Several small or single large? Quantifying the catchment-wide performance of on-site wastewater treatment plants with inaccurate sensors

Schneider, Mariane Yvonne[a, c]; Harada, Hidenori[d]; Villez, Kris[e]; Maurer, Max[b,c]

[a] Next Generation Artificial Intelligence Research Center & School of Information Science and Technology, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan.
[b] Eawag, Swiss Federal Institute of Aquatic Science and Technology, 8600 Dübendorf, Switzerland.
[c] Institute of Civil, Environmental and Geomatic Engineering, ETH Zürich, 8093 Zurich, Switzerland.
[d] Graduate School of Asian and African Area Studies, Kyoto University, Yoshida-Shimoadachi, Sakyo, Kyoto 606-8501, Japan
[e] Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA.

**Contact Address: Mariane Schneider,* Next Generation Artificial Intelligence Research Center & School of Information Science and Technology, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan.
E-mail: myschneider@isi.imi.i.u-tokyo.ac.jp (M. Y. Schneider)

This document is the accepted manuscript version of the following article:
Schneider, M. Y., Harada, H., Villez, K., & Maurer, M. (2023). Several small or single large? Quantifying the catchment-wide performance of on-site wastewater treatment plants with inaccurate sensors. Environmental Science and Technology.
https://doi.org/10.1021/acs.est.2c05945

Abstract

On-site wastewater treatment plants (OSTs) often lack monitoring, resulting in unreliable treatment performance. They thus appear to be a stopgap solution despite their potential contribution to circular water management. Low-maintenance but inaccurate soft sensors are emerging that address this concern. However, how their inaccuracy impacts the catchment-wide treatment performance of a system of many OSTs has not been quantified. We develop a stochastic model to estimate catchment-wide OST performances with a Monte Carlos simulation.

In our study, soft sensors with a 70% accuracy improved the treatment performance from 66% of time functional to 98%. Soft sensors optimized for specificity, indicating the true negative rate, improve the system performance, while sensors optimized for sensitivity, indicating the true positive rate, quantify the treatment performance more accurately. This new insight leads us to suggest programming two soft sensors in practical settings with the same hardware sensor data as input: one soft sensor geared to high specificity for maintenance scheduling and one geared to high sensitivity for performance quantification. Our findings suggest that a maintenance strategy combining inaccurate sensors with appropriate alarm management can vastly improve the mean catchment-wide treatment performance of a system of OSTs.

Short synopsis statement: We take a systems perspective to show that a maintenance strategy incorporating inaccurate soft sensors is useful for improving the treatment performance of many on-site wastewater treatment plants in a catchment.

Keywords: parameter optimization, sensor-based maintenance, soft sensors, WRRF

1 Introduction

Water is essential to growing cities, and dwindling resources require a substantial increase in water efficiency. One option is to deploy modular systems of many on-site wastewater treatment plants (OSTs) to increase local reuse potential.[1–3] The term OSTs refers here to small plants, typically at household level. The term systems refers throughout this article to the wastewater treatment solution for an entire catchment. This may be either a large number of OSTs in a modular approach or one central wastewater treatment plant (WWTP), also called a water resource recovery facility (WRRF). The aim is to enable locally appropriate recovery solutions.[4] Such a paradigm shift from network-based wastewater management towards more modular or hybrid systems is urgently needed, as noted by Larsen et al.[5] Increasing the flexibility of highly networked and investment-intensive infrastructure will increase planning and operating flexibility[6] and is another reason to study systems of OSTs. However, the OSTs' spatial distribution is expected to lead to lower performance than centralized approaches. Therefore, no consensus yet exists on what degree of decentralization is optimal.

The question about the optimal degree of decentralization also referred to as "single large or several small," appears in many areas, including ecology[7] and the provision of infrastructure services such as electricity, heat, and, as in this case, wastewater.[8] For systems of OSTs, "several small" means a large number of wastewater treatment plants spatially distributed in a catchment. Importantly, comparisons must be made at catchment-wide level and not at individual unit level.

Measuring the units' performance enables the long-term overall treatment performance to be quantified. However, it also has a recursive effect on performance by allowing targeted maintenance strategies. Here, we focus on sensor-measurement-based maintenance, also called sensor-driven maintenance. Our key hypothesis is that a sensor-based maintenance strategy can enable systems of OSTs to achieve verifiable performance levels that guarantee environmental and human health.

To our knowledge, online quality monitoring is not commonly practiced, potentially leading to unnoticed failures.[9,10] Instead, maintenance takes place periodically. This time-driven maintenance scheme without online monitoring is currently the standard management strategy for OSTs in real-world practice.[11–14] Typical intervals are four times per year in Australia,[12,15] two to three times per year in Germany,[16–18] usually three or more times per year in Japan,[13,19] and at least once per year in Switzerland.[11] therefore one to four interventions per year currently seem a feasible range of interventions for OSTs. With one to four grab samples per OST per year during maintenance as observed performance, the true treatment performance of an OST cannot be estimated sufficiently accurately, creating a risk of undetected failures.

Few measurement campaigns have assessed the performance of OSTs by analyzing grab samples,[20–23] and even fewer use online sensors for close-to-continuous measurements.[24–26] Thus, information that characterizes performance, OST reliability, and failure rates remains sparse, despite calls for centralized remote monitoring and control of OSTs as early as 1998.[27] Therefore, a decisive factor for the successful implementation of OSTs is an online monitoring concept.[28–30]

Traditional wastewater monitoring emphasizes sensor accuracy with a correspondingly high frequency of maintenance, typically once a week. The equivalent maintenance effort for a large number of OSTs is so high as to render monitoring infeasible. Olsson (2013)[31] stated that OSTs need less accurate and therefore potentially less frequently maintained sensors than centralized WWTPs but did not quantify these requirements. Hug and Maurer (2012)[10] showed that inaccurate sensors could lead to a discrepancy between true performance and the performance reported by the sensors, subsequently referred to as observed performance.

In previous work, Schneider et al.[32] suggested that hardware sensors combined with software models, together termed soft sensors, could predict target variables from signals from unmaintained physical sensors. These soft sensors solve the issues of labor-intensive maintenance and the costs of measurement accuracy. Therefore, they can potentially resolve the performance issue[9] of OSTs by offering real-time data on their status. Soft-sensor accuracy, calculated as the number of correct predictions divided by the number of predictions, has been identified as 0.83-0.85 for four pH sensors and 0.80 and 0.85 for two dissolved oxygen sensors.[33] However, it is unclear whether the inaccurately measured performance suffices to quantify the true performance of a large number of OSTs and whether a maintenance strategy is feasible despite inaccurate measurements.

