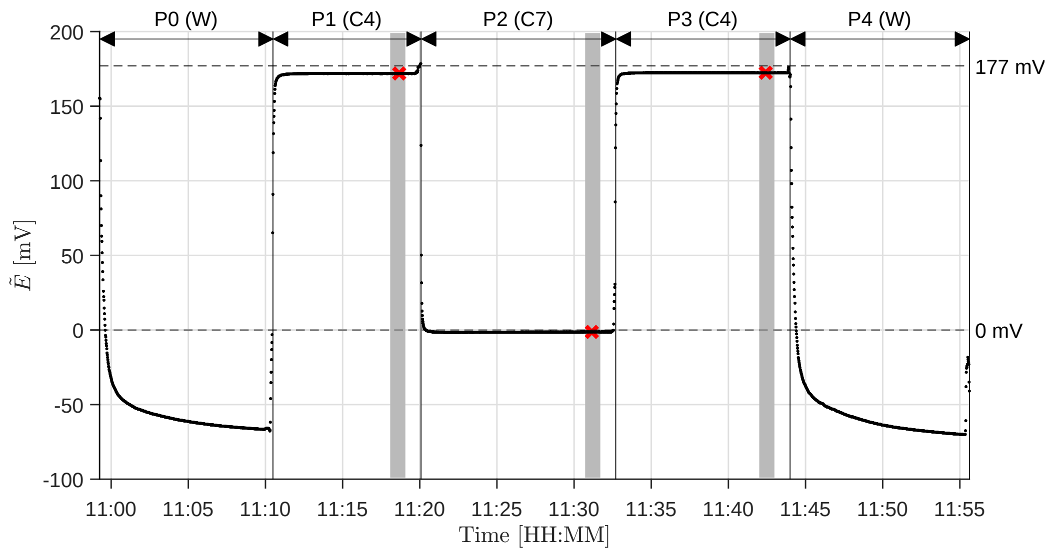

Figure 2: **Exemplary sensor characterization test.** Raw data obtained in the first sensor characterization test with sensor T1a. The measured potential decays during P0, P2, and P4, while it increases during P1 and P3. Steady state is reached quickly in P1, P2, and P3. The theoretical potential values for P1, P2, and P3 are indicated with dashed horizontal lines. Grey shading indicates the data used to obtain the potential measurements (2 to 1 minute before phase change). The selected median potential values are shown with red crosses.

in P1, P2, and P3 between 2 and 1 minutes before the start of the next phase (P2, P3, and P4). Refer to these values as E^{P1}, E^{P2}, and E^{P3}

2. The sensor offset is defined as $\tilde{E}^0 = \tilde{E}^{P2}$.
3. The decay potential sensitivity is defined as $\tilde{S}_D = \frac{\tilde{E}^{P1}-\tilde{E}^{P2}}{7.00-4.01} = \frac{\tilde{E}^{P1}-\tilde{E}^{P2}}{2.99}$.
4. The decay potential sensitivity is defined as $\tilde{S}_R = \frac{\tilde{E}^{P3}-\tilde{E}^{P2}}{7.00-4.01} = \frac{\tilde{E}^{P3}-\tilde{E}^{P2}}{2.99}$.

These steps are demonstrated below with a practical example.

2.5. Drift model

The results shown below indicate that the offset significantly varies over time while the sensitivity remains remarkably stable in all studied sensors.

We describe the observed drift of the offset by means of two models.

2.5.1. Model 1 - Constant trend followed by linear trend

For the first model, we apply a modified version of the excessive drift model proposed for the BSM family (Rosén et al., 2008). This model simulates $E^0(t)$, the sensor offset, as:

$$E^0(t) = d_o + r_d \, H\,(t - t_f) \tag{3}$$

with d_o the initial offset, r_d the drift rate parameter, $H\,(\cdot)$ the Heaviside function ($H(a) = 1$ if $a \geq 0$, $H(a) = 0$ otherwise), t the time since sensor installation, and t_f the time of the drift onset. The applied modification consists of adding the parameter d_o. To fit this model, the offset measurements, $\tilde{E}^0(t_h)$, collected at discrete time instants t_h, are assumed to exhibit independently and identically distributed measurement errors, ϵ_h, drawn from a normal distribution with zero mean and standard deviation, σ_ϵ:

$$\tilde{E}^0(t_h) = E^0(t_h) + \epsilon_h, \, \epsilon_h \sim N(0, \sigma_\epsilon) \tag{4}$$

Values for the 4 parameters d_o, t_f, r_d, and σ_ϵ are obtained independently for all sensors through maximum likelihood estimation (MLE). Once calibrated, the model is used to obtain the estimated mean and point-wise standard deviations for the sensor offset, $\mu_1(t) = \mathbb{E}\left(E^0(t)\right)$ and $\sigma_1(t)$, while using the estimates of t_f and σ_ϵ as fixed hyperparameter values.

