To calibrate or not to calibrate, that is the question

Oscar Samuelsson a, b, *, Erik U. Lindblom b, c, Anders Björk b, Bengt Carlsson a

a Uppsala University, Sweden
b IVL Swedish Environmental Research Institute, Sweden
c Lund University, Sweden

A R T I C L E   I N F O

Keywords:
Sensor maintenance
Process control
Wastewater treatment plant
Measurement error
Condition-based maintenance
Interlaboratory test

A B S T R A C T

Sensors used for control have become widespread in water resources recovery facilities during the strive for resource efficient operations. However, their accuracy is reliant on uncertain laboratory measurements, which are used for calibration and, in turn, to correct for sensor drift. At the same time, current sensor calibration practices are lacking clear theoretical understanding of how measurement uncertainties impact the final control action. The effects of a customarily, and ad hoc, applied calibration threshold are unknown, leading to the current situation where many wastewater treatment processes are controlled by measurements with unknown accuracy. To study how sensor accuracy is affected by calibration, including varying calibration thresholds, we developed a simple theoretical model with closed-form expressions based on the variance and bias in sensor and laboratory measurements. The model was then simulated to yield the results, which showed no practical gain of minimizing laboratory measurement uncertainties during calibration, which otherwise directly impact operations. Due to these findings we strongly advice shifting mindset from considering calibration as a sequential detection and correction approach, towards an estimation approach, aiming to estimate bias magnitude and drift speed.

1. Introduction

Biased sensors operating in a feedback control loop have been shown to have negative impact on the treatment and resource efficiency in water resource recovery facilities (WRRFs) (Samuelsson et al., 2021). At the same time, on-line sensors and advanced control are technologies that enable optimized resource efficient operations (Stentoft et al., 2021). The challenge is to manage the increasing abundance of sensor data, and in particular, their accuracy. We here refer to accuracy as the combination of trueness (lack of systematic errors, bias) and precision (lack of random errors) in line with standardized nomenclature (International Organization for Standardization, 1994).

The evaluation of sensor measurement quality and accuracy is known as validation, which is commonly followed by calibration. These are the most important actions for reaching a high accuracy in practice, together with the mandatory sensor cleaning routines. The aim of this paper is to show how validation and calibration impact the sensor accuracy, and whether these actions can be tailored to achieve a desired accuracy.

Sensor calibration in WRRFs can be conducted by linear regression (Rieger et al., 2005). This is also a widely accepted method in general when the measurement used for validation (also denoted reference measurement) has a low precision, as compared to the sensor (BIPM, 2008). Alternative methods exist that consider precision in both sensor and reference (Orear, 1982), although they are scarce in practice. A drawback with the linear regression approach is that at least two reference samples are needed. Preferably, three or more samples are desired that should further be spread out in the full measurement range. Possibly due to these time-consuming requirements, a simpler approach has become standard in the WRRF domain, which is described next.

The current practice for sensor calibration is a two-step sequence, commonly executed on a fixed time interval. This approach applies to most sensors such as total suspended solids (TSS), ammonium, nitrate, and dissolved oxygen (DO). The first step – validation - is conducted by
comparing the sensor value with a reference measurement. In practice, the more common reference is a grab sample, e.g., for TSS, ammonium, and nitrate. It is possible to also use a standard solution with a known concentration if this is available (common for pH-sensors). This study focuses on evaluating current practices, and we here assume that the reference measurement is a grab sample. The validation step is intended to decide whether the sensor provided a ‘good enough’ measurement or needs calibration. What good enough means is subjective and application specific and is encoded in a so-called calibration threshold \( \gamma \) (henceforth referred to as the threshold). It is common to only calibrate the sensor when the difference during validation exceeds the threshold. The adjustment is then executed via the sensor signal processing unit, which we here assume to consist of a simple offset adjustment.

Despite the prevalent usage of thresholds, we have, to the best of our knowledge, not seen a theoretical explanation for how the threshold \( \gamma \) should be selected to produce a desired accuracy. On the contrary, we have seen a wide range of thresholds in practice, instead motivated by practical experience, ad hoc procedures, or by sensor supplier information (unpublished results by Andersson, 2019). A common suggestion is that a too tight threshold (or even omitting the threshold) would lead to excessive calibrations with increased variations in the measured signal, because of the more frequent calibrations. Thus, the threshold is not only intended to control the desired accuracy.

A second challenge with the current validation and calibration practice is that uncertainties in the reference measurement are not considered in a systematic manner. One or two grab laboratory (lab) samples are commonly used for validation and calibration. Lab samples contain both systematic and random errors through the sampling and analytical procedures (Eurachem, 2019). For example, a systematic error is present if the concentration always differs between two measurement points such as the surface of an activated sludge tank and 0.5 m below the surface. Similarly, influent variations causing dynamic variations in the tank concentration would induce a random sampling error if the sensor and lab sample are not taken simultaneously at the same location.

However, these errors are not quantified or considered during the validation and calibration procedure. Therefore, the final sensor accuracy after calibration is unclear. This leads to the situation where the treatment process is controlled based on measurements with unknown accuracy. This further lowers the incentive to improve existing calibration routines since any improvement would go unnoticed.

Ultimately, current calibration practices are lacking a clear theoretical understanding of how the lab uncertainties impact the final sensor accuracy, and whether a threshold can be applied to improve or control the accuracy. Therefore, we analysed calibration from both a practical and theoretical perspective to clarify what an optimal threshold value may be. Then, we analysed the impact from the uncertainties during calibration and how these are transferred to the sensor, as well as the process, when the sensor is used in a feedback control loop or for monitoring. More specifically, we developed a model describing the instant steps during calibration, which was then simulated and analysed theoretically. The calibration model was illustrated with a fictive total suspended solids sensor used to control a constant total suspended solids concentration in an activated sludge tank.

2. Materials and methods

A general discrete-time model for calibration is introduced in Section 2.2, which is illustrated by an example containing a TSS sensor measuring in an activated sludge tank (Section 2.1). The calibration model was then simulated and evaluated as described in Section 2.3 to assess the impact of threshold values on the sensor accuracy, and indirectly, the controlled TSS concentration.