The work reported in this article spans this gap by comparing a monitoring concept for the observation of such inaccurate soft sensors with the true treatment performance of all OSTs in a catchment and investigating how well this true performance can be quantified despite the inaccuracy of the sensors. To do this, we developed a novel stochastic model. This model is used to explore the links between sensor-based monitoring, intervention intensity, and OST reliability in a Monte Carlo simulation. We used the insights gained in our previous studies[32,33] to specify the parameters of the model but did not limit the model to the previous studies. The following questions are addressed with this model:

Q1: How do soft-sensor accuracy, reliability of OST, and alarm management influence the true treatment performance and the number of interventions required?

Q2: How much can the performance of a large number of OSTs in a catchment be improved by online monitoring?

Q3: How does optimizing the parameters of the soft sensor influence the treatment performance and its quantification?

2 Methods

2.1 Model description and data

Where feasible, we used data from a previous study[32,33] to develop the stochastic model. The sensors were for measuring pH and dissolved oxygen, were commercially available, and predicted one target variable through feature engineering: the ammonium effluent concentration. This ammonium effluent concentration was assumed to represent the overall treatment performance, and therefore grab samples were analyzed for ammonium and served as labels for parameter optimization. It is essential to understand two characteristics of the previous study: (i) To serve a large number of OSTs, Schneider et al. (2019) left the sensors unmaintained, leading to inaccurate measurement and prediction of system performance. (ii) Schneider et al. (2019) used a binary classification to predict treatment performance. The two classes predict the completeness of ammonium oxidation in a biological process in a sequencing batch reactor (SBR). The binary states were defined as ammonium effluent concentrations below and above 1 $g_N m^{-3}$, which are here termed the "up" and "down" states respectively.

2.1.1 True and observed performance

The soft sensor's inaccuracy causes a discrepancy between the OSTs' true and observed performance. Therefore, we separated the model into two parts: true performance and observed performance (see Figure 1). Eq. 1 is the conceptual definition of the two performances:

$$true\ performance = observed\ performance + inaccuracy \quad (1)$$

Such inaccuracy can be caused by a lack of maintenance, as is the case for the soft sensors tested, or a complete lack of observations. The observed performance is the performance that human operators can estimate based on observations such as sensor measurements and inspections. The true performance is what an error-free sensor would measure. In the real world, the true performance can only be estimated from the observed performance. The model developed here is based on the assumption that the true performance has previously been modeled, and we quantify how well the observation with inaccurate sensors represents this true performance.

2.1.2 Model modules

The model consists of five modules (see Figure 1). The unit failure module (I) defines when an OST is operating normally and when a fault or failure occurs. The output of the unit failure module is the OST state. Typical failures we observed were clogging of the effluent, pump failure, and removal of all sludge from the plant (see supporting information S6 of Schneider et al. 2020[32]). The true performance module (II) uses the unit failure state as an input to quantify the true performance. In the previous study, this information was unknown too but was approximated with the measured ammonium concentration. The monitoring module (III) represents all observations available from both soft sensors and human inspection. The observed performance is determined by using the modeled true performance as input and considering the sensor accuracy. In this article, the observed performance is the result of the binary classification based on the measured signal with the unmaintained pH and dissolved oxygen sensors. The output of the monitoring module is used to quantify the observed treatment performance in a catchment. The alarm management module (IV) takes the observed performance as input to decide which OSTs are flagged for intervention. In the previous study, no alarm management was implemented. The intervention module (V) defines the frequency and types of intervention.

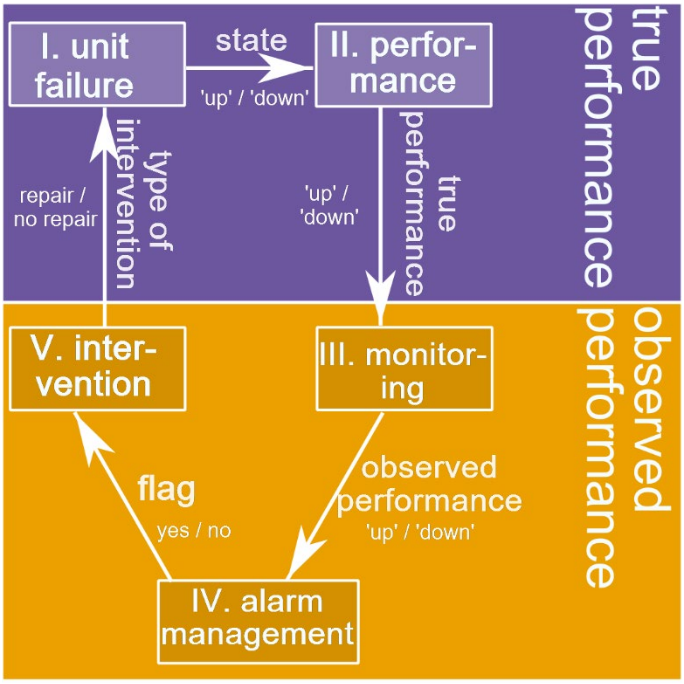

Figure 1: Systematic representation of the treatment performance model of one OST. I. and II. represent the true treatment performance of one OST. Therefore, they represent the actual OST. III., IV., and V. the observable performance by sensors and inspection. The arrows indicate the input and output flow from one module to another. I. is the failure of the OST; a failure or drift of the sensor is included in the monitoring module. The bold labels on the arrows are the type of input from one module to another; the small labels are the actual binary input from one module to another. In the Python scripts, the input from one module to another is implemented as 0 and 1. To estimate the performance of a system of OSTs, a large number of such units are modeled.

2.2 Model implementation

To keep computational costs low and increase usability, we implemented the model in the simplest way possible to address the questions presented in the introduction. The implementation below is the specific implementation based on the previous studies.[32,33]

At every time step, the model passes through all modules in Figure 1 in the order indicated by the arrows. The first simulation starts with a newly constructed OST (unit age $t = 0$ days) in the failure module and then passes through all the modules in one time step. A time step is defined as one SBR cycle, which we chose to be one day for simplicity, but ultimately is the frequency at which new soft-sensor-based observations become available. The model is implemented in Python 3,[34] is available under a GNU common license (Schneider[35] on https://gitlab.com/Viraja/systemperformance/), and is based on Hug and Maurer (2012).[10]

Failure module (I): The failure module defines whether an OST is fully functioning, here called the up state, or in a failure situation, the down state. The failure module is a function of the unit age, and for this article, we use the same binary state as in Schneider et al. (2019).[33] Therefore, up, implemented into the Python code as 1, means all ammonium is oxidized, down, implemented as 0, means not all ammonium is oxidized. We implemented a failure situation as an irreversible event with a high probability of insufficient performance that is only resolved by an intervention: sensor-driven maintenance. This means that within the failure module, an up state can change to a down state but not vice versa. To go from down to up, an intervention by the intervention module is required.

