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

# Developing soft-sensors for a digital twin to mitigate GHG emissions at the Viikinmäki WWTP

H. Haimi*, A. Kiran**, T. Larsson*, K. Blomberg***, F. Elvander**, E. Petäjä*, A. Awaitey*, K. Sahlstedt***, A. Mikola*

* Department of Built Environment, Aalto University, P.O. Box 15200, FI-00076 AALTO, Espoo, Finland, {henri.haimi;timo.larsson;eero.petaja;alexis.awaitey;anna.mikola}@aalto.fi

** Department of Information and Communications Engineering, Aalto University, P.O. Box 15200, FI-00076 AALTO, Espoo, Finland, {anmol.kiran;filip.elvander}@aalto.fi

*** Helsinki Region Environmental Services Authority HSY, P.O. Box 100, FI-00066 HSY, Helsinki, Finland, {kati.blomberg;kristian.sahlstedt}@hsy.fi

**Abstract:** A digital twin to optimize the carbon balance of the Viikinmäki wastewater treatment plant in Finland is being developed in the DIGICARBA project. However, not all the water quality variables required for generating input files of the process model are measured online, or their data quality is insufficient. In this work, we present the design of data-driven soft-sensors to estimate $NH_4$-N and COD concentrations, which will be applied as process model inputs. Ordinary Least Squares regression and autoregressive model with exogeneous variables are investigated for soft-sensor applications. Moreover, the feasibility of dedicated models for different weather conditions is studied.

**Keywords**: Digital Twin; Soft-sensors; Wastewater treatment

## INTRODUCTION

Mitigation of the greenhouse gas (GHG) emissions from the biological wastewater treatment processes has become a major objective for many water utilities that have set goals for achieving carbon neutrality. Digital decision-making tools can support maintaining favorable operative conditions for low process emissions (*e.g.*, Daneshgar et al., 2024). Such decision-making tools are typically based on risk assessment, but not on predictive modelling of GHG emissions.

The target of the DIGICARBA project is developing a digital twin (DT) of the wastewater treatment process at the Viikinmäki plant (1 185 000 P.E.) in Finland. DT distinguishes itself from a simulation model of its existing physical counterpart (*e.g.* a wastewater treatment process) by a continuous, automated data connection (Torfs et al., 2022). The main purpose of the Viikinmäki DT is optimization of wastewater treatment's carbon balance, utilizing the historical nitrous oxide ($N_2O$) emission data. We focus on the work package 1.4 of the project (Figure 1.1) in this abstract.

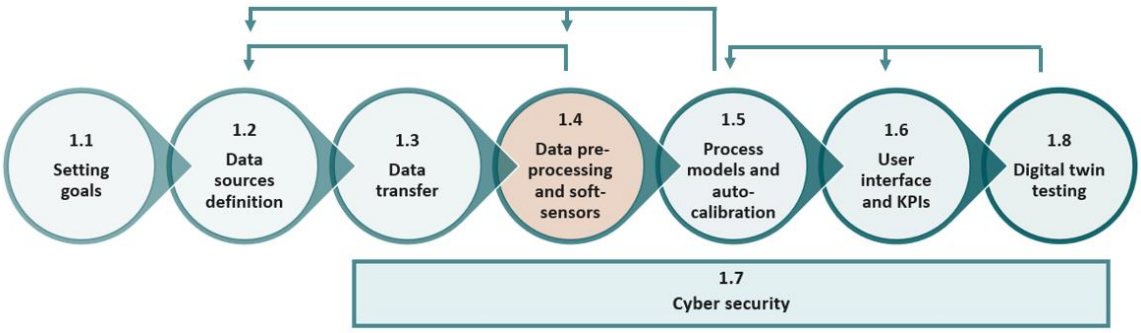


**Figure 1.1** Workplan and the technical work packages of the DIGICARBA project.

In the first development step in DIGICARBA, a DT of one activated sludge process (ASP) line is focused on, and in the future steps the scope of the DT will be extended. The biomechanistic models are constructed using the SUMO software (Dynamita). In this work, we present developing soft-sensors for predicting two key variables: primary effluent $COD_{Cr}$ and $NH_4$-N. Wastewater process data are associated with problems such as missing data, drifting data, constant values, obvious outliers and data outside of the instrument range (Haimi et al., 2013). Therefore, the raw input data of the soft-sensor models are also quality-assessed and reconstructed when necessary.

## PROCESS AND DATA DESCRIPTION

The wastewater treatment line of the Viikinmäki WWTP consists of influent tunnel and pumping