ICA2025OSLO
14th IWA Specialised Conference on Instrumentation, Control and Automation

# Automated influent data generation for digital twin applications

C. Wärff*,**, O. Samuelsson***, H. Aydin****, M. Arnell*,**, U. Jeppsson**

* RISE Research Institutes of Sweden, Sven Hultins Plats 5, 41258 Gothenburg, Sweden (E-mail: christoffer.warff@ri.se; magnus.arnell@ri.se)

** Division of Industrial Electrical Engineering and Automation (IEA), Department of Biomedical Engineering, Lund University, P.O. Box 118, 22100 Lund, Sweden (E-mail: ulf.jeppsson@iea.lth.se)

*** IVL Swedish Environmental Research Institute, Valhallavägen 81, Stockholm, 114 28, Sweden, (E-mail: oscar.samuelsson@ivl.se)

**** Nordvästra Skånes Vatten och Avlopp, P.O. Box 2022, 25002 Helsingborg, Sweden (E-mail: hilat.aydin@nsva.se)

**Abstract:** Digital twins of WRRFs require high quality and high time resolution data for accurate simulations, but such sensors are expensive. The goals of this study are to automate WRRF influent data generation and reconciliation and evaluate the accuracy of the methods. A COD and an ammonium sensor were installed at the WRRF inlet and automatically fault checked and calibrated using daily composite samples. $PO_4$-P was estimated through linear relationship to $NH_4$-N. Organically bound nitrogen and phosphorus fractions were estimated from plantwide model calibration. Laboratory values were fault checked using ratios, statistics and a moving average to adapt to seasonal variations. The validation showed good results for most variables (around 10 % error), while larger errors were found for the sCOD and TSS variables. The described method can be used to provide high frequency data for a low investment and maintenance cost, particularly for nitrogen variables.

**Keywords**: Digital twins; influent data generation; data reconciliation.

## Introduction

Digital twins of water resource recovery facilities (WRRF) (Torfs et al., 2022) are gaining increasing interest as tools to improve plant performance and evaluation. Digital twins are process models of an existing plant, with an automated data feed from the real process and have potential of being used for, e.g., regular, automated process optimisation (Sparks et al., 2023; Jonhson et al., 2024), forecasting and what-if scenarios (Daneshgar et al., 2024). Access to accurate high time resolution influent data is a crucial step for achieving accurate digital twin predictions. However, due to the difficult measurement environment at the WRRF influent with high concentrations of solids and bacteria, accurate high frequency measurements are often time consuming and expensive. Robust and accurate low-cost methods for influent measurements are needed, but so far, such methods are scarce in the scientific literature.

Several methods are available for measuring compounds in wastewater, with a different level of accuracy, maintenance need and cost. Ion-selective sensors are low-cost alternatives available for measuring, e.g., ammonium nitrogen, but have problem with signal drift which can sometimes occur rapidly. Furthermore, re-calibration of the sensor is often difficult and can lead to errors (Pedersen et al., 2020). A more accurate alternative is using a chemical analyser, but for a higher initial investment cost and with lower data frequency than an ion-selective sensor. Using such a sensor at the WRRF inlet is difficult, however, due to the pre-filtering required before analysis which can rapidly clog due to the high concentration TSS in the influent wastewater. For COD, ultraviolet-visible (UV-Vis) optical sensors are a lower cost alternative, with the downside that they are easily affected by fouling (Matsché & Stumwöhrer, 1996) and that not all organic compounds are detectable with this method. Online analysers are a higher cost alternative, and while analysers using the standard dichromate method exists and are (probably) the most accurate, they require handling a reagent containing mercuric sulfate ($HgSO_4$) which is not permitted in Sweden. Other options include thermal combustion analyser, which combusts the sample and measure the oxygen consumed, and *total organic carbon* (TOC) analysers, although the TOC analyser will require a conversion factor to convert the measurement data to COD. For

orthophosphate and total nitrogen and phosphorus, there are (to our knowledge) no readily available options to chemical analysers, meaning that they, especially for filtered samples, are subject to the difficult measurement environment previously mentioned. For those, it can therefore be advantageous to use a soft sensor instead.