2.1. Example system and limitations

The calibration model is illustrated with a TSS sensor used for monitoring and control of the TSS in the bioreactor of an activated sludge process. The results are also applicable to a general sensor-control system where sensor validation and calibration is conducted with uncertain reference measurements. A deliberate limitation is that we here only considered off-set calibration, which is sufficient to correct a constant error, i.e., an error that is independent of the measured value. However, for a sensitivity error (slope error) where the error magnitude depends on the measured value, two-, or multipoint calibration is needed (e.g. an ultraviolet light sensor). The implications of this limitation, with relation to the results, are discussed in Section 4.

It should be noted that available calibration methods differ among sensor makes. It is also common to have the choice of either conducting an off-set or slope calibration. This study is exemplified by a TSS-sensor calibrated with an offset, which resembles, e.g., a Ceric ITX sensor.

A list of the used model variables and related abbreviations is given in Table 1.

2.2. Sensor validation and calibration model

A model for the sequential steps of sensor validation and calibration is successively introduced in this section together with an overview of how the model was simulated (Section 2.2.10).

An important model feature is that one time step \( k \) in the model corresponds to one validation-calibration sequence (henceforth referred to as a calibration occasion). This is commonly executed at a fixed interval (weekly or monthly) although this is not a requirement for the method.

2.2.1. True concentration update

The true concentration \( x_t \) was assumed to be constant between two calibration occasions \( k \) and \( k+1 \), apart from an adjustment, \( \delta_k \),

\[
    x_t(k+1) = x_t(k) + \delta_k
    \tag{1}
\]

The adjustment \( \delta_k \) represents the effect of the control signal on the true concentration. The control signal can either be manual or automatically executed via a feedback controller and corresponds here to a change in amount of wasted activated sludge. For simplicity, other disturbances acting on the true sludge concentrations, such as variability in influent flow rate, concentrations, temperature, and operational changes in the return activated sludge flow were excluded. It is, however, straightforward to extend the model with such additional factors.

Table 1

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( k )</td>
<td>Discrete time index enumerating calibration occasions</td>
</tr>
<tr>
<td>( x_t )</td>
<td>True concentration</td>
</tr>
<tr>
<td>( x^* )</td>
<td>Reference measurement (incl. bias and noise)</td>
</tr>
<tr>
<td>( x_0 )</td>
<td>Reference measurement (excl. bias and noise)</td>
</tr>
<tr>
<td>( b_0 )</td>
<td>Bias in reference measurement</td>
</tr>
<tr>
<td>( v_n(k) )</td>
<td>Reference noise value at time ( k )</td>
</tr>
<tr>
<td>( R_v )</td>
<td>Variance of reference measurement</td>
</tr>
<tr>
<td>( R_s )</td>
<td>Total variance of sensor measurement</td>
</tr>
<tr>
<td>( R_{e_s} )</td>
<td>Variance of apparent sensor noise</td>
</tr>
<tr>
<td>( R_{r_e} )</td>
<td>Variance of sensor repeatability</td>
</tr>
<tr>
<td>( R_{r_{re}} )</td>
<td>Variance of sensor reproducibility</td>
</tr>
<tr>
<td>( x_c )</td>
<td>Sensor measurement (incl. bias and noise)</td>
</tr>
<tr>
<td>( x_s )</td>
<td>Sensor measurement (excl. bias and noise)</td>
</tr>
<tr>
<td>( x_{s_{cal}} )</td>
<td>Calibrated sensor value</td>
</tr>
<tr>
<td>( v_n(k) )</td>
<td>Sensor noise value at time ( k )</td>
</tr>
<tr>
<td>( \gamma )</td>
<td>Calibration threshold</td>
</tr>
<tr>
<td>( \delta_k )</td>
<td>Offset adjustment during calibration</td>
</tr>
<tr>
<td>( \delta_0 )</td>
<td>Change induced by the controller</td>
</tr>
<tr>
<td>( \delta_0 )</td>
<td>Validation difference</td>
</tr>
<tr>
<td>( sp )</td>
<td>Controller setpoint</td>
</tr>
</tbody>
</table>
2.2.2. Reference measurement

In general, the so-called accepted reference value is the most accurate reference value that can be used for comparison, i.e., sensor validation and calibration (International Organization for Standardization, 1994). Here, the concentration was assumed to be measured with one or several lab measurements in the media close to the sensor, i.e., the measurement 1994). Here, the concentration was assumed to be measured with one or several lab measurements in the media close to the sensor, i.e., the measurement

\[ x^*_k = x_k + b + v_k = x_k + v_k. \]

(2)

was modelled with a bias \( b \) and measurement noise \( v \), which is assumed to be Gaussian with zero mean and variance \( R_c \). Furthermore, \( x \) is the noise-free (but biased) reference measurement. The measurement noise is the sum of all random errors including sampling and analytical errors. Similarly, we assumed systematic errors due to sampling and analysis to be summed in the bias term \( b \).

2.2.3. Sensor measurement

The sensor measurement

\[ x^s_k = x_k + v_s(k). \]

(3)

was also modelled with a noise-free component \( x \) and white Gaussian noise \( v \) with variance \( R_s \). Note that the noise variance is measured with respect to \( k \), and not conventional time \( t \) in (3). Therefore, \( R_s \) represents the sensor variance between validation occasions.

The apparent measurement noise is the conventional measurement noise and is the variance when reading the sensor value used for the validation. Commonly, the sensor value during a validation is obtained by manually reading the sensor display during a short time interval.

Mathematically, the apparent sensor measurement noise, \( R_{a,s} = \left( \frac{x_s}{\sigma_r} \right)^2 \), is a function of the \( N \) sensor measurements executed to assess the sensor mean value during one validation occasion, where \( \sigma_r \) is the standard deviation of \( R_{a,s} \). Hence, \( R_{a,s} \) approaches zero when \( N \rightarrow \infty \).

In addition to the apparent sensor noise, random errors may add due to the so-called repeatability and reproducibility of the sensor validation. The repeatability is the minimum variance for repeated measurements at constant conditions. The reproducibility is the maximum variance of the precision when different sensor and operator individuals reproduce equivalent measurements in the same media. Therefore, the repeatability, \( R_{\text{rept}} \), sets the lower bound and the reproducibility, \( R_{\text{rept}} \), the upper bound for the added variance. Thus, the total variance, \( R_t \), lies between \( R_{a,s} + R_{\text{rept}} \leq R_t \leq R_{a,s} + R_{\text{rept}} \), depending on the sensor validation routines. For example, well-trained staff executing documented routines with a high reproducibility \( R_{\text{rept}} \) would yield lower random errors, as compared to ad hoc actions conducted without caring for the sensor’s condition.