2.5.2. Model 2 - Integrated Brownian motion for a single sensor

In model 2, we assume instead that the recorded offset measurements are generated by an integrated Brownian motion. This is a continuous-time stochastic process, which reflects that the drift rate is subject to unmeasured disturbances:

$$\dot{r}_d(t) = \gamma(t)dt,\ r_d(0) = r_{d,o},\ \gamma(t) \sim N(0, \sigma_\gamma), \tag{5}$$

$$\dot{E}^0(t) = r_d(t)dt,\ E(0) = d_o, \tag{6}$$

$$\tilde{E}^0(t_h) = E^0(t_h) + \epsilon_h,\ \epsilon_h \sim N(0, \sigma_\epsilon) \tag{7}$$

This model also includes 4 parameters: the initial drift rate ($r_{d,o}$); the initial offset (d_o); an input noise standard deviation controlling the rate by which the drift rate changes (σ); and an output noise standard deviation (σ_ϵ). As with model 1, parameter values are obtained through MLE. This is achieved by formulating the above process as a Gaussian process (Rasmussen and Williams, 2006). This also enables to compute expected values and associated point-wise standard deviations, $\mu_2(t) = \mathbb{E}\left(E^0(t)\right)$ and $\sigma_2(t)$, with the estimates of σ_γ and σ_ϵ now used as fixed hyperparameter values.

2.5.3. Model 3 - Integrated Brownian motion for multiple sensors

A third model is derived from Eqs. 5-7 by considering that two sensors of the same type may be characterized by distinct initial conditions ($r_{d,o}$, d_o) but the same noise parameters (σ_ϵ, σ_γ). This lead to a model with six parameters (d_o^a, d_o^b, $r_{d,o}^a$, $r_{d,o}^b$, σ_ϵ, σ_γ), instead of two models with 4 parameters each. Their values are again obtained via MLE and used to obtain calibrated predictions ($\mu_3(t) = \mathbb{E}\left(E^0(t)\right)$, $\sigma_3(t)$) , once again using the estimates of σ_γ and σ_ϵ as fixed hyperparameter values.

2.5.4. Model evaluation

The proposed models are evaluated through visual inspection of the measurements, predictions, and residuals between the measurements and predictions. In the present case, such a visual inspection is considered sufficient to select a suitable model.

2.5.5. Implementation

All data collected during the sensor characterization tests and all code necessary to reproduce our results is added in the *Supplementary Information (Section A)*.

3. Results

3.1. Sensor characterization tests: Example

Fig. 2 shows the data obtained in the first sensor characterization test with sensor T1a on Oct. 6th, 2016 (day 3). The raw potential measurement decreases during P0, increases to a steady value in P1, decreases to a steady value in P2, increases to a steady value in P3, and decreases again in P4. The time intervals used for computation of $\tilde{E}^{P1}$, $\tilde{E}^{0}$, and $\tilde{E}^{P3}$ (in calibration medium, pH = 4, 7, and 4) are indicated by grey shading. One can see that the measured offset $\tilde{E}^{0}$ is slightly below 0 mV (−1.30 mV). The values for $\tilde{E}^{P1}$ and $\tilde{E}^{P3}$ are slightly lower than their ideal value (171.9 and 172.4 mV). The measured rise and decay sensitivities are therefore $\tilde{S}_D = 57.73$ and $\tilde{S}_R = 57.90$ mV per pH unit. The results of every sensor characterization test are visualized in the *Supplementary Information (Section G.2)*.