To overcome the issue with drift and erroneous re-calibration of the ion-selective ammonium sensor at a WRRF inlet, Pedersen et al. (2020) developed a method to adjust the measured value based on measured daily composite samples. A correction factor is applied to days with composite samples available, and for historical data interpolation of correction factors can be used to account for sensor drift between sampling points. The method could successfully improve the sensor signal for 87 % of the days where there were samples available. Nivert et al. (2009) used another method for adjusting sensor signals from ion-selective and optical sensors. They used break-point curves, i.e., sections of several linear relationships between the sensor reference value and the process value, to cover a wider range of concentration values more accurately. The linear relationships were updated with data from the $N$ last grab samples using a recursive least squares (RLS) algorithm to allow lower influence of older samples. Although this method was proven reliable (Lumley et al., 2013), a large amount of grab samples were used (2-5 per sensor per week). In summary, the use of low-cost sensors combined with automatic correction appear to be a viable alternative to more expensive equipment.

To limit the extra workload when installing new sensors at the inlet of a WRRF for a digital twin application, it would be advantageous if routine sampling frequency does not need to be increased but can rely on the current sampling frequency of composite daily samples. Further, a robust fault detection method is required to assure that the sensor signal and daily composite values are free from faults before the sensor signal is corrected. The goals of this study were therefore to: 1) Develop a method for automated data fault detection of daily composite sample data; 2) Develop a method to generate hourly influent data for a digital twin at a WRRF in Helsingborg, Sweden, using low-cost sensors and soft sensor measurements combined with automated sensor adjustments from daily composite samples; and 3) Evaluate the accuracy of the developed methods.

## Material and methods

*Composite samples fault detection*

A lab data fault detection method was developed based on two steps. In the first step, daily influent load was calculated by multiplying the measurements with the daily average flow rate. A period with assumed good data quality in terms of load variations was identified for each variable. As the data was assumed correct for this period and was close to normally distributed (data not shown), 99 % confidence intervals could be calculated and error limits defined relative the mean load. Calculated load values at day $i$ ($L_i$) were considered potential errors if the following conditions were met:

$$\frac{(\mu + 2.58\sigma)}{\mu} \leq L_i \leq \frac{(\mu - 2.58\sigma)}{\mu} \tag{1}$$

where $\mu$: mean load; $\sigma$: standard deviation. For the full dataset (2020–2024), rolling averages and corresponding rolling error limits were calculated. This way, the observed seasonal variations in load could be accounted for. Data points outside of the error limits were classified as potential faults.

In the second step, ratios were calculated between variables (Rieger et al., 2013), for $NH_4$-N/TN, $NH_4$-N/TCOD, TN/TCOD, TSS/TCOD and fCOD/TCOD. Error limits were determined

based on outlier detection with the interquartile range method, where the calculated ratio at day $i$ ($r_i$) that fulfil the following condition were considered potential errors:

$$P_{25} - 1.5IQR > r_i > P_{75} + 1.5IQR \quad (2)$$

where $P_{25}$: 25th percentile of the data; $IQR$: interquartile range; $P_{75}$: 75th percentile. For COD error detection, the last four ratios listed above were evaluated, while the first one was used for ammonium. Data points that were flagged as potential faults in both the load analysis and in the ratio analysis (in any of the used ratios) were removed and not used for calibration of the sensors.

*Automated influent generation method*

Two new sensors were installed at the plant: an optical UV-Vis sensor providing values for total COD (TCOD) and soluble COD (sCOD); and an ion-selective electrode (ISE) ammonium sensor. The sensors were placed in a small cylindrical tank (approximately 25 cm in diameter and a water level of 75 cm) where wastewater was pumped (4 L/s) from the plant influent to the bottom of the tank, with an overflow at the top. The ammonium sensor is a WTW AmmoLyt Plus SET/Comp. Initial calibration of the sensor was done by the supplier on site after installation. The COD-sensor is a WTW CarboVis 701 IQ TS, which has a 1 mm measuring gap and measures UV-Vis in the 200–720 nm range. In addition to the measurement channel, the sensor has a reference channel which automatically corrects for aging optics in the sensor. In theory, sensor drift should therefore only occur due to fouling. The output from the sensor is a raw signal and a calibrated signal, both separate for TCOD and sCOD. The raw signal is corrected based on a linear equation to obtain the measurement. Initial calibration of the sensor was made with grab samples.