The sensor was further assumed to get a bias \( b \), between each calibration occasion \( k \) and \( k + 1 \),

\[ x_{k+1} = x_k + b + x_{k+1} - x_k, \]

(4)

where \( x_{k+1} - x_k \) from (1) shows the change in true concentration due to the impact of the control signal. Note that when the true concentration is constant, i.e., \( x_{k+1} = x_k \), the sensor measurement is only changed by \( b \) between the calibration occasions.

To simplify, the bias \( b \) was assumed to reach the same value between each calibration occasion, although a time varying bias, \( b_t(k) \), is likely and could straightforwardly be included in the model.

2.2.4. Validation

During sensor validation, the noisy sensor measurement is compared with the reference measurement. The difference, \( \delta_s \), was identified as

\[ \delta_s(k) = x^s_k - x^*_k = x_k + b + v_s(k) - (x_k + v_k), \]

(5)

2.2.5. Calibration

Having identified \( \delta_s \), the sensor was adjusted with a constant, \( \delta_\alpha \), if the difference was larger than a user defined calibration threshold \( \gamma \),

\[ |\delta_s(k)| > \gamma, \delta_\alpha(k) = \delta_\alpha(k) \]

(6a)

\[ |\delta_s(k)| \leq \gamma, \delta_\alpha(k) = 0 \]

(6b)

The adjustment \( \delta_\alpha \) in (6a) is known as an off-set calibration and provides a calibrated sensor \( x_{\text{calib}}(k) \) as

\[ x_{\text{calib}}(k) = x_k + \delta_\alpha(k) \]

(7a)

Recall from (4) that between every calibration occasion, the sensor gets a bias in addition to the change in true concentration

\[ x_{k+1} = x_{\text{calib}}(k) + b + \delta_s(k) = x_k(k) + \delta_\alpha(k) + b_s(k) \]

(7b)

Eq. (7b) describes how the sensor signal changes when calibration is applied by including \( \delta_s(k) \), as compared to (4), in which the sensor is not calibrated but only validated (i.e. \( \delta_s(k) = 0 \)).

2.2.6. Sensor calibration in a feedback control loop

As described in (1), the true concentration was changed via the effect of a manual or automatic control signal, \( \delta_s \). For the situation with feedback control (e.g. using a PI-controller), the sensor’s accuracy will impact the true concentration via the controller, as described below.

First, the controller setpoint, \( sp \), was assumed to be met prior each calibration occasion. That is

\[ x_s(k) = sp \]

(8)

i.e., a perfect control without constraints on the control signal. Note that in (8) we used the noise-free measurement in (3) and assume that the sensor measurement noise does not influence the controller’s ability to reach the setpoint. This assumption is valid for PI-controllers when the sensor’s apparent measurement noise frequency is much larger than the controller’s speed (i.e., its integration time), which is the case for most controlled and slow wastewater treatment processes.

The main consequence of (8) is that \( \delta_s(k) \) will depend on the sensor drift and the adjustment during calibration

\[ \delta_s(k) = sp - (x_s(k) + \delta_\alpha(k) + b_s(k)) \]

(9)

This can be understood by recalling the sequence of made changes in the sensor during one calibration occasion. First, the sensor \( x_s(k) \) was adjusted during calibration with \( \delta_\alpha(k) \) as described by (6a–7a). Next, the sensor receives a bias (7b) and the controller compensates for the changes to fulfill (8), that is \( x_s(k+1) = sp \). Finally, the impact on the true concentration with a feedback controller is again described by (1), with \( \delta_s \) from (9).

For validation (omitting calibration), (9) was revised by simply setting \( \delta_\alpha(k) \) to zero according to (6b).

2.2.7. Validation and calibration errors during monitoring and feedback control

The final expressions for how calibration (7b) or only validation (4) impact sensor accuracy is given by (10) and (11), based on (1)–(9) with a complete derivation in Appendix A.

Sensor used for monitoring. When the sensor is only used for monitoring and not control, the validation and calibration errors were here defined as the difference between the true and measured concentration. Validation error (no calibration)

\[ x_s(k+1) - x^s_k(k+1) = x_s(k+1) - x_s(k+1) - v_s(k+1) \]

(10a)

Calibration error (calibration)

\[ x_s(k+1) - x^s_k(k+1) = -b_s - v_s(k+1) - b_s - v_s(k+1) \]

(10b)
Sensor used for feedback control. For a sensor operating in a feedback control loop, the validation and calibration errors are instead defined as the difference between the true and desired setpoint concentration, i.e., how the sensor error is transferred to the process.

Validation error (no calibration)

\[ x_t(k+1) - sp = x_t(k) - x_t(k) - b_t \]  \hspace{1cm} (11a)

Calibration error (calibration)

\[ x_t(k+1) - sp = b_t - v_t(k) + v_t(k) - b_t \]  \hspace{1cm} (11b)

Note that the only difference between (10) and (11) is the sensor noise \( v_t(k+1) \), which is part of the sensor monitoring example but not the feedback control example. This is a direct consequence of the assumption made in (8).

2.2.8. Calibration rule

The decision to calibrate or not to calibrate can be answered by assessing the error magnitudes in (10) and (11). When the error is larger during validation than for calibration, the sensor should be calibrated and vice versa. Interestingly, the same simple decision rule was obtained for both (10) and (11), namely

\[ |x_t(k) - x_t(k)| > |b_t - v_t(k) + v_t(k)|, \text{ Calibrate} \]

\[ |x_t(k) - x_t(k)| < |b_t - v_t(k) + v_t(k)|, \text{ Validate} \]  \hspace{1cm} (12)

That is, when the absolute sensor bias, \( |x_t(k) - x_t(k)| \), is larger than the bias in the reference measurement plus the noise realizations, calibration should be executed.

Taking expectation of the right side of (12) yields

\[ E[|b_t - v_t(k) + v_t(k)|] = E[b_t], \]  \hspace{1cm} (13)

which demonstrates the logic decision to only calibrate when the reference measurement is more accurate than the sensor. However, the expected value of the left side of (12) is not obvious and depend on the calibration threshold in (6). For this reason, the model was simulated as described in the next sections.