We used reliability theory, assuming that the failures are stochastic,[36] and chose a Weibull-distributed failure rate.[37] The Weibull distribution is often used in infrastructure management. A random number between 0 and 1 was drawn from a uniform distribution to check whether an OST that is up fails in the current time step. This number was compared with the Weibull distributed hazard rate in Eq. 2. If the random number is smaller than the hazard rate, the OST state is changed to down. If the random number is greater than or equal to the hazard rate, the state stays unchanged.

$$hazard\ rate(t, k, \lambda) = \frac{k}{\lambda}\left(\frac{t}{\lambda}\right)^{k-1}. \qquad (2)$$

Unit age is represented with *t* in days, *λ* is the scale factor in days, and *k* is the dimensionless shape parameter. Thus, *λ* and *k* define the Weibull distribution. For our screening, we kept the shape parameter *k* constant at 2, as in a previous study.[10] We varied the scale factor *λ*, which stands for the mean expected survival time in the Weibull distribution, in a sensitivity analysis because the performance is sensitive to *λ* and we currently lack sufficient data to estimate *λ.* The assumption was that OSTs fail on average every four months and six years (88 to 2216 days). Other failure modes, such as periods without inflow, defective installations, or contamination with harmful substances, probably follow other distributions and would need the implementation of additional failure probability functions.

True performance module (II): In the current implementation, the failure module's up state leads to an up state in the true performance module. Likewise, the failure module's down state translates into a down performance.

Monitoring module (III): Analogously to the failure and true performance module, the sensor-based monitoring module also records a binary state based on the information from the true performance module with a defined accuracy between 0.5 and 1.0. The lower bound is 0.5 because if, instead of measuring, a random state were drawn from the same distribution as the performance, this random soft sensor would achieve a 0.5 accuracy; one with an accuracy of 1.0 represents a perfect soft sensor. The following sensing outcomes are possible for the individual OSTs:

- True positive (TP): The soft sensor indicates up, and the true performance is up.
- False negative (FN): The soft sensor indicates down, and the true performance is up. The probability for the FN outcome is 1 minus the TP. The FN potentially causes false alarms.
- True negative (TN): The soft sensor indicates down, and the true performance is down.
- False positive (FP): The soft sensor indicates up, and the true performance is down. The probability for this outcome is 1 minus the TN. The FP represents undetected failures.

The sensitivity (Eq. 3) is used when the true performance is up and the specificity (Eq. 4) when the true performance is down to determine whether the soft sensor makes a true or false prediction. This means TP is the sum of all true positive predictions of the individual OSTs over all time steps, and the same is true for FN, TN, and FP.

$$sensitivity = \frac{\sum TP}{\sum FN + \sum TP} \tag{3}$$

$$specificity = \frac{\sum TN}{\sum TN + \sum FP} \tag{4}$$

A specificity of 0.98 and a sensitivity of 0.71 were identified through experimental validation under realistic conditions for the pH-based soft sensor.

Furthermore, we assumed that a complete soft-sensor failure either is noticed immediately or does not occur, based on our experience[33,38] where more than ten commercial pH sensors were tested, this is a valid assumption.

The *alarm management module (IV)* converts the input from the monitoring module into a flag for OST, where action needs to be taken. In the previous study, no alarm management was implemented. Therefore, we used the model to test various alarm management strategies. The most straightforward approach is to flag a unit as needing an intervention as soon as a soft sensor identifies the performance as down. We call this alarm management strategy (a). An approach to improving confidence is to wait for *n* consecutive cycles labeled *down* before flagging an OST for an intervention: strategy (b) where n = 4 and strategy (c) where n = 14. A more elaborate method, strategy (d), uses previously identified soft-sensor accuracy to calculate the probability of the unit being down from the signal of the current and previous steps. As soon as this probability has surpassed a threshold (we used 98% as a boundary close to100%), the unit is flagged for intervention, and an intervention is issued for the unit. To estimate the probability of the unit being down, we calculate the conditional probability with a sequential probability ratio test:[39]

$$P_t(perf = down|signal = negative) = \frac{P_{t-1}(perf=down)\cdot TN}{P_{t-1}(perf=down)\cdot TN+(1-P_{t-1}(perf=down))\cdot(1-TP)}, \quad (5)$$

and correspondingly,

$$P_t(perf = down|signal = positive) = \frac{P_{t-1}(perf=down)\cdot(1-TN)}{P_{t-1}(perf=down)\cdot(1-TN)+(1-P_{t-1}(perf=down))\cdot TP}. \quad (6)$$

$P(perf_{time} = down|signal_{time} = negative)$ represents the probability that the true performance of an OST is down given that the soft sensor shows a failure (TN or FN) at time *t* based on the probability of a down state in the previous time step (*t* - 1). The probability that an OST is down is updated for every time step using Eq.s 5 and 6. We limit the probability interval to [0.001, 0.999] with a simple stopping rule to avoid numerical issues and a memory effect.

Intervention module (V): The intervention is implemented so that an intervention is triggered every time a flag is raised in the alarm module. To minimize model complexity, we do not distinguish between different types of interventions nor consider spatial efficiencies. The assumption is that the intervention always repairs the OST. If the unit state is down and an intervention is issued, the true performance will be reset to up. Note that an intervention executed when the state is up does not impact the unit performance or the sensor accuracy. Clearly, such interventions only raise the maintenance cost without any benefit. In addition to sensor-driven interventions, a time-driven inspection is implemented. The time-driven inspection reveals the true performance of an OST in the time step during which the inspection is performed and repairs a unit with down status in the same manner as a sensor-driven intervention. The model can model sensor-driven or time-driven interventions or a combination of the two.

2.2.1 Mean performance in a catchment

To answer Q2, we modeled a system with 10 000 OSTs for ten years with individual reliabilities: a random uniform distribution of λ between 400 and 2200 days. The catchment-wide mean true performance is calculated from the true performance of the individual OSTs in the true performance module. Correspondingly, the observed performance is calculated from the output of the monitoring module. The overall true performance is the mean of the performance of all individual OSTs in the catchment. Likewise, the overall observed performance is estimated by calculating the mean of all soft-sensor measurements in the monitoring module (III). Three base scenarios were modeled for TN numbers of periodic interventions per year: (i) TN = 0 without any intervention, (ii) TN = 1, and (iii) TN = 3. The output of the modeling is the mean availability of the 10 000 OSTs over 10 years.