The raw sensor data were checked for faults and gaps, inspired by Daneshgar et al., (2024). Faults were identified through: comparison to thresholds where few dynamics could be distinguished (including zero values); and constant signal values. Noise was removed through: 1) a rate of change analysis, based on the change in concentration from one minute to the next; 2) deviation from the filtered signal value. Threshold values were defined for both noise identification methods, were values higher than these were considered faults (and set to NaN).

The ammonium sensor data was automatically adjusted based on the method from Pedersen et al. (2020), matching composite sample values to daily flow proportional mean values calculated from the sensor. A similar approach was used for the COD-sensor, where the offset parameters in the linear equations were adjusted to obtain the correct daily average as compared to the measured daily composite data. Only routine operational data classified as non-faults in the fault detection were used in this step (except sCOD, which was not measured in the routine analyses). An "online" and an "offline" method for correcting the sensor values were used, according to the definition in Pedersen et al. (2020).

A simple orthophosphate soft sensor was constructed through observed linear relationship to the ammonium nitrogen load. The influent was samples every other hour between 2023-09-04 09:00 and 2023-09-05 07:00 (in total 12 samples) and sent to an external laboratory for analysis of $NH_4$-N and $PO_4$-P. The measured influent flow rate values were extracted for the same hours, and the load of each variable calculated. A linear relationship was established according to (2):

$$L_{PO4-P} = 0.12\, L_{NH4-N} - 33255.43 \quad (3)$$

where $L_{PO4-P}$: orthophosphate phosphorus load [g P $d^{-1}$]; and $L_{NH4-N}$: ammonium nitrogen load [g N $d^{-1}$]. The calculated orthophosphate load could then be divided by the flow rate to obtain

the orthophosphate concentration. The calculation could then be done using the corrected ammonium sensor data to obtain high frequency orthophosphate data.

Influent wastewater characterisation was done for four separate daily composite samples between 2021-10-27 and 2022-05-30, using a modified version of the STOWA protocol (Roeleveld & van Loosdrecht, 2002). A plantwide model for Öresundsverket WRRF was constructed in the commercial simulation platform Sumo22 (Dynamita SARL, France) and calibrated using the Sumo2S biokinetic model on averaged data for January 2024. Organically bound nitrogen and phosphorus fractions were adjusted in the calibration procedure and the final values related to the content in the different COD-fractions (i.e., expressed as g N g $COD^{-1}$ and g P g $COD^{-1}$). These were then assumed constant. By combining the COD-fractions with sensor and soft sensor data, a complete dynamic influent characterisation could be constructed and values of TN, TP, VSS and TSS could be calculated.

*Validation data*

To validate the sensor measurements, an intensive measurement campaign was conducted where daily composite influent samples were collected between 2024-09-29 – 2024-10-31. Five samples per week were collected, for a total of five weeks (25 sampling days). For each sampling day, the following parameters were analysed in triplicate (using Hach LCK vial tests for chemical analyses): TCOD, sCOD (flocculated + filtered through 0.45 µm filter), TN, $NH_4$-N, TP, $PO_4$-P, TSS and VSS. The hourly $PO_4$-P and $NH_4$-N data collected to develop the $PO_4$-P soft sensor were used to validate the $NH_4$-N sensor and $PO_4$-P soft sensor data dynamically as well.