3.2. Long-term trends in the offset measurements within the warranty period

Fig. 3 displays the measured offsets in all sensors throughout the experimental period. The recorded values collected within the warranty period (1 year) range from approximately 0 mV (no offset) to roughly −70 mV. All commercially available sensors (T1-T4) produce a decaying trend in the offsets. The firstly recorded offsets for the T1-T3 sensors are small in magnitude and concentrate around 0 mV. In contrast, the T4 sensors offset values indicate a shock effect producing a shift of −20 and −45 mV (T4a, T4b) within days from installation. This is explained by the manufacturer as an effect of the high ammonium concentration in the medium and should only be expected for this specific type of sensors. The accumulated drift in the T1 sensors is at most −25 mV after one year while the T2 and T3 sensors exhibit an offset of −75 mV after one year. Without calibration, this means the T1 sensors can produce a pH value as high as 7.4 when the true pH is 7. The T2 and T3 sensors will produce a pH value as high as 8.3 in the same circumstances. Due to failure of T5d, no offsets could be measured for this sensor. The remaining prototypes (T5a/b/c) do not produce a significant offset at any time, except for T5b which produces a dramatic shift in the offset during three sensor characterization tests executed prior to replacement. A detailed inspection of the T5b measurements revealed that the first symptoms of sensor degradation can be observed on August 31st, 2017 (day 332). This is however only obvious when comparing these measurements with the simultaneous T1b/T2b/T3b measurements (see the *Supplementary Information, Section D*). In all cases, except for the T4 and T5a/b pairs, the difference between offsets in sensors of the same type remains rather small with 1 year of installation, with a maximal difference of 16.7 mV recorded with the T2 sensors. Taking the 0.1 pH threshold discussed above as a guideline, one could propose to validate and calibrate the sensors when their potential measurements are 5.9 mV apart. This happens for the first time for the T1, T2, and T3 sensors on day 127, 79, and 309. By these times, the absolute offsets are already larger than this accepted threshold so that the relative difference between sensors of the same type is unlikely a good measure to trigger sensor maintenance.

Fig. 4 shows offsets for the sensors T1a, T3a, and T3b collected in the first year of the experiment as a function of the difference in the offset between T1a and T3a (left panel) and T3b and T3a (right panel). The left panel

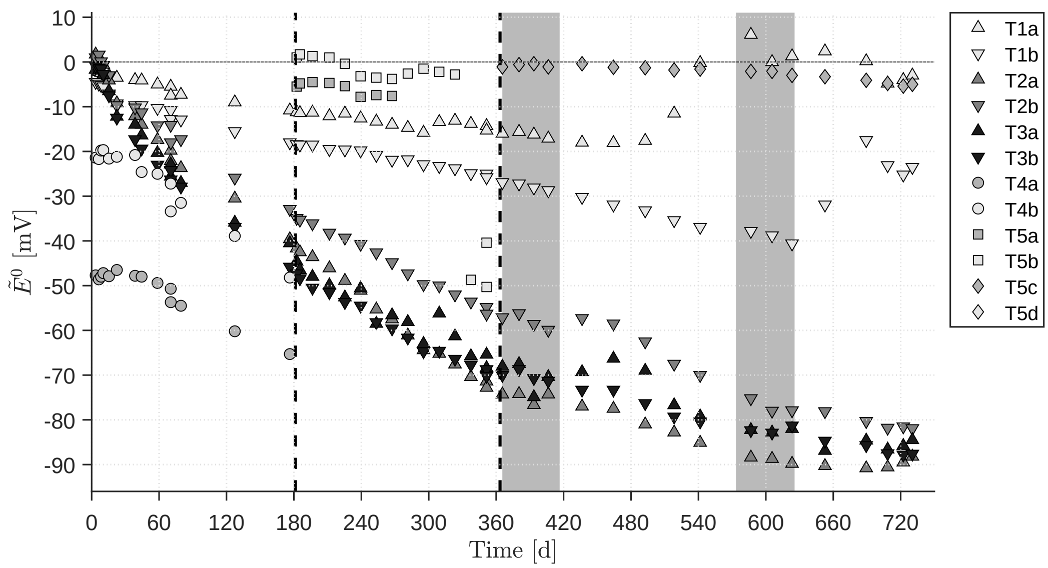

Figure 3: **Offset in all studied sensors as a function of time.** Vertical lines indicate a change of installed sensors (see Fig. 1). Grey bands indicate a change of reactor medium (see Section 2.3). The commercially available sensors (T1-T4) exhibit drift from the start of installation while the prototypes (T5) exhibit close to no drift when otherwise functioning properly. A significant shock effect is observed for the T4 sensors at the start of the experiment but not for any other sensor.

suggests that offset difference between sensors can be predictive of the offset in an individual sensor. The right panel shows that this is less likely to be successful for sensors of the same sensor type, as also described above. This is considered an important opportunity for further research, which we discuss further below.