2.2.9. Motivation of model parameters

To the best of our knowledge, we have not found any studies or practical validation results that provide error estimates of the needed model parameters, that is, the bias and variance of TSS sensor and reference measurements. Therefore, these parameter values were assumed according to our reasoning in the next paragraphs.

Firstly, the reference measurement variance was assumed to be the sum of the analytical and sampling error variance for a lab measurement. A typical standardized analytical protocol specifies 10–15 % measurement uncertainty for TSS, which includes both systematic and random errors. The sampling errors are less studied, and the only example we found was Rossi et al. (2011), which indicates a 20–100 % sampling error, depending on flow conditions. Here, we selected a variance error for the reference measurement of 150 mg/L. This corresponds to a 10 % measurement uncertainty completely based on random errors, which is a deliberate underestimation of the true value. Based on the limited data about systematic proportion of sampling and analytical errors we used the same error value for both the reference bias, as for the variance.

Secondly, the sensor noise variance was assumed to be smaller than the reference measurement variance. As indicated in Section 2.2.3, the sensor variance can be reduced when performing a careful measurement, e.g., during constant conditions in a bucket with wastewater, and during an extended time. We selected the sensor noise variance to one third of the assumed of reference measurement variance.

Lastly, we considered two bias levels, 100 mg/L and 350 mg/L. These correspond to values that would be considered ‘small’ (requiring no action) and ‘large’ (needing a calibration) in practice.

Although this study uses the TSS sensor as example, the challenge to assess random and systematic errors is general for most water quality sensors.

2.2.10. Model simulations and software

The simulated actions during one calibration occasion, i.e., the simulated calibration model, are outlined in Table 3 with related parameter values in Table 2. The model was implemented in MATLAB version R2020 and is available as Supplementary materials.

2.3. Performance evaluation

How well \( \delta_g \) measures the true difference \( x_t - x_t \), i.e., whether a bias can be detected, was analysed with the receiver operating characteristics (ROC) and a two-sample Z-test. The improvement in accuracy by conducting several repeated lab measurements for the reference measurement (replicates) as well as for the sensor was analysed for one, three and five repeated samples. The ROC visualizes the detection and false alarm rates as a function of calibration threshold value (Kay, 1998).

The impact from calibration thresholds, on the true concentration, was analysed by simulating the calibration model for \( K = 1 \) 000 time steps and \( M = 10 \) 000 Monte Carlo repetitions with the settings in Table 2. The root mean squared error (RMSE)

\[ RMSE = \sqrt{\frac{1}{M} \sum_{m=1}^{M} \left( \frac{1}{N} \sum_{t=1}^{N} (x_t(k) - sp)^2 \right)}, \]  \hspace{1cm} (14a)

for a sensor used for control was then used to assess the accuracy and impact from calibration thresholds of the calibration and validation error in (11).

The RMSE was also analysed by its bias and variance proportions. In general, the mean squared error (MSE) is the sum of the squared bias and variance, see e.g. (Gustafsson, 2000)

\[ MSE(x) = E[x]^2 + E[x - E[x]]^2 \]  \hspace{1cm} (14b)

Further, the impact from calibration thresholds on the variation in true concentration \( x_t \) and controller correction \( k_t \) was evaluated by their respective standard deviations during the simulations. Finally, the expected value of (11) when \( \gamma = 0 \) was derived as reference to the numerical estimate in (13) obtained from the simulations.

Table 2

<table>
<thead>
<tr>
<th>Variable/parameter</th>
<th>Value</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>( sp )</td>
<td>3 000 mg/L</td>
<td>A common target TSS concentration in an activated sludge plant</td>
</tr>
<tr>
<td>( R_d )</td>
<td>(150 mg/L)</td>
<td>2</td>
</tr>
<tr>
<td>( R_s )</td>
<td>(50 mg/L)</td>
<td>2</td>
</tr>
<tr>
<td>( b_t )</td>
<td>150 mg/L</td>
<td>An assumed systematic error magnitude due to sampling and analysis with the same magnitude as the random error.</td>
</tr>
<tr>
<td>( b_t )</td>
<td>100 mg/L</td>
<td>An assumed drift between each calibration occasion. Two bias values were studied illustrating a small and large bias.</td>
</tr>
<tr>
<td>( 350 mg/L(k) )</td>
<td>An alternative calibration threshold in practice is between 10 and 20 % deviation of the reference measurement (300–600 mg/L).</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable/parameter</th>
<th>Value</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \gamma )</td>
<td>0–1000 mg/L</td>
<td>A common calibration threshold in practice is between 10 and 20 % deviation of the reference measurement (300–600 mg/L).</td>
</tr>
</tbody>
</table>
3. Results

The results are presented in the same order as calibration is executed in practice. That is, first, the bias detection performance during sensor validation is analysed (Section 3.1). Next, the selection of an optimal calibration threshold is studied (Section 3.2). Last, the impact from calibration thresholds on the process variations is analysed (Section 3.3).

3.1. Sensor bias detection during validation

The calibration rule in (12) requires that the difference $x_i(k) - x_j(k)$ can be quantified, i.e., that the sensor bias can be detected. This is in practice assessed via the validation and the difference $\delta_k(k) = x_i(k) - x_j(k)$ in (5). How well bias of 100 and 350 mg/L were detected during sensor validation with simulations and the settings in Table 2 is shown in Fig. 1(a) as the receiver operating characteristics.

In Fig. 1(a), a false alarm is when the difference during validation calls for a calibration, although there is no true difference, i.e., no bias. The opposite situation is denoted detection, i.e., when a bias is correctly detected during validation. The optimal detection performance is when the detection ratio is one and the false alarm rate is zero for all threshold values. The lowest acceptable performance is indicated by the diagonal grey line in Fig. 1(a), which is obtained when the decision to calibrate is decided randomly. It is possible to get an even lower performance if deliberately violating the goal of validation. However, as a benchmark, we should expect any validation method to be better than just flipping a coin.