**Results and discussion**

*Fault detection*

Figure 1 shows comparisons between the values used for calibrating the sensors (ordinary routine analyses, apart from sCOD), including samples that were identified as faults, and the mean value from the triplicate analyses during the measurement campaign. Only two samples were identified as faults during this period, one for TCOD and one for $NH_4$-N. The one for TCOD, on October 14th, deviate substantially from the mean of the triplicate analysis, indicating that it was correctly identified as a fault. The one for $NH_4$-N on October 10th though, is nearly identical to the mean of the triplicate analysis. The flow rate during this day (Figure 1 (d)) show a substantial wet weather event during this day, which caused both the measured load and the $NH_4$-N/TN ratio to be slightly lower than the error detection limit. Even without calibration to data from this day, the daily average calculated from the $NH_4$-N sensor was near identical to the mean from the measurement campaign (Figure 2 (d)). This indicates that the error detection method misclassified the $NH_4$-N data point as an error. There is a balancing act between using a more aggressive error detection, and risking false positives, and having a less aggressive one and risking false negatives.

For this case, where $NH_4$-N is normally analysed five times per week, excluding one value has little impact on the sensor calibration. For the TCOD, which is only measured once per week, the impact is larger as a removed value means that there will be two weeks between re-calibration of the sensor. On the other hand, TCOD measurements are more prone to substantial errors due to the influence of solids and challenge of correctly homogenise the sample before analysis. If the TCOD error on October 14th had not been detected, the sensor would have been re-calibrated to a value nearly twice as high as the more likely value. This would cause a large overestimation in the digital twin process model until the next measurement, which would occur at least one week later. Therefore, the level aggressiveness used appear sufficient for this application. If needed, the error limits can be manually tweaked to increase or decrease the detection frequency. Due to the

sensitivity of the TCOD-sensor to measurement errors, use of the fault detection directly in the laboratory (to enable re-analysis if an error is detected) or increasing the number of samples per week to at least two would increase the robustness of the method.