3.3. Long-term trends in the offset measurements beyond the warranty period

The offset measurements obtained after the warranty period expired exhibit two phenomena that are surprising (Fig. 3). The first phenomenon is the rise of the offset of the T1a sensor after 480 days of exposure and a similar

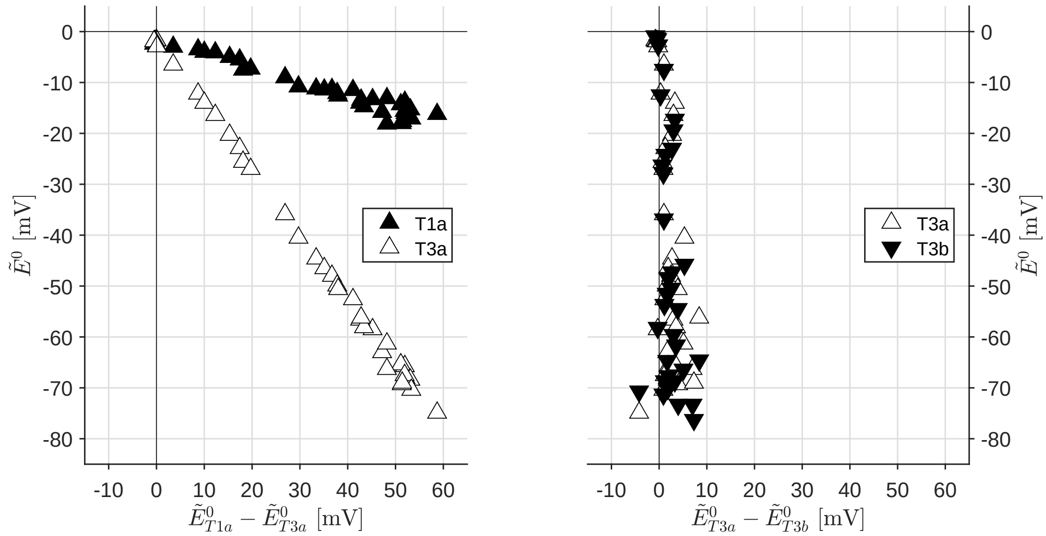

Figure 4: **Offset measurements as a function of relative deviations in the offset measurements.** *Left panel:* Offsets of sensor T1a and T3a as a function of the difference of these offsets. These data are suggestive of a close to linear relationship between sensor offsets and the offset difference. *Right panel:* Offsets of sensors T3a and T3b relative to the difference of these offsets. The difference in offset remains small and there is no obvious relationship in this case.

rise of the offset of the T1b sensor after 630 days of exposure. Considering that this appears at distinct times in the lifetime of the T1 sensors, this cannot be explained as a direct effect of medium composition changes. Based on information provided by the sensor manufacturer, this type of drift rate sign reversal is unique for the T1 sensors and is unlikely to be observed with any other sensor type covered in this study. It is the opinion of the authors that the time for this reversal is difficult to predict in advance. For this reason, this phenomenon is best handled as an unmeasured process disturbance.

The second phenomenon consists of the rather flat to increasing profile of the offset measurements in the T2 and T3 sensors between day 360 and day

480. Before and after this period, the drift rate in these sensors are visually similar. Given the synchronicity of this effect between 4 pH sensors, it is hypothesized that this change in the drift rate is influenced by the deliberate addition of nitrite in the form of $NaNO_2$ salt to the reactor contents from day 366 to 417. The nitrite addition affected the biomass concentration and the concentrations of all dominant nitrogen species (ammonia, nitrite, nitrate, see *Supplementary Information, Section C*) and may also have affected the ion strength and conductivity of the reactor contents. Due to this combination of effects, the available data only offers an incomplete understanding of the complete chain of causes and effects between the nitrite addition and the observed changes in the sensor drift rates. For this reason, the effects of changing media composition on the sensor drift rate is best also considered an unmeasured process disturbance.