2.2.2 Model concept

The stochastic model simulates the true and observed performance of one or many individual OSTs in a catchment, as displayed in Figure 1. These results allow the catchment-wide performances to be quantified. Therefore, the model can be scaled to various catchment sizes with different numbers of OSTs. This article presents the specific implementation of the previous study with inherently inaccurate soft sensors. However, the model concept is broadly applicable and can be refined with more or improved information on OSTs' performance and failures and the behavior of unmaintained sensors. The model concept is explained in detail in the supporting information (S1 model concept).

2.2.3 Operational costs

Differences in operational costs are simply represented as the number of interventions.

3 Results and Discussion

3.1 Q1: How do soft-sensor accuracy, reliability of OSTs, and alarm management influence the true treatment performance and the number of interventions required?

Figure 2.1 shows the mean fraction of time that a single OST's performance is up, or functioning, in ten years. Every pixel represents an OST modeled with reliability (the mean survival time λ) on the x-axis and soft-sensor accuracy on the y-axis with identical inputs for sensitivity and specificity. The target is to have both high performance (Figure 2.1) and a low number of interventions (Figure 2.2). Therefore, these two subplots need to be considered jointly. To discuss the results, we highlight two meaningful regions in the subplots of Figure 2.1: Y represents high performance, which is considered as a feasible operation, and X represents low performance, which is infeasible. Similarly, we divided Figure 2.2 into two areas: A represents a low number of interventions, which is feasible, and B represents a high number of interventions, which is infeasible. We use these areas solely to guide the reader through the results and discussion.

In Figure 2.1a, the entire plot belongs to area Y, which means that, independently of the sensor accuracy and the mean survival time, the true treatment performance is nearly always high. The alarm management strategy for (a) is that an intervention is issued as soon as the soft sensor shows a failure (TN or FN). Unsurprisingly, this leads to a very high number of units in the up state. Interestingly, even under these very intervention-intensive circumstances, the performance does not reach 100% for a very low mean survival time and low sensor accuracy. The very large area B in Figure 2.2a shows the high number of interventions, and therefore costs, this alarm management entails. Only with very accurate soft sensors (>98%) and elevated OST reliability does the intervention demand decrease to a reasonably low level.

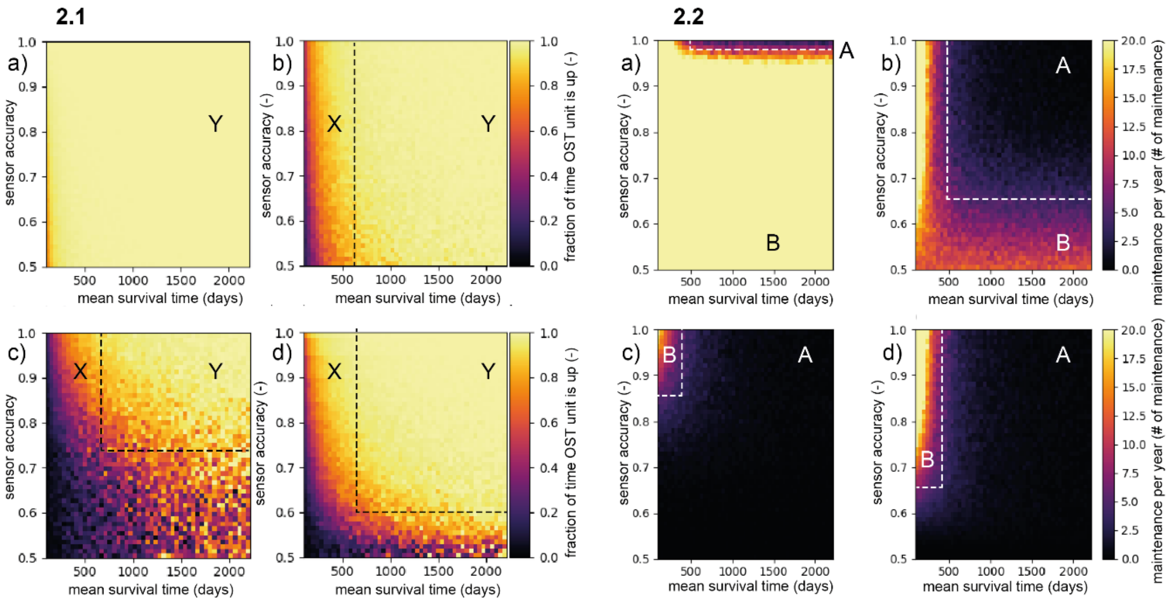

Figure 2: Sensitivity of the fraction of time an OST is up (2.1, true performance of an OST) and maintenance per year (2.2) to two system design parameters, i.e., sensor accuracy and mean survival time (λ). Every pixel represents an OST with a specific sensor accuracy and λ. The difference between the four subplots is the alarm management strategy; an intervention is issued after the soft sensor shows a failure (TN or FN) for (a) 1 day, (b) 4 consecutive days, (c) 14 consecutive days, and (d) conditional probability of a failure > 98%. The x-axis is from 88 to 2216 days. The subplots of 2.1 are schematically divided into two sections: Y for high performance and X for low performance. The subplots of 2.2 are schematically divided by the number of interventions per year into A for feasible and B for infeasible.

Figure 2.1 shows the impact of the alarm management on the up time of the unit. The alarm management defines how many consecutive time steps a down signal needs to be recorded in the monitoring module to trigger an intervention: (a) 1, (b) 4, and (c) 14. The area of Y is smaller in (b) than in (a). This means that a small delay of four time steps of the intervention, as in Figure 2.1b, leads to a substantial decrease in up time for units with low reliability. However, in Figure 2.2b, the area of A is larger than in 2.2a. This means management strategy (b) is preferable over (a) for the number of interventions. This trend continues in Figure 2.2c, with the difference that the effect of low-accuracy sensors is clearly visible as a reduction in area Y in Figure 2.1c. Soft sensors with low or intermediate accuracy have a low probability of showing a series of 14 consecutive negative signals and therefore are prone to considerable intervention delays, even when the down performance is apparent, as in our example. Thus, the optimal value for consecutive negative signals lies between 4 and 14 days. Figure 2.1d shows the impact of more sophisticated alarm management based on the confidence in the monitoring signal. Interventions are triggered when the estimated probability (Eq.s 5 and 6) of a unit having a down state exceeds 0.98. This approach assumes that the soft-sensor accuracy is known and links it with the alarm management. High up time for units is achieved for all but the least reliable units and with medium or higher accuracy soft sensors.