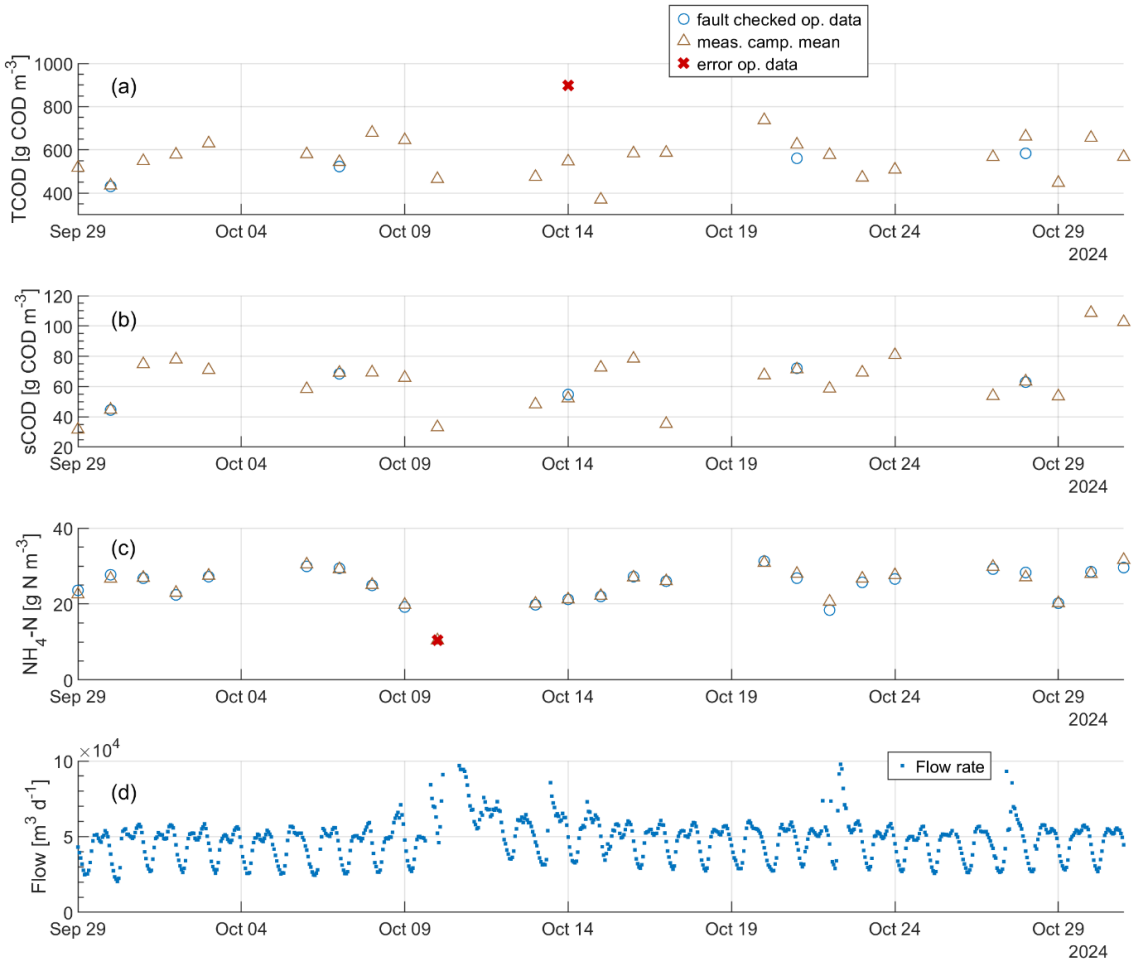


Figure 1 Comparison between daily composite sample analyses used for sensor calibration (based on routine operational data) and mean values from triplicate analyses in the measurement campaign, for TCOD (a), sCOD (b) and $NH_4$-N (c), as well as hourly influent flow rate during the period (d). Note that no routine sampling of sCOD had started at the time of the campaign, the first sCOD of the triplicate samples during the same day as TCOD was measured in the routine analyses were therefore used for the automated sensor calibration.

*Sensor based influent generation*

Figure 2 shows sensor based (calculated, through the online and offline methods) and measured composite samples, with error bars based on 95$^{th}$ % confidence intervals (± 1.96 σ) calculated from the triplicate analyses. Statistics regarding absolute errors during the whole evaluation period is shown in Table 1. In general, most of the sensor-based calculations are close to the measured data, with mean absolute errors slightly above or below 10 %.

sCOD has the highest percentual error rate, although the absolute error is still small compared to the total COD due to the low sCOD concentration. The standard deviation of the error is also large, displaying the range in error values (from insignificant to large) that can be seen in Figure 2 (b). The impact from this on use in a process model is therefore expected to be small. The error for the soluble COD sensor is partly due to negative values occurring at a few of the days, mainly during the rain event on October 10$^{th}$. This occurs due to the very low concentration values which are out of the range that the sensor was calibrated at. This could probably be solved by using dilution to obtain artificially lower values during the calibration procedure. A break point curve might also be beneficial to use if the sensor parameters are significantly different in the low range.