The detection performance (high detection rate and low false alarm rate) increased with an increasing bias (Fig. 1(a)). Thus, it was easier to correctly detect a large bias, which is expected. Likewise, a lowered threshold value improved the detection rate, but at the cost of an increased false alarm rate. This is indicated by the arrow in Fig. 1(a). When the threshold was at its minimum, i.e., zero, the detection and false alarm rate matched the random detection performance (upper right corner in Fig. 1(a)). That is, both the maximum detection and false alarm rates were obtained, with equal proportions of detections as false alarms. Close to optimal detection performance was obtained when three replicate measurements were taken during validation for a 350 mg/L bias (thick black solid line in Fig. 1(a)).

An alternative to the ROC is to adopt a statistical test framework and evaluate sensor validation in the view of a statistical hypothesis test. In this regard, the detection rate is known as the test’s power and the false alarm rate equals the significance level (Kay, 1998). An appropriate test was to assess the absolute difference during validation like a two-sample Z-test. The null hypothesis (H0) was that: there is no difference between the sensor and reference measurement (no bias, or a too small bias). The alternative hypothesis (H1) was that a certain difference exists, e.g., that the bias equals 350 mg/L or above. Fig. 1(b) illustrates the probability distributions for such test with the same settings as used in Fig. 1(a), where a bias of 350 mg/L was assumed to be present when the alternative hypothesis is true.

Fig. 1(b) explains the difference in detection performance for the three black lines in Fig. 1(a). First, the power approaches one as the threshold is lowered to cover H1, see the arrow in Fig. 1(b). Next, if H0 and H1 are perfectly separated, e.g., due to an assumed large bias in H1 (>350 mg/L), optimal detection performance can be obtained with power one and zero significance level. Last, the overlap between H0 and H1 is also decreased if the variance of either hypothesis is decreased. This was the reason for the improved detection performance with three repeated measurements in Fig. 1(a). In general, the standard deviation of a Gaussian decreases with $\sigma/\sqrt{n}$ when replicate samples are taken, where $\sigma$ is the original standard deviation and $n$ is the number of replicate samples. Table 4 shows the dramatic improvement in power when three or more samples were used.

Ultimately, Fig. 1 and Table 4 indicates that a sufficiently large bias (>350 mg/L) or repeated measurements during validation are needed for a reliable detection (a power above 0.8 is commonly required in statistical tests) for the settings in Table 2.

3.2. Identifying the optimal calibration threshold

How the calibration threshold comes into play and impact the sensor accuracy and true concentration is now analysed by evaluating simulations with the settings in Table 2. Also, the impact from different threshold values with respect to the root mean squared error (RMSE) is analysed.

3.2.1. Simulating the calibration model

The output from simulating the calibration model with the sensor operating in a feedback control loop is shown in Fig. 2, which is intended to demonstrate that the calibration model produced logic results. The top part, Fig. 2(a), shows the that

- The sensor measurement varies around the setpoint regardless of the true concentration.

![Fig. 1. (a) Receiver operating characteristics (ROC) for detecting a bias during sensor validation with the settings in Table 2. The ROC is shown for two bias values (100 mg/L, dashed line and 350 mg/L solid black lines) and one set of measurements during validation (n = 1) and triplicate measurements (n = 3). The random detector (grey solid line) sets the lower performance bound. Threshold values (γ) that produce five percent false alarms and 95% detection rate follow the grey dashed lines and are indicated with white and black circles, respectively. The grey arrow indicates the improved detection rate when the threshold value is lowered. (b) Probability distributions corresponding to the black solid line in (a) with n = 1. The grey arrow indicates how a lowered threshold increases the power of when the alternative hypothesis is true with a bias of 350 mg/L (H1: bias of 350 mg/L, grey solid line), and the significance level of the null hypothesis (H0: no bias, black line). The dashed grey line shows the original Gaussian distribution for $\delta_k$, which is folded above zero due to the absolute difference in (6).]
• The reference measurement varies around the true concentration, but with an offset equal to its bias.
• The true concentration drops in line with the accumulated sensor drift.

Further, Fig. 2 (b) shows that

• The periods with no sensor adjustments or controller corrections (e.g., \( k = 3-6 \)) coincides with the consistent sensor drift in Fig. 2 (a).
• When the true concentration has dropped low enough so that the difference between the sensor and the reference is larger than the calibration threshold, the sensor is adjusted during calibration (e.g., \( k = 7 \) and \( k = 12 \)).

A complete set of figures describing the scenarios where the sensor was used for monitoring, and with/without bias and with/without noise in the sensor and reference measurements are provided as complement in the Supplementary Materials.

3.2.2. Root mean squared error for different threshold values

The impact of RMSE with different threshold values when the sensor was used for feedback control is shown in Fig. 3 (a,b). The RMSE increased with increasing threshold (Fig. 3 (a)). The RMSE consists of a bias and a variance part. These are however more easily analysed in terms of their relative proportions of the mean squared error (MSE) since the MSE is the sum of these two components, recall (14b). The bias proportion of the error can thus be obtained by computing the ratio between the bias and the MSE.

The bias proportion was constantly about 75 % of the MSE, regardless of the threshold value. This result indicates that the squared bias and variance errors increased equally much when the threshold was increased. i.e., that the variance proportion did not increase for a zero threshold value (when the sensor was biased).

However, the MSE bias proportion dramatically dropped for \( \gamma > 400 \) mg/L in Fig. 3 (b), which illustrates the same situation as in Fig. 3 (a) but with an unbiased sensor. This observation simply reflects that the unbiased sensor never will be calibrated if a sufficiently large threshold is applied (recall Table 3) and that the true concentration and the sensor was initialized to \( x_s = x_t = x_{sp} \). This was also the only situation when a lower RMSE was obtained with a calibration threshold larger than zero, i.e., to keep the sensor ‘as is’. When the sensor is perfect there is no gain in calibrating it with uncertain reference measurement (recall the calibration decision rule in (12)). Ultimately, the results in Fig. 3 (a, b) indicate that the lowest RMSE, and thus the highest accuracy, was obtained when calibration was consistently executed. In this respect, the optimal threshold value was zero when the impact from a biased sensor operating in feedback control should be minimized during calibration. The one exception (to the best of our knowledge) for when a calibration

![Fig. 2. Simulation of the calibration model with settings in Table 2 and a calibration threshold of 310 mg/L when the sensor is used for feedback control. (a) Measurements and the controller setpoint. (b) Sensor and controller adjustments and the sensor bias, i.e., the drift. Top and bottom figure share y-axis label.](image)

![Fig. 3. (a) Root mean squared error (RMSE) (black solid line) and the percentage of mean squared error (MSE) due to bias (grey dashed line) for calibration thresholds in the range 0–1 000 mg/L and the settings in Table 2. (b) Same as in (a) but without sensor bias. (c,d) Standard deviation (SD) of the true concentration (\( x_t \), black solid line) and controller induced correction (\( \delta_c \), grey solid line) with (c) and without bias (d).](image)
which reduce the variance, is also shown. The difference for varying number of replicate samples (n), measured as RMSE. The accuracy. More specifically, the RMSE decreased with decreasing variance. The deviation in true concentration and RMSE in Fig. 3 (a,b) was indicated the settings in Fig. 1 (thin black solid line in (a)).