3.4. Long-term trends in the sensitivity measurements

Fig. 5 displays the computed sensitivity measurements for the potential rise ($\tilde{S}_R$) during the complete experimental period. These measurements do not exhibit strong trends in any particular direction. The sensitivity measurements fall between 54.9 and 62.1 mV per pH unit. This means that one can expect to measure a pH value between 5.95 and 6.08 when *(i)* the true pH value is 6 and *(ii)* any offset is corrected for. The same graph also shows the theoretical value of the sensitivity according to (2) and the recorded temperature. This profile is very similar to the recorded sensitivity profiles and explains most of the variations in the sensitivity measurements, which are small anyway. The same conclusions are drawn from the computed sensitivity measurements for the potential decay ($\tilde{S}_D$, see *Supplementary Information, Section E*).

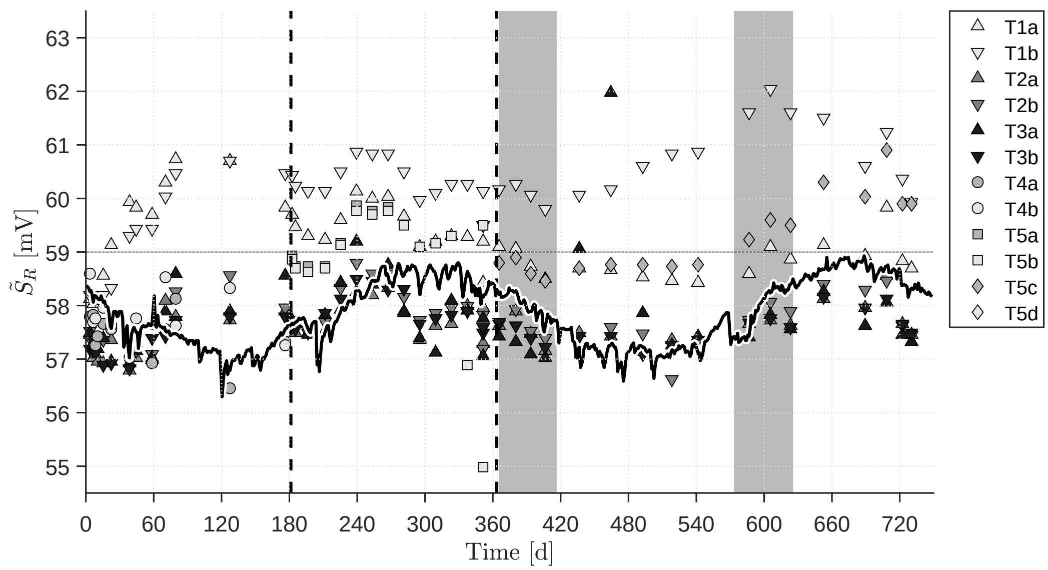

Figure 5: **Sensitivity measurements for the potential rise as a function of time.** Vertical lines indicate a change of installed sensors (see Fig. 1). Grey bands indicate a change of reactor medium (see Section 2.3). A black line shows the theoretically expected sensitivity computed with (2). Variations in the sensitivity are small and follow the theoretical sensitivity closely.

3.5. Drift models

For practical intents and purposes, the sensitivity – when corrected for temperature variations – can be considered constant for the considered process and sensors. We therefore focus on further analysis of the offset measurements.

The left panel of Fig. 6 shows the offset measurements for the T2a and T2b sensor together with the model predictions and their confidence bounds. The right panel of Fig. 6 shows the prediction residuals. With Model 1, the time of the drift onset (t_f) is always identified as a time before the first measurement was obtained (2.1 and 2.3 days), suggesting that drift occurs throughout the experiment. The same kind of result is obtained with every other commercially available sensor type (T1-T4), except for the T1a sensor (see the *Supplementary Information (Section F)*). More importantly however is that Model 1 offers a rather poor description of the data. The confidence intervals are wide and the residuals are clearly auto-correlated. In contrast, Models 2 and 3 provide narrower confidence intervals and residuals that do not suggest presence of autocorrelation. There are no clear differences in performance between these two models so that Model 3, which has fewer free parameters, is preferred. The modeling results for the T1 and T3 sensors lead to the same conclusions. For these results and all parameter estimates, we refer to the *Supplementary Information (Section F)*. For the T4 sensors, all model types delivered the same, adequate performance. This may indicate that (a) the T4 sensors exhibit a drift which is influenced less by unmeasured disturbances and therefore occurs with a close to constant rate or (b) that the shortened exposure – 6 months in this case – was too short to capture the long-term effects of unmeasured disturbances.