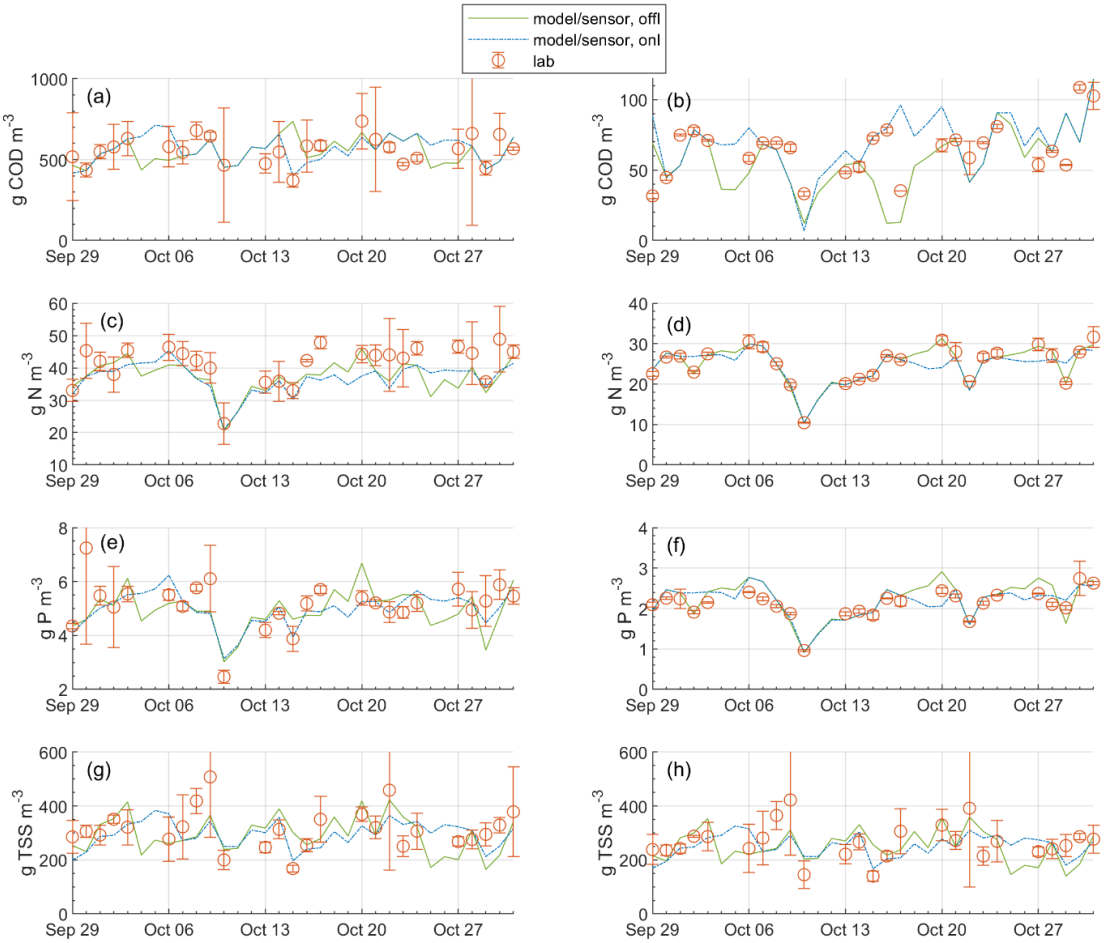


Figure 2 Calculated and measured influent composite samples during 2024-09-29 – 2024-10-31, for TCOD (a), sCOD (b), TN (c), $NH_4$-N (d), TP (e), $PO_4$-P (f), TSS (g) and VSS (h). Error bars based on mean ± 1.96 std (95 % conf. interval) based on triplicate analyses.

The mean TSS and VSS errors are close to 20 %. This is probably due a combination of factors: 1) they are directly affected by the COD-sensor error and assumption on the composition; 2) there is a large variation in many of the triplicate samples, showing uncertainty in the analysis; and 3) calculations of particulate and colloidal COD to VSS ratio (data not show) show large variations in this ratio (most values between 1.4 and 2.5). This means that there could be either substantial variation in the particulate and colloidal COD-fractions, or in the VSS to COD conversion parameters (which are assumed constant), or both.

The errors in calculated TN and TP are low, around 10 %. Like the TSS variables, these will be affected by the different COD fractions and errors in the TCOD sensor. However, the low errors seen here indicate that the assumption of constant N and P content in the different COD variables is reasonable, given the different times of year between the calibration and evaluation data. If they are found to vary, which could be the case for example during vacation periods with changing demographics, an automatic calibration on laboratory data could be used. For example, through optimising the N- and P- content in $X_B$ to fit the daily composite TN and TP data. In a digital twin setting, where the process model is automatically simulated regularly, this can also be automated by keeping track of the model fit to different N and P related measurements. Depending on the objective of the digital twin, it could also be reasoned that the difference might be small enough to not have a significant enough impact on the results to warrant adjusting.