The only exception where a threshold can decrease RMSE can be seen in Fig. 3(b), which applies for the case with an unbiased sensor. Applying γ > 800 mg/L indicates a RMSE equal to zero (Fig. 3(b)). This corresponds to a sufficiently large threshold where no calibrations are executed due to random variations in the sensor and reference measurements. Complementary computations (data not shown) showed a linear correlation between the total sensor and reference standard deviation and the threshold producing zero RMSE as $γ = 4.8/R + R^2$ (97.74 %). Thus, a rule-of-thumb, a threshold needs to be at least five time times larger than the standard deviation of the random errors if calibration-on-noise should be avoided. Note, for such large threshold, the detection rate is below the random detector performance (Fig. 1(a, b)) and incapable of detecting a bias of 350 mg/L.

### 3.2.3. Using a threshold to control the calibration error

The deviation in true concentration and RMSE in Fig. 3(a,b) was induced by the sensor and reference measurement errors as described by (11b) and denoted calibration error. Similarly, the calibration error for the sensor, i.e., its accuracy, was described in (10b). How the sensor accuracy was affected by different thresholds and replicate samples is shown in Table 5.

The sensor’s accuracy reflected the accuracy of the controlled true concentration (Table 5). This was expected since the model is based on similar expressions (compare (10) and (11)), where only the sensor noise differs. Further, Table 5 shows that an increasing threshold corresponds to an increased RMSE, which was also shown in Fig. 3. However, the threshold’s absolute value does not necessarily reflect the achieved accuracy. More specifically, the RMSE decreased with decreasing variance error (increasing number of samples), see Table 5. Further, the threshold was shown to only impact the false alarm rate (Fig. 1). Altogether, these results indicate that a threshold cannot control the sensor accuracy, although it will have a direct (negative) impact on the final sensor accuracy.

### 3.2.4. Expected values for zero threshold value

The RMSE for a zero threshold obtained from simulations can be verified by comparing the expected value for the calibration error in (10b) and (11b) with the RMSE value for $γ = 0$ mg/L in Fig. 3(a,b). As Hastie et al. (2009) describe, the standard expression for the mean squared error is $MSE(X) = E[X^2] + E[X]E[X]$, which can then be used to quantify the accuracy in (10b) and (11b) as

$$RMSE(x_t(k + 1) - x^*_t(k + 1)) = \sqrt{(b_n - b)^2 + R + 2R_n}$$  (15a)

$$RMSE(x_t(k + 1) - sp) = \sqrt{(b_n - b)^2 + R + 2R_n}$$  (15b)

where the bias proportion of the MSE is $(b_n - b)^2$ and the variance part is $R + 2R_n$ for the sensor accuracy and $R + R_n$ for the controlled concentration due to the assumption in (8). Note, that (15) provides a closed-form expression that quantifies the accuracy as a function of the bias and variance of sensor and reference samples, when no threshold is used.

### 3.3. Process variations due to sensor calibration

One hypothesized complication of using $γ = 0$ mg/L (mentioned in Section 1) was the potential variations in the true concentration or in the change induced by the control signal δc. Fig. 3(c,d) shows how the variance in δc and the true concentration were affected by different calibration thresholds in the range 0–1 000 mg/L. Similarly, as for the RMSE, the variance in both the true concentration and δc increased with increasing threshold values in Fig. 3(c). In the end, neither a decreased RMSE, nor an increased variance, was indicated as an issue when omitting the calibration threshold.

### 4. Discussion

#### 4.1. Targeting sensor bias detection or estimation?

Not surprisingly, the results showed that it is challenging to detect a
small bias when the reference measurement has a large uncertainty. The
detection rate could be increased by lowering the calibration threshold,
and optimally reached one when set to zero - at the expense of an
increased false alarm rate. Then, the detection performance approached
the random detector’s performance with 50 % false alarms. Half of the
executed calibrations would then be carried out, despite that the sensor
is unbiased, and hence, in no need for calibration. In practice, this would
be a poor strategy if calibration is time-consuming or costly. However,
for most water quality sensors, the preceding validation and reference
measurement sample is the more costly action. The calibration effort
only amounts to manually entering the validation data in the sensor’s
digital processing unit (and even this step has been automated in some
Swedish WRRFs). In the end, negligible negative effects can be assumed
from excessive calibrations in practice.

The logic calibration rule in (12)-(13) stated that calibration should
only be executed when the reference measurement is more accurate than
in the sensor. One such situation in practice when bias detection, as a
pre-step for calibration, could be motivated is when a sensor has been
calibrated to a very high accuracy, e.g., when directly supplied from the
supplier. One example is airflow rate sensors, which can be adjusted to
a high accuracy in factory (where a good reference sensor is available).
Then, calibrating the same sensor with on-site routines (lower accuracy)
would deteriorate the supplier calibration in line with Eq. (15) and the
results for the unbiased sensor (Fig. 3(b)). If instead a calibration
threshold value is used, the sensor should only be calibrated when it has
lost its supplier provided accuracy and obtained a sufficiently large bias.
Then, the rule-of-thumb in Section 3.2.3 could be used to select a suf
ficiently large threshold. Note, however, that a large bias is required for
a reliable detection when the reference measurement has a low accuracy
(Fig. 1). Thus, it would be a gamble (and a guess) to know when the
sensor has lost enough accuracy to gain from on-site calibration.

Disregarding the unlikely situation in the previous paragraph, the
results showed that both the highest accuracy as well as the smallest
variation, in true concentration were obtained when no threshold was
applied, i.e., when calibration always was executed. This contrasts the
common perception that a calibration threshold can be used to somehow
improve the accuracy in the sensor, or reduce the variations induced
during calibration.