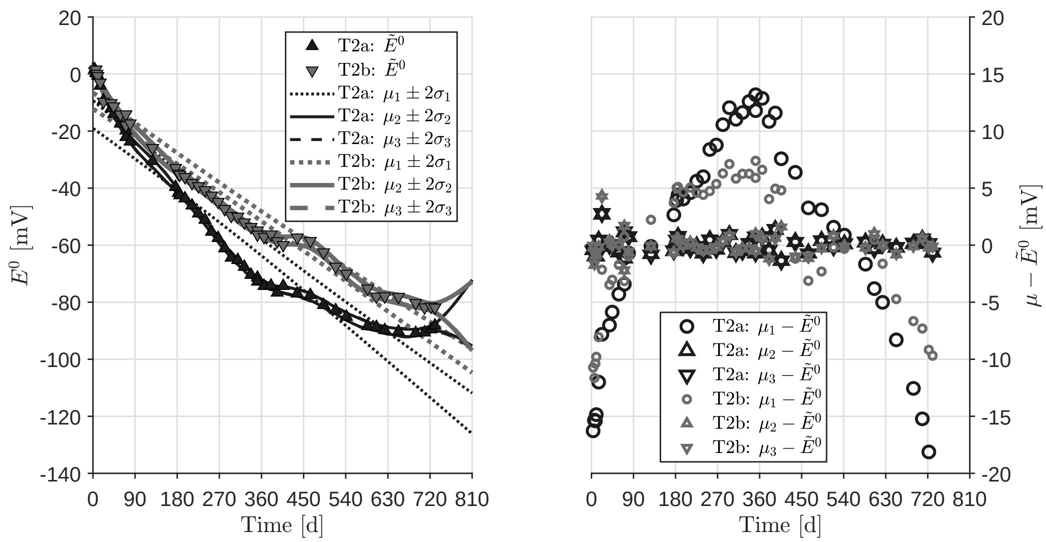

Figure 6: **Modeling results for the T2 sensors.** *Left panel:* Offset confidence bounds ($\mu \pm 2\,\sigma$) obtained with models 1 (μ_1, σ_1), 2 (μ_2, σ_2), and 3 (μ_3, σ_3). *Right panel:* Residuals between expected values (μ) and measured potentials ($\tilde{E}^0$). Model 1 does not describe the data well, leading to larger confidence bounds and auto-correlated residuals. Models 2 and 3 fit the data well and their predictions are hard to distinguish from each other.

4. Discussion

This study present the first peer-reviewed results with which the effect of long-term wear-and-tear on water quality sensors deployed in wastewater treatment plants is assessed and evaluated in a systematic manner and at this scale (12 sensors). The experimental results reveal that commonly held assumptions regarding the occurrence of sensors faults and fault symptoms are false. First, it is demonstrated that drift in pH sensors occurs simultaneously in all commercially available sensors. Second, it is demonstrated that drift occurs as soon as a sensor is deployed in the measured medium. In some cases, the immediate onset of drift is paired by a significant shift in the offset. Importantly, the data needed to compute the offsets and sensitivities as a function of time are also available in modern pH instruments in the form of a calibration logbook that can be accessed through standardized communication protocols (e.g., Modbus).

These observations have important consequences for the development of methods for fault detection and identification (FDI). Indeed, *(i)* one cannot assume that faults appear independently in distinct sensors and *(ii)* one cannot assume to have access to a fault-free historical data set. Naturally, this also holds in the context of simulation-based benchmarking of FDI methods. Consequently, it is our opinion that the development of FDI methods and model-based benchmarking should be focused on methods that do not rely on such assumptions.