OSTs are operated optimally somewhere in the intersection of areas Y (Figure 2.1) and A (Figure 2.2) ($Y \cap A$), where the true treatment performance is high and the number of interventions remains low. The $Y \cap A$ of alarm management strategy (a) is restricted by A. Figure 2.2a shows that an alarm management that triggers an intervention as soon as the sensor indicates a negative signal will lead to many FN and, therefore, the high costs that many interventions would entail. The exception to this high number of false alarms is to build very accurate sensors with more than 98% accuracy, where none of the alarm management strategies lead to a high number of interventions. Conversely, $Y \cap A$ for strategy (c) is only curtailed by Y, the true treatment performance displayed in Figure 2.1c, suggesting that too few interventions are triggered in strategy (c). In this case, the threshold used to trigger an alarm is set too high, leading to delayed inspections and reduced performance.

The most successful alarm management strategies are (b) and (d). Interestingly, both strategies produce a similar $Y \cap A$ area, although their underlying causes differ. For (b), a high number of interventions are needed (Figure 2.2, area A), whereas the performance is high for any sensor accuracy between 0.5 and 1.0 (Figure 2.1, area Y). For (d), the opposite can be observed: the number of interventions is low for all sensor accuracies, but the true treatment performance drops for sensor accuracies below 0.6.

Assuming that strategies (b) and (d) are both preferable alarm strategies, we can learn that a minimal OST robustness is required or a replacement strategy[40] needs to be set in place if a high number of failures is observed at the end of life of OST units. In (b) and (d), this robustness threshold is around a one-year mean survival time if the failure follows a Weibull distribution (Eq. 6). Soft sensors have no such clear threshold because we assume they are frequently replaced and do not reach the end-of-life phase before replacement.

3.2 Q2: How much can the performance of a large number of OSTs in a catchment be improved by online monitoring?

In this section, we explain how the true performance of each OST (Q1) translates into an overall systems performance of a large number of OSTs (10 000) in a catchment, referred to here as system performance, and how the true system performance compares to the observed one. Furthermore, we discuss how sensor-driven interventions triggered by online monitoring compare to the same number of periodic time-driven maintenance interventions. Figure 3 shows the mean true system performance and the system performance observed from monitoring with inaccurate sensors. The mean number of interventions per unit per year is indicated with N for time-driven interventions and SN for sensor-driven interventions. The exact results (shown in Figure 4 as horizontal, dashed lines) as mean availability are (i) 0.31 of the time, or 31% up time of the 10 000 OSTs out of 10 × 365 days for N = 0; (ii) 0.66 for N = 1; and (iii) 0.83 for N = 3. Figure 3 shows that the observed performance's deviation from the true performance depends on the sensor accuracy and, except in the extreme case of a nearly perfect sensor, is substantially lower than the true performance.

Figure 3 further shows the influence of various soft-sensor accuracies on the true system performance and how accurately the true performance is observed. The observed system performance improves nearly linearly with soft-sensor accuracy. This is the behavior that we would expect in a system with close to 100% true performance. The performance observed in this special case directly reflects soft-sensor accuracy. The true performance is consistently above 90%, even with inaccurate soft sensors. However, high performance despite low sensor accuracy comes at the cost of maintenance frequency. SN = 12 interventions is three times higher than the usual frequency of one to four interventions per year in use today (see introduction). Figure 3 shows that the soft-sensor accuracy should be 0.7 or higher to keep the number of interventions equal to or lower than three per year. The modeled treatment performance for a 0.7 soft-sensor accuracy is 97% (with 0.12% variance, see supporting information). The previously developed pH-based soft sensors have an overall accuracy of 0.80–0.85,[33] well above this minimum of 0.7. These results suggest that the broad adoption of the sensor technology currently available would significantly improve treatment performance. Figure 3 suggests an improved treatment performance from 66% (when N = 1) to 98% (with SN = 1–2 and 0.8 soft-sensor accuracy). Therefore, when changing from the current paradigm of regular, time-driven interventions towards measurement-based, sensor-driven interventions, the system potentially shifts from a mediocre to a decent true performance even with inaccurate soft sensors.

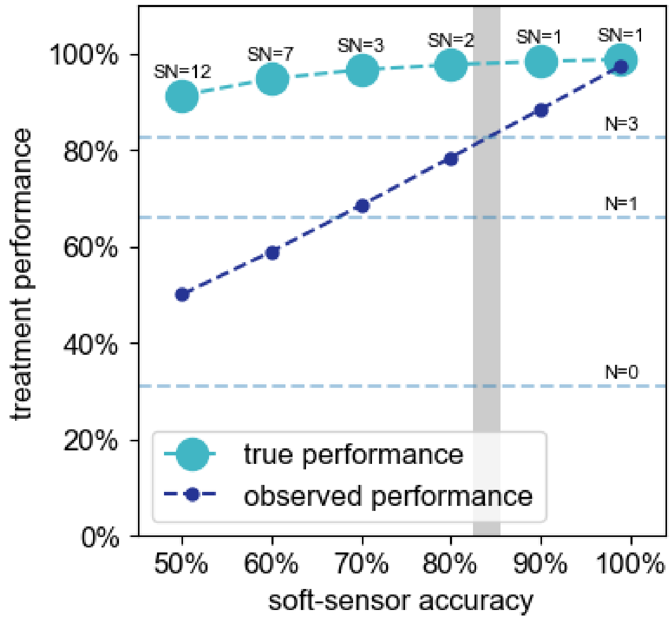

Figure 3: Comparison of true (large circles with the corresponding number of sensor-driven interventions: SN) and observed (small circles) system performance of 10 000 OST for sensor-based maintenance schemes for various soft-sensor accuracies. The horizontal dashed lines are the reference to maintenance schemes without sensors (time-driven interventions: N). They indicate the true system treatment performance for N = 0, 1, and 3. The vertical shaded area indicates the current real-world soft sensor and shows that a true treatment performance of around 98% would be reached with the current sensors, while the observed performance would be expected to be around 80% (intersection of the gray shaded area with the interpolation of the true, respectively observed performance).

3.3 Q3: How does optimizing the parameters of the soft sensor influence the treatment performance and its quantification?

Table 1 shows two diverging aspects of sensor accuracy. The modeling results imply that optimization for specificity improves the system treatment performance but reduces the accuracy of the performance quantification. Monitoring with high specificity is especially relevant for the high inflow variability such small OSTs encounter.[20] Therefore, the difference between true and observed system performance in Figure 3 is sensitive to the parametrization of the soft sensor; the soft sensors in the previous study were optimized for specificity.[33] The benefit of this conservative approach is that it results in more reliable information for alarm management, which in turn improves the true system performance. However, this specificity-optimized soft sensor leads to less accurate quantification of the observed performance.