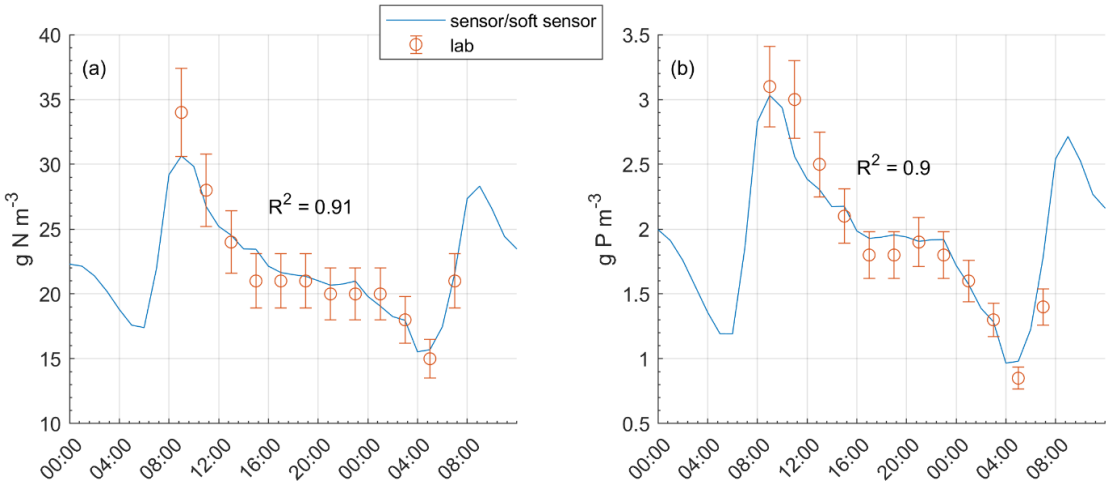


Figure 3 Measurements of hourly composite samples of $NH_4$-N (a) and $PO_4$-P (b), together with sensor and soft sensor data. Error bars represent uncertainty given by the external laboratory, corresponding to ± 2σ. Data from 2023-09-04 00:00 – 2023-09-05 12:00.

Table 1 Errors between daily composite values measured in laboratory and calculated from sensor/model data during the measurement campaign.

| | | TCOD | sCOD | TN | $NH_4$-N | TP | $PO_4$-P | TSS | VSS |
|---|---|---|---|---|---|---|---|---|---|
| Online method | | | | | | | | | |
| Mean abs. error | g m$^{-3}$ | 72 | 17 | 4.6 | 1.4 | 0.50 | 0.18 | 63 | 51 |
| | % | 13 | 34 | 11 | 5.3 | 9.3 | 8.4 | 20 | 19 |
| Standard dev. | g m$^{-3}$ | 51 | 17 | 3.0 | 1.8 | 0.54 | 0.12 | 37 | 34 |
| | % | 9.0 | 47 | 6.2 | 6.9 | 8.2 | 5.4 | 9.4 | 12 |
| Offline method | | | | | | | | | |
| Mean abs. error | g m$^{-3}$ | 80 | 16 | 4.7 | 0.61 | 0.63 | 0.20 | 63 | 52 |
| | % | 15 | 28 | 11 | 2.4 | 12 | 9.4 | 22 | 22 |
| St.dev. abs. error | g m$^{-3}$ | 76 | 16 | 2.7 | 0.58 | 0.58 | 0.14 | 52 | 45 |
| | % | 19 | 30 | 5.9 | 2.4 | 10 | 5.9 | 26 | 28 |

A comparison between the online and offline method shows errors in a similar range, but with the offline method in general showing slightly higher mean errors. The largest discrepancy between the methods can be seen for the COD-sensor data, with certain periods with large variation between the two methods. This is coinciding with a period with substantial drift of the sensor, probably due to fouling, with a sudden decrease in the sensor signal when the sensor is cleaned (data not shown). Depending on when this occurs and when the daily composite data is available for re-calibration, there can be large deviations in calibrated values. This highlights the need of regular cleaning of the sensor and measurement of influent COD more often than once per week to make sure drift due to fouling is avoided and if it happens that it is rapidly compensated for.