Fortunately, our results also reveal a number of opportunities for the use and maintenance of ion-selective measurements. First, the prototype sensors tested in this study exhibit a remarkably stable offset. While these sensors appear prone to failure, as one might expect from a prototype, this suggests that practically drift-free yet economical pH sensors will enter the market soon. Second, the recorded sensitivity measurements in all sensors hover around the ideal values and are remarkably stable throughout the experimental period. Such a stable sensitivity lends support for advanced monitoring and control strategies which are inherently robust to changes in the offset but still assume a rather stable sensitivity (Villez and Habermacher, 2016; Thürlimann et al., 2018a,b). Third, it was shown that the offset difference between two pH sensors in the same medium can be predictive of the offset of the individual pH sensors, however only if two sufficiently distinct sensor types are selected. Combined with a stable sensitivity, this means that the deviation between two online pH sensor signals could be used as a proxy for the deviation in each individual sensor. Such a proxy measurement could be very useful for remote sensor quality assessment and predictive sensor maintenance, especially since one can compute such deviations between on-line sensor signals while the sensors remain in their normal measurement location in the monitored reactor.

The obtained offset measurements were studied in more detail by comparing the fit of 3 models. From this, it is concluded that the excessive drift model included in the BSM family (Rosén et al., 2008; Gernaey et al., 2014) cannot adequately describe the naturally occurring drift in ion-selective electrodes. Instead, the proposed stochastic model, specifically an integrated Brownian process, delivers a good description of the obtained data sets. In the authors' opinion, such a model should be included in the BSM family for realistic simulation of measurements obtained through ion-selective measurement principles. The obtained model also enables prediction of the expected offset measurement and associated confidence intervals beyond the last measurement. This means that such a model can be used for predictive sensor maintenance, e.g., by planning a new sensor validation and/or calibration before the predicted confidence interval exceeds a predetermined tolerance, each time also updating the parameters of the stochastic model. For this, confidence intervals for the reference potential (E^0) rather than for the measurements ($\tilde{E}^0$) are expected to be most useful. Exploring the utility of this idea is considered for future research.

5. Conclusions

Despite the abundance of literature of fault detection and identification (FDI) methods, little is actually known about the cause-and-effect relationships between the exposure of water quality sensors to harsh conditions, such as wastewater media, and the occurrence of sensor faults and failures. This first long-term study of the ageing of 12 individual pH sensors gives valuable insight into this challenge. First, it is concluded that commonly held assumptions in FDI method development and evaluation, such as the availability of fault-free historical data and independent onsets of sensor faults, are invalid for pH sensors based on the ion-selective measurement principle. In addition, the effects of offset drift in redundant sensors is unlikely to be identified early if these sensors are of the exact same type and exposed to the same medium. A stochastic model is shown to offer a good description of the observed drifts of the sensor offsets and perform better than a previously established drift model. Finally, our results suggest that newly developed pH sensors which exhibit stable offsets will enter the commercial market soon.

6. Acknowledgements

This research was made possible by the Swiss National Foundation (Project: 157097). Mr. Ohmura's contributions are financially supported by Toshiba, Tokyo, Japan. Dr. Carbajal's contributions are financially supported by Eawag Discretionary Funds (grant no.: 5221.00492.011.05, project: DF2017/EMUmore). All sensors were provided at no cost by Endress+Hauser. We thank Daniel Iten and Stefan Vogel from Endress+Hauser for their valuable advice and input to this study. The stochastic models were calibrated with the GPML toolbox v4.2 (Rasmussen and Nickisch, 2005).

References

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

Barty, M., Patton, R., Syfert, M., de las Heras, S., Quevedo, J., 2006. Introduction to the DAMADICS actuator FDI benchmark study. Control Engineering Practice 14 (6), 577–596.

Buck, R., Rondinini, S., Covington, A., Baucke, F., Brett, C., Camoes, M., Milton, M., Mussini, T., Naumann, R., Pratt, K., Spitzer, P., 2002. Measurement of pH. definition, standards, and procedures (iupac recommendations 2002). Pure and Applied Chemistry 74 (11), 2169–2200.

Carr, J. J., 1993. Sensors and circuits. PTR Prentice Hall.

Corominas, L., Garrido-Baserba, M., Villez, K., Olsson, G., Cortes, U., Poch, M., 2018. Transforming data into knowledge for improved wastewater treatment operation: A critical review of techniques. Environmental Modelling and Software 106, 89–103.

Corominas, L., Villez, K., Aguado, D., Rieger, L., Rosén, C., Vanrolleghem, P. A., 2011. Performance evaluation of fault detection methods for wastewater treatment processes. Biotechnology and Bioengineering 108 (2), 333–344.

Downs, J. J., Vogel, E. F., 1993. A plant-wide industrial process control problem. Computers and Chemical Engineering 17 (3), 245–255.