Table 1: Influence of soft-sensor specificity, sensitivity, and reliability on mean true performance, the mean observed performance, and mean number of interventions per unit per year for a system of 10 000 OSTs. The category "as Figure 3" uses the same random uniformly distributed λ in days as Figure 3 (400 and 2200): "low"(100 to 400), and "high" (800 to 2200).

sensor accuracy	OST reliability	mean true performance	mean *observed* performance	mean number of interventions
specificity 0.98, sensitivity 0.7	as Figure 3	0.99	0.69	3.5
	low	0.82	0.54	23
	high	0.99	0.71	2.5
specificity 0.7, sensitivity 0.98	As Figure 3	0.96	0.95	1.4
	low	0.64	0.71	13
	high	0.98	0.96	0.80

The modeling results suggest that the soft sensor should be used with two sets of optimized parameters: one geared toward high specificity for alarm management and one toward high sensitivity to quantify system performance. Both can be used with the same calibration data as input. Realizing this possible double use of data to obtain different outputs is an important change in mindset and a strength of any soft-sensor approach.

3.4 Alarm management

Two alarm strategies have been identified as preferable to the other two strategies investigated. One is very simple: waiting for four consecutive negative cycles; the other uses a conditional probability that depends on sensor accuracy. The latter is the more elaborate method. However, it has two disadvantages: Firstly, an error in the estimated sensor accuracy could lead to inadequate system performance quantification. Secondly, if the accuracy is lower than estimated, the performance is lower, but the number of interventions will not rise as required because of the overestimated accuracy (Figure 2.1d and Figure 2.2d). In reality, if the low performance of an individual OST occurs due to wrong estimation of sensor accuracy, this low performance would probably not be discovered because performance is estimated from this same inaccurate sensor. In contrast, a sensor with lower accuracy would simply cause more interventions (Figure 2.1b and Figure 2.2b). A rising number of interventions is likely to be noticed by an operator. This shows that when choosing an alarm management strategy, the ease with which an unfavorable state of the system can be identified should also be considered.

3.5 Simplifications, assumptions, and potential model extensions

The current model aims to answer the scoping questions presented at the end of the introduction and is implemented in the most straightforward way, requiring several strong simplifications and assumptions. However, the model[35] is designed in such a way that every module can be refined for applications that go beyond the investigated aspects in this article, such as planning monitoring strategies, and supporting system design decisions such as balancing the costs of investment in sensor hardware and maintenance against the costs of interventions caused by false alarms.

A strong simplification is the applied binary performance model, which assumes that a plant either fulfills the requirements or not. This ignores the possibility that overperforming plants could compensate for minor malfunctions or that a unit self-recovers from a shock. Our conservative model's assumption of stochastic failure behavior therefore represents the lower boundary of system performance.

A second strong simplification is the alarm management implementation presented here. The signals from the sensors are treated as statistically independent, leading to a very simple and, again, conservative result. Possible improvements include using machine learning algorithms for anomaly detection in the online time series and spatially optimizing an intervention strategy that also considers OST locations.

The third very relevant simplification is the constant accuracy of the sensors between the intervention intervals. Our own experience shows that this is reasonably realistic. However, much more research in soft sensors is needed to show how sensor accuracy can be modeled more realistically.

We conclude that most simplifications and assumptions lead to conservative results that indicate the minimal performance achievable for the given inputs.

3.6 Possible model extensions

For further studies, improving the model with more differentiated failure modes, such as biology failure due to prolonged periods without influent, and effective performance distributions would give a more detailed quantitative estimation of the achievable performance of real-world systems of OSTs. An interesting extension would be to couple our performance model with an agent-based modeling approach and a detailed physical-biological treatment plant model. This could also capture various socioeconomic behaviors that affect the performance of systems of OSTs. Furthermore, the model presented in this article could also be modeled with a stochastic Petri net with aging tokens.[41,42]

3.7 Single large or several small?

To decide which system is the most efficient for treating wastewater across an entire catchment, quantifying the performance of a system of OSTs is essential. We showed that accurate quantification remains challenging for OSTs and will require more research on soft sensors. However, when comparing

OST and centralized systems, it is remarkable that even for centralized systems, many unknowns remain about the true treatment performance in an entire catchment: Firstly, losses via combined sewer overflows and faulty connections are only partly monitored;[43] in Switzerland, for instance, 2.7% of the yearly dry-weather flow is estimated to be discharged untreated during rain events via combined sewer overflows.[43] Secondly, losses from sewers due to suboptimal infrastructure management are another unknown. In Europe, sewer losses are estimated to be from 5% to 20%,[44] which is low compared, for example, to the 77% estimated for Vietnam.[45] These two issues highlight that the true system performance of centralized WWTPs needs to be quantified similarly to that of OSTs to achieve a fair comparison of system performance.

This article shows that systems of OSTs with adequate monitoring have the potential to achieve high performances at catchment level, even with conservative assumptions. We believe that this justifies the need for more research into the remote operation and monitoring of systems of decentralized, modular, and unstaffed WWTPs with particular emphasis on resource recovery and reuse.

Lastly, we want to emphasize the importance of evaluating environmental impact from the system perspective. Without professional monitoring, maintenance, and repair, no technical system works properly.[46] This applies equally to OSTs and centralized systems. Therefore, to further scale up systems of OSTs, centralized management[47] and appropriate monitoring strategies are essential.

Abbreviations

N	number of interventions
OST	on-site wastewater treatment plant
SBR	sequencing batch reactor
WRRF	water resource recovery facility
WWTP	wastewater treatment plant

Supporting information

Detailed description of the model concept including model walkthrough, influence of the modeled number of OSTs on the variance.

Acknowledgments

We want to thank Peter Vanrolleghem for his detailed feedback; Rachel Barrett, Philipp Beutler, Lisa Deutsch, Liliane Manny, Matthew Moy de Vitry, Lena Mutzner, Andreas Scheidegger, and Omar Wani for input; and Simon Milligan for language advice. We thank two anonymous reviewers for seeing the value in this manuscript and contributing significantly to improving it with their detailed, constructive comments.

Author contributions

MYS did the modeling and led the project and the writing of the article. MM and KV provided ideas and expertise. HH provided experience on OSTs in Japan. All authors contributed to the writing of this document.

Funding Sources

MYS received funding from the Japanese Society for Promotion of Science (JSPS) Grant GR19105 and P20763.

Note

An earlier version of this manuscript has been published in the PhD thesis "Monitoring and quantifying the treatment performance of on-site wastewater treatment plants"[48] and as a preprint.[49]

References

(1) Gikas, P.; Tchobanoglous, G. The Role of Satellite and Decentralized Strategies in Water Resources Management. *J. Environ. Manage.* **2009**, *90* (1), 144–152. https://doi.org/10.1016/j.jenvman.2007.08.016.