The $NH_4$-N sensor performs well when compared to both daily composite samples and the hourly data (Figure *3* (a)). The $PO_4$-P soft sensor, which is based on the $NH_4$-N sensor signal, also performs well based on both hourly and daily composite samples, showing the suitability of this simple soft sensor. It should be noted that the good correlation between $NH_4$-N and $PO_4$-P found here is likely due to the low amounts of phosphorus contained in detergents, where a strict limit has been imposed in Sweden since 2008. In countries where this is not the case, it might not be possible to find such a clear correlation.

**Conclusions**

A method to automate generation of high frequency data in real time for a digital twin simulation model has been developed, based on sensor data with automatic calibration and model calibration. The following conclusions can be drawn:

- A combination of evaluation of load statistics and ratio between measurements could be used to successfully identify substantially deviating operational data but identified a probable correct data point as faulty during a heavy rain event.
- The validation of the sensor-based measurements showed good agreement between most variables, with daily average errors around 10 %. A clear exception was TSS and VSS, the reason for this deviation is yet to be established.
- The simple $PO_4$-P soft sensor based on the observed relationship between $NH_4$-N and $PO_4$-P showed high accuracy compared to both hourly and daily composite data.

The described method can be used to provide high frequency influent data for digital twins for a low investment and maintenance cost.

**REFERENCES**

Daneshgar, S., Polesel, F., Borzooei, S., Sørensen, H.R., Peeters, R., Weijers, S., Nopens, I & Torfs, E. (2024). A full-scale operational digital twin for a water resource recovery facility – A case study of Eindhoven water resource recovery facility. *Wat. Env. Res.* **96**(3), e11016.

Johnson, B.R., Yang, C., Rieger, L., Registre, J., Lesnik, K., Menniti, A., Villez, K., Pienta, D., Stewart, H., Johnson, T., Miletic, I., Rickerman, J., Oristan, M. & Peterson, T. (2024). Full-scale development and piloting of a hybrid digital twin for wastewater operations optimization. *Proc. 11th IWA Water Resource Recovery Modelling Seminar (WRRmod2024)*, Notre Dame, USA, 6-10 April 2024.

Lumley, D., Lindell, P., Lindqvist, P. & Heinonen, M. (2013). Experience using auto-adjustment for improving sensor signal robustness at two large wastewater treatment plants. *Proc. 11th IWA conference on Instrumentation, Control and Automation (ICA)*, Narbonne, France, 18-20 Sept. 2013.

Matsché, N. & Stumwöhrer, K. (1996). UV absorption as control-parameter for biological treatment plants. *Wat. Sci. Tech.* **33**(12), 211-218.

Nivert, G., Lindqvist, P., Lumley, D., Rosenqvist, F. & Gunnarsson, J. (2009). Implementing auto-adjustment and auto-validation of on-line instrument signals. *Proc. 10th IWA conference on Instrumentation, Control and Automation (ICA)*, Cairns, Australia, 14-17 June 2009.

Pedersen, J.W., Larsen, L.H., Thirsing, C. & Vezzaro, L. (2020). Reconstruction of corrupted datasets from ammonium-ISE sensors at WRRFs through merging with daily composite samples. *Wat. Res*. 185, 116227.

Rieger, L., Gillot, S., Langergraber, G., Ohtsuki, T., Shaw, A., Takács, I. & Winkler, S. (2013). *Guidelines for using activated sludge models*. IWA Scientific and Technical Report no. 22. IWA Publishing, London, UK.

Roeleveld, P.J. & van Loosdrecht, M.C.M. (2002). Experience with wastewater characterization in the Netherlands. *Water Science & Technology* **45**(6), 77-87.

Sparks, J.A., Vanrolleghem, P.A. & Bott, C.B. (2023). Full scale digital twin with integrated hybrid model predictive controller for ammonia based aeration control. *Proc. 11th IWA Symposium on Modelling and Integrated Assessment,* Québec City, Québec, Canada, 23-27 Sept. 2023.