Gernaey, K. V., Jeppsson, U., Vanrolleghem, P. A., Copp, J. B., 2014. Benchmarking of control strategies for wastewater treatment plants. Scientific and Technical Report No. 23. IWA Publishing.

Haimi, H., Mulas, M., Corona, F., Vahala, R., 2013. Data-derived soft-sensors for biological wastewater treatment plants: An overview. Environmental Modelling and Software 47, 88–107.

Jeppsson, U., Pons, M. N., Nopens, I., Alex, J., Copp, J. B., Gernaey, K. V., Rosén, C., Steyer, J., Vanrolleghem, P. A., 2007. Benchmark Simulation Model No. 2: General protocol and exploratory case studies. Water Science and Technology 56(8) (67-78).

Le, Q. H., Verheijen, P. J., van Loosdrecht, M. C., Volcke, E. I., 2018. Experimental design for evaluating WWTP data by linear mass balances. Water Research 142, 415–425.

Nopens, I., Batstone, D. J., Copp, J. B., Jeppsson, U., Volcke, E., Alex, J., Vanrolleghem, P. A., 2009. An ASM/ADM model interface for dynamic plant-wide simulation. Water Research 7, 1913–1923.

Rasmussen, C. E., Nickisch, H., 2005. Gaussian processes for machine learning (GPML) toolbox.
URL `https://gitlab.com/hnickisch/gpml-matlab/`

Rasmussen, C. E., Williams, C. K., 2006. Gaussian processes for machine learning. MIT press.

Rieger, L., Langergraber, G., Siegrist, H., 2006. Uncertainties of spectral in situ measurements in wastewater using different calibration approaches. Wat. Sci. Technol. 53(12), 187–197.

Rieger, L., Takács, I., Villez, K., Siegrist, H., Lessard, P., Vanrolleghem, P. A., Comeau, Y., 2010. Data reconciliation for wastewater treatment plant simulation studies – planning for high-quality data and typical sources of errors. Water Environment Research 82, 426–433.

Rieger, L., Thomann, M., Gujer, W., Siegrist, H., 2005. Quantifying the uncertainty of on-line sensors at WWTPs during field operation. Water Research 39 (20), 5162–5174.

Rosén, C., Rieger, L., Jeppsson, U., Vanrolleghem, P. A., 2008. Adding realism to simulated sensors and actuators. Water Science and Technology 57 (3), 337–344.

Spindler, A., 2014. Structural redundancy of data from wastewater treatment systems. determination of individual balance equations. Water Research 57, 193–201.

Spindler, A., Vanrolleghem, P. A., 2012. Dynamic mass balancing for wastewater treatment data quality control using CUSUM charts. Water Science and Technology 65 (12), 2148–2153.

Taylor, B. N., Mohr, P. J., Douma, M., 2007. The NIST reference on constants, units, and uncertainty.
URL `physics.nist.gov/cuu/index`

Thomann, M., 2008. Quality evaluation methods for wastewater treatment plant data. Wat. Sci. Technol. 10, 1601–1609.

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

Thürlimann, C. M., Dürrenmatt, D. J., Villez, K., 2018a. Soft-sensing with qualitative trend analysis for control in full-scale wastewater treatment plants. Control Engineering Practice 70, 121–133.

Thürlimann, C. M., Udert, K. M., Morgenroth, E., Villez, K., 2018b. Assessment of two qualitative trend analysis tools for process control. In: 4th IWA Specialized International Conference "Ecotechnologies for Wastewater Treatment" (EcoSTP2018), London, ON, Canada, June 25-27, 2018.

Thürlimann, C. M., Udert, K. M., Morgenroth, E., Villez, K., Submitted. Handling sensor drift for stabilizing nitrite control during nitrification of high strength wastewater by means of qualitative trend analysis.

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

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

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

Villez, K., Habermacher, J., 2016. Shape anomaly detection for process monitoring of a sequencing batch reactor. Computers & Chemical Engineering 91, 365–379.

Volcke, E. I., van Loosdrecht, M. C., Vanrolleghem, P. A., 2006. Continuity-based model interfacing for plant-wide simulation: A general approach. Water Research 15, 2817–2828.

Westcott, C., 2012. pH measurements. Elsevier.

Wolpert, D. H., 1996. The lack of a priori distinctions between learning algorithms. Neural Computation 8 (7), 1341–1390.