It is unclear why the usage of calibration thresholds has reached wide
acceptance in the wastewater domain. One reason could be saving time
if validation is quicker than calibration. Another reason could be that it
is a consequence of considering validation as a bias detection test and
assess whether calibration is needed, i.e., a detection problem. In this
regard, statistical hypothesis testing is a widely taught method that
produces a threshold via the decided significance level. However, a
statistical test is commonly executed for more than one observation
when the mean and variances are estimated from data. This can cause
confusion for how such a test should be executed for one sample (which
is common practice during validation) and may be one reason for the
lack of documented procedures for choosing the optimal threshold.
The results showed that the significance level and the false alarm rate could
be controlled by the threshold, but not the detection rate or the sensor
accuracy.

An alternative view is to instead consider validation and calibration
as an estimation problem. This is logical since most water quality sensors
are indirect measurements and rely on calibration (estimation) via
reference measurements. In such estimation viewpoint, all efforts are
focused around estimating the bias and maximizing the sensor accuracy
for the given data, rather than worrying about whether the sensor is
sufficiently biased or not. By continuously using all available data for
calibrating the sensor, the lowest RMSE is obtained on average (Fig. 3),
while accepting that variations in the adjustments in Fig. 2(b) are a
natural consequence from the uncertain reference measurement and
calibration procedure.

It also becomes clear in the estimation viewpoint why thresholds can
be problematic. For example, excluding part of the measurements in a
dataset used during estimation by applying a threshold would violate
the best estimation practice and bias the training dataset. It is only
customary to remove outliers. This is the opposite effect to what a
threshold induces. For example, when applying a 310 mg/L threshold,
only values and outliers outside this limit are used for calibration. The
reason for these outliers is more likely caused by inaccurate reference
measurement and not a large sensor bias. Also, reference measurements
close to the sensor’s (within the threshold range) are not used for cali
bration and effectively wastes part of the available data for estimation.

In the end, we could ask what the main goal with sensor validation
and calibration is. Do we want to know whether the sensor has a bias
(yes/no), or do we want to produce a sensor measurement as accurately
as possible? We recommend the latter and to shift our mindset towards
regarding sensor validation and calibration as an estimation problem
rather than a detection problem.

4.2. Optimizing the practical trade-off in accuracy

When regarding calibration as an estimation problem, a natural
follow-up question is what the optimal trade-off is with regards to
achieved accuracy for spent resources. The (15) gave a closed-form expression for the final accuracy,
depending on the noise and bias errors in the sensor and the reference
measurement. First, the equation is a good tool to assure that the desired
accuracy is reached. For example, a previous simulation study showed
that the bias above 500 mg/L in a total suspended solids sensor was
problematic for the nitrogen removal efficiency (Samuelsson et al.,
2021). The equation can be used to decide the maximum acceptable bias
and variance in the sensor and reference measurement to maintain the
desired treatment efficiency. Cost can be traded for accuracy during
calibration, and thereby identify the minimum needed effort. For
example, the improved accuracy from triplicate reference samples can
be compared to the increased analysis cost, or an alternative action. The
improvements for revised calibration strategies can be analysed through
the bias and variance components in (15). A good example is the
frequently debated ‘bucket approach’, and whether it is worth the efforts
during calibration.

The bucket approach refers to when the sensor validation and cali
bration measurements are taken in a bucket filled with wastewater,
instead of directly from the process. When analysing the bucket
approach from (15), we can see that it improves the accuracy in several
ways. First, the sensor’s apparent noise can be reduced to zero if the
wastewater concentration is constant and the measurement is conducted
sufficiently long (recall the sensor measurement Eq. (3)). Second, the lab
sampling error decreases, as compared to when a grab sample is taken in
the process. As a reference, the sampling error in sewers is indicated to
be between 20 % (good turbulent sampling conditions) and up to 100%
(laminar flow) in (Rossi et al., 2011). Last, note that a reduced sampling
error impacts both the bias and variance in the reference measurement,
when the sampling error contributes to both $b$ and $R$. However,
remaining key challenge is to quantify the sampling er
ors, as well as the bias and variance proportions in the analytical errors.
Without good error estimates, also the estimated accuracy in (15) be
comes unreliable. Therefore, we recommend both on-site and academic
research efforts to bridge the knowledge gap in lab sample error quan
tification. The following actions are recommended

(1) Routinely execute interlaboratory assessments (‘round-robin’) to
identify systematic and random errors in the analytical chain.
(2) Quantify the sensor validation variance $R$, in practice via the
reproducibility and repeatability for the used validation
approach.
(3) Assess the variation in different sampling methods, using, e.g.,
the variogram approach as described in (Petersen and Esbensen,
2005).
Differentiate between systematic and random errors in the analyses when assessing the lab coverage factor (confidence interval). This may require new uncertainty quantification methods.

4.3. Aiming towards a best sensor calibration practice

An assumption in the calibration model was that the same bias was obtained at each calibration occasion. This reflects the ideal situation when calibration is executed just-in-time to maintain a desired accuracy. A challenge, which was not considered here, amounts to estimate the drift speed. In practice, the sensor drift varies for different reasons, although only few studies have made detailed assessments about the exact relationships (Cecconi and Rosso, 2021; Ohmura et al., 2019; Samuelsson et al., 2018). Further, the calibration interval \( k \) is commonly fixed, leading to different bias values for a fixed time, depending on the drift speed. Thus, assessing the time to reach a certain bias is critical in addition to the actual bias quantification. A good thing is that the drift speed can be estimated from historic calibration data when including validation measurements information about how the sensor signal was adjusted during calibration. Tracking these so-called metadata are essential for assessing the sensor’s condition. Today, routines are unfortunately lacking in practice on how to fully exploit the values in metadata, which could, when used, be a game-changer for best sensor calibration practices. The importance of this topic is emphasized by the on-going task group on Metadata collection and organization (MetaCO, supported by International Water Association [IWA]).