(2) Hoffmann, S.; Feldmann, U.; Bach, P. M.; Binz, C.; Farrelly, M.; Frantzeskaki, N.; Hiessl, H.; Inauen, J.; Larsen, T. A.; Lienert, J.; Londong, J.; Lüthi, C.; Maurer, M.; Mitchell, C.; Morgenroth, E.; Nelson, K. L.; Scholten, L.; Truffer, B.; Udert, K. M. A Research Agenda for the Future of Urban Water Management: Exploring the Potential of Non-Grid, Small-Grid, and Hybrid Solutions. *Environ. Sci. Technol.* **2020**, acs.est.9b05222. https://doi.org/10.1021/acs.est.9b05222.

(3) Langergraber, G.; Muellegger, E. Ecological Sanitation—a Way to Solve Global Sanitation Problems? *Environ. Int.* **2005**, *31* (3), 433–444. https://doi.org/10.1016/j.envint.2004.08.006.

(4) Sedlak, D. *Water 4.0: The Past, Present, and Future of the World's Most Vital Resource*; Yale University Press: New Haven, CT, USA, 2014.

(5) Larsen, T. A.; Hoffmann, S.; Lüthi, C.; Truffer, B.; Maurer, M. Emerging Solutions to the Water Challenges of an Urbanizing World. *Science* **2016**, *352* (6288), 928–933. https://doi.org/10.1126/science.aad8641.

(6) Neumann, M. B.; Rieckermann, J.; Hug, T.; Gujer, W. Adaptation in Hindsight: Dynamics and Drivers Shaping Urban Wastewater Systems. *J. Environ. Manage.* **2015**, *151*, 404–415. https://doi.org/10.1016/j.jenvman.2014.12.047.

(7) Simberloff, D. S.; Abele, L. G. Island Biogeography Theory and Conservation Practice. *Science* **1976**, *191* (4224), 285–286. https://doi.org/10.1126/science.191.4224.285.

(8) Eggimann, S.; Truffer, B.; Maurer, M. To Connect or Not to Connect? Modelling the Optimal Degree of Centralisation for Wastewater Infrastructures. *Water Res.* **2015**, *84*, 218–231. https://doi.org/10.1016/j.watres.2015.07.004.
(9) Moelants, N.; Janssen, G.; Smets, I.; Van Impe, J. Field Performance Assessment of Onsite Individual Wastewater Treatment Systems. *Water Sci. Technol.* **2008**, *58* (1), 1–6. https://doi.org/10.2166/wst.2008.325.
(10) Hug, T.; Maurer, M. Stochastic Modeling to Identify Requirements for Centralized Monitoring of Distributed Wastewater Treatment. *Water Sci. Technol.* **2012**, *65* (6), 1067–1075. https://doi.org/10.2166/wst.2012.945.
(11) Baeriswyl, P.; Carisch, M.; Etter, B.; Holliger, U.; Hug, T.; Konz, N.; Lehmann, S.; Maron, T.; Oetterli, M.; Schneider, M. Y.; Tinner, B.; Udert, K. M. *Abwasser im ländlichen Raum (wastewater in rural area, in German)*; VSA: Glattbrugg, 2017.
(12) Levett, K. J.; Vanderzalm, J. L.; Page, D. W.; Dillon, P. J. Factors Affecting the Performance and Risks to Human Health of On-Site Wastewater Treatment Systems. *Water Sci. Technol.* **2010**, *62* (7), 1499–1509. https://doi.org/10.2166/wst.2010.434.
(13) Ministry of the Environment. *Johkasou Act (Tentative Translation)*; Act No. 43, Revised version Act No. 69; Japan, 1983.
(14) U.S. Environmental Protection Agency. Handbook for Managing Onsite and Clustered (Decentralized) Wastewater Treatment Systems. **2005**, EPA-832-B-05-001.
(15) Government of South Australia. On-Site Wastewater Systems Code, 2013.
(16) DIBt. Zulassungsgrundsätze Für Allgemeine Bauaufsichtliche Zulassungen Für Kleinkläranlagen (in German). **2012**, 1–15.
(17) DWA. *DWA-M 221 Grundsätze für Bemessung, Bau und Betrieb von Kleinkläranlagen mit aerober biologischer Reinigungsstufe (in German)*; Hennef, 2012.
(18) Schalk, T. Möglichkeiten zur Behandlung saisonal anfallender Abwässer in ländlich strukturierten Gebieten durch Anpassung der Bemessungsgrundlagen und temporäre Aktivierung interner Reserven (in German). Dresdener Berichte, Band 43, TU Dresden, Dresden, 2017.
(19) Ministry of the Environment, Japan. *Kankyohshou Kankei Johkasouhou Sekohkisoku (in Japanese, Enforcement Regulations for the Johkaso Act Related to the Ministry of the Environment)*; Tokyo, 1984.
(20) Lowe, K. S.; Tucholke, M. B.; Tomaras, J. M. B.; Conn, K.; Hoppe, C.; Drewes, J. E.; McCray, J. E.; Munakata-Marr, J. *Influent Constituent Characteristics of the Modern Waste Stream from Single Sources*; IWA Publishing, 2010. https://doi.org/10.2166/9781780403519.
(21) Lowe, K. S.; Tucholke, M. B.; Tomaras, J. M. B.; Conn, K.; Hoppe, C.; Drewes, J. E.; McCray, J. E.; Munakata-Marr, J. *Influent Constituent Characteristics of the Modern Waste Stream from Single Sources: Final Report*. http://web.archive.org/web/20190216024758/http://www.decentralizedwater.org/research_project_04-DEC-1.asp (accessed 2020-04-28).
(22) Schranner, T. Kleinkläranlagen Bewähren Sich in Der Praxis (in German). *Korresp. Abwasser Abfall* **2014**, *61* (8), 695–700. https://doi.org/10.3242/kae2014.08.003.
(23) Klinger, M.; Ulrich, L.; Ramprasad, C.; Wolf, A. T.; Reynaud, N.; Narayan, A. S.; Siemsen, P.; Lüthi, C.; Philip, L. *Technology, Implementation and Operation of Small-Scale Sanitation in India–PerformanceAnalysis and Policy Recommendations*; 4S Project Report Vol. I; 2020.
(24) Straub, A. Einfache Messmethoden zur Charakterisierung sowie Maßnahmen zur Erhöhung der Zuverlässigkeit und Leistungsfähigkeit biologischer Kleinkläranlagen (in German). Schriftenreihe Siedlungswasserwirtschaft und Umwelt: Heft 17, Cottbus, 2008.
(25) Abegglen, C.; Joss, A.; McArdell, C. S.; Fink, G.; Schlüsener, M. P.; Ternes, T. A.; Siegrist, H. The Fate of Selected Micropollutants in a Single-House MBR. *Water Res.* **2009**, *43* (7), 2036–2046. https://doi.org/10.1016/j.watres.2009.02.005.
(26) Schneider, M. Y.; Furrer, V.; Sprenger, E.; Sterkele, B.; Richter, S.; Rottermann, K. Data for: Benchmarking Soft-Sensors for Remote Monitoring of on-Site Wastewater Treatment Plants. **2020**. https://doi.org/10.25678/000194.
(27) Crites, R.; Tchobanoglous, G. *Small and Decentralized Wastewater Management Systems*; McGraw-Hill Ser. in Water Resources and Environmental Engineering; McGraw-Hill: Boston, 1998.
(28) Coleman, J. J.; Abrams, I. J. Mathematical Model for Operational Readiness. *Oper. Res.* **1962**, *10* (1), 126–138. https://doi.org/10.1287/opre.10.1.126.
(29) Rieger, L.; Vanrolleghem, P. A. MonEAU: A Platform for Water Quality Monitoring Networks. *Water Sci. Technol.* **2008**, *57* (7), 1079–1086. https://doi.org/10.2166/wst.2008.135.
(30) Thomann, M.; Rieger, L.; Frommhold, S.; Siegrist, H.; Gujer, W. An Efficient Monitoring Concept with Control Charts for On-Line Sensors. *Water Sci. Technol.* **2002**, *46*, 107–116.