Assuming an ideal situation with quantified lab sample uncertainties, metadata governance systems in place, and extending the results and discussions in this study, we can sketch a simple six-step-recipe best calibration practice for mitigating sensor drift with off-set calibrations:

1. Decide the needed sensor accuracy using process expertise or simulations.
2. Assess and quantify the errors in (15).
3. Decide the most cost-effective way to achieve the needed accuracy in step 1, e.g., by replicate samples during calibration or tailored sampling approaches.
4. Execute calibrations at an initially fixed interval, e.g., monthly, and record the estimated sensor drift speed during different seasons.
5. Extrapolate the drift speed and identify the calibration interval corresponding to the maximum accepted bias and desired accuracy.
6. Iterate step 4–5 and continuously improve the calibration routine while evaluating the metadata produced during calibration.

The calibration practice could also be adopted for a slope calibration if one expects the bias to vary with varying concentrations. This would require adjustments of step 1–2 and step 5 in how the bias and variance are computed. The bias may change depending on the considered concentration.

Regardless of calibration method, the suggested calibration strategy aims at correcting slow and minor sensor drift, which will likely result in extended calibration intervals.

As a complement to the low frequency calibration, we suggest using data-driven methods such as the software sensor-based approaches reviewed in (Hisimi et al., 2013), for high frequency (real-time) monitoring and to detect sudden large bias.

4.4. Limitations and future work on the calibration model

The calibration model assumes an offset calibration, which is also used for the assumed TSS-sensor make. Such calibration is appropriate for an offset error, i.e., an offset bias. When the bias is instead caused by a slope error or a nonlinearity in the sensor, an offset correction will only fully correct the bias at the specific concentration where the calibration is executed. This is true when the controller is perfect and maintains the concentration at the setpoint, which applies for the studied model. However, in practice, the controller will not be perfect, and an offset calibration will be inappropriate for a slope error. Further, the bias direction may shift between positive and negative direction, as well as with different magnitudes, e.g., depending on season. Altogether, the effects of these realistic issues need to be further analysed, especially during the time between the calibration occasions that were not considered here. For such analysis, dynamic simulations are desired. It is likely that these issues will induce deviations in the desired true concentration, and potentially also create larger variations in the control signal. Still, we hypothesize that such negative effects won’t be reduced by applying a calibration threshold in the light of the presented results in this study. Rather, executing the correct mitigating action (such as a slope calibration for a slope error) should be strived for. Such informed decision is reliant upon accurate validation data and metadata as described in the previous sections.

Note that (15) is generally applicable regardless of calibration strategy. Here, an offset calibration was assumed, but also a slope calibration for a slope error would give the same results - as long as the bias is corrected during the calibration. This is because we here evaluated the RMSE at the calibration instants. However, as a complement, it would be interesting to also study how the RMSE differs when the time between calibration occasions is considered. Then, the effects of using an appropriate calibration strategy could be studied (in a dynamic model) where the time to attain the ideal setpoint in (8) would impact the results.

An additional, and timely, study would analyse how the errors in reference measurements and the inherited sensor calibration errors, are transferred during model calibration. Assessing model accuracy is becoming a burning question when real-time process models gaining wide-spread usage, popularly denoted digital-twins. There are similarities between model and sensor calibration, but more reference measurements and constraints during model calibration complicates assessment of the final model accuracy. It would therefore be interesting to see whether the findings here could be extended and revised to also benefit model calibration.

5. Conclusion

Uncertainties in the lab measurements used for sensor calibration induced errors in the sensor as well as the process controlled by a controller with the very same sensor. This calibration error could not be mitigated by using a calibration threshold, which on the contrary, decreased the sensor accuracy for large threshold values. Thus, the best choice to reach maximum sensor accuracy was to violate current practices and skip the calibration threshold, and instead consistently execute calibration. The only realistic situation that motivated a threshold was when the needed time for calibration was large, as compared to validation. Setting the calibration threshold to zero did neither increase variations in the process nor in the control signals, which has been a common concern.

Further, the sensor’s calibration error was transferred to the process as a similar deviation between the setpoint and true concentration (when controlled by a feedback control loop). The transfer of calibration errors was seen in simulations as well as demonstrated analytically. The expected sensor accuracy and the errors transferred to the process were quantified via a closed-form expression when the bias and variance are known for the sensor and reference measurement. This provides a theoretical tool to assess and improve calibration strategies in practice.

Finally, the bias detection rate during validation was low due to the uncertainties in the lab measurement. More knowledge about the proportions of random and systematic sampling and measurement errors are needed to optimize both validation and calibration procedures.

Because of the findings, we argue that the current calibration viewpoint needs to shift focus from the existing sequential detection and
correction approach, towards an estimation approach with the goal to instead estimate bias magnitude and drift speed. In this respect, methods for automatically quantifying the apparent measurement errors based on metadata will be central for adopting best calibration practices, which, in effect, are needed to optimize resource efficient operations.

Declaration of Competing Interest
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability
The model used for producing the simulation and related data is available as supplementary materials.

Supplementary materials
Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.watres.2022.119338.

Appendix A: derivation of validation and calibration errors

Derivations of the final four expressions Eqs. (10a) and (10b) and Eqs. (11a) and (11b) are given here using the same equation numbers as in Section 2.

A) Validation of sensor used for monitoring by combining (1)–(4)
\[
x_i(k+1) - x'_i(k+1) = x_i(k) + \delta_i(k) - (x_i(k) + v_i(k+1)) = x_i(k) - v_i(k+1)
\] (10a)

B) Calibration of sensor used for monitoring by combining (1)–(7)
\[
x_i(k+1) - x'_i(k+1) = x_i(k) + \delta_i(k) - (x_i(k) + b_i + v_i(k+1)) = x_i(k) + \delta_i(k) + v_i(k+1) = -b_i - v_i(k) - b_i - v_i(k+1)
\] (10b)

When the same sensor is used for control, the error in the true concentration as a consequence of the sensor accuracy is

C) Validation of sensor used for control
\[
x_i(k+1) - sp = x_i(k) + \delta_i(k) - sp = x_i(k) + (sp - x_i(k) - b_i) - sp = x_i(k) + sp - x_i(k) - b_i - sp = x_i(k) - x_i(k) - b_i
\] (11a)

D) Calibration of sensor used for control
\[
x_i(k+1) - sp = x_i(k) + \delta_i(k) - sp = x_i(k) + (sp - x_i(k) - b_i) + \delta_i(k) - sp = x_i(k) + sp - x_i(k) - b_i + v_i(k) + x_i(k) - b_i - sp = -b_i - v_i(k) - b_i + v_i(k)
\] (11b)

References