(31) Olsson, G. Chapter 13: The Potential of Control and Monitoring. In *Source Separation and Decentralization for Wastewater Management edited by T. A. Larsen, K. M. Udert, J. Lienert*; IWA Publishing: Londen, UK, 2013; pp 179–192.
(32) Schneider, M. Y.; Furrer, V.; Sprenger, E.; Carbajal, J. P.; Villez, K.; Maurer, M. Benchmarking Soft Sensors for Remote Monitoring of On-Site Wastewater Treatment Plants. *Environ. Sci. Technol.* **2020**, *54* (17), 10840–10849. https://doi.org/10.1021/acs.est.9b07760.
(33) Schneider, M. Y.; Carbajal, J. P.; Furrer, V.; Sterkele, B.; Maurer, M.; Villez, K. Beyond Signal Quality: The Value of Unmaintained PH, Dissolved Oxygen, and Oxidation-Reduction Potential Sensors for Remote Performance Monitoring of on-Site Sequencing Batch Reactors. *Water Res.* **2019**, *161*, 639–651. https://doi.org/10.1016/j.watres.2019.06.007.
(34) Van Rossum, G.; Drake, F. L. *Python 3 Reference Manual*; CreateSpace: Scotts Valley, CA, 2009.
(35) Schneider, M. Y. Systemperformance Module Python 3, 2020.
(36) Barlow, R. E.; Proschan, F.; Hunter, L. C. *Mathematical Theory of Reliability*; Classics in applied mathematics; SIAM: Philadelphia, 1996.
(37) Weibull, W. A Statistical Distribution Function of Wide Application. *ASME J. Appl. Mech.* **1951**, 293–297.
(38) Ohmura, K.; Thürlimann, C. M.; Kipf, M.; Carbajal, J. P.; Villez, K. Characterizing Long-Term Wear and Tear of Ion-Selective PH Sensors. *Water Sci. Technol.* **2019**, *80* (3), 541–550. https://doi.org/10.2166/wst.2019.301.
(39) Wald, A. Sequential Tests of Statistical Hypotheses. *Ann. Math. Stat.* **1945**, *16* (2), 117–186. https://doi.org/10.1214/aoms/1177731118.
(40) Lotka, A. J. A Contribution to the Theory of Self-Renewing Aggregates, With Special Reference to Industrial Replacement. *Ann. Math. Stat.* **1939**, *10* (1), 1–25. https://doi.org/10.1214/aoms/1177732243.
(41) Volovoi, V. Modeling of System Reliability Petri Nets with Aging Tokens. *Reliab. Eng. Syst. Saf.* **2004**, *84* (2), 149–161. https://doi.org/10.1016/j.ress.2003.10.013.
(42) Volovoi, V. System-Level Maintenance Policies via Stochastic Petri Nets with Aging Tokens. In *2007 Annual Reliability and Maintainability Symposium*; 2007; pp 89–94. https://doi.org/10.1109/RAMS.2007.328049.
(43) Mutzner, L.; Staufer, P.; Ort, C. Model-Based Screening for Critical Wet-Weather Discharges Related to Micropollutants from Urban Areas. *Water Res.* **2016**, *104*, 547–557. https://doi.org/10.1016/j.watres.2016.08.003.
(44) Rutsch, M.; Rieckermann, J.; Krebs, P. Quantification of Sewer Leakage: A Review. *Water Sci. Technol.* **2006**, *54* (6–7), 135–144. https://doi.org/10.2166/wst.2006.616.
(45) Watanabe, R.; Harada, H.; Yasui, H.; Le, T. V.; Fujii, S. Exfiltration and Infiltration Effect on Sewage Flow and Quality: A Case Study of Hue, Vietnam. *Environ. Technol.* **2019**, 1–11. https://doi.org/10.1080/09593330.2019.1680739.
(46) Graham, S.; Thrift, N. Out of Order: Understanding Repair and Maintenance. *Theory Cult. Soc.* **2007**, *24* (3), 1–25. https://doi.org/10.1177/0263276407075954.
(47) Reymond, P.; Abdel Wahaab, R.; Moussa, M. S.; Lüthi, C. Scaling up Small Scale Wastewater Treatment Systems in Low- and Middle-Income Countries: An Analysis of Challenges and Ways Forward through the Case of Egypt. *Util. Policy* **2018**, *52*, 13–21. https://doi.org/10.1016/j.jup.2018.03.008.
(48) Schneider, M. Y. Monitoring and Quantifying the Treatment Performance of On-Site Wastewater Treatment Plants. Doctoral Thesis, ETH Zurich, 2020. https://doi.org/10.3929/ethz-b-000445338.
(49) Schneider, M. Y.; Harada, H.; Villez, K.; Maurer, M. How Well Can Inaccurate Sensors Quantify and Improve the Performance of a Fleet of On-Site Wastewater Treatment Plants? engrXiv September 10, 2021. https://doi.org/10.31224/osf.io/4myz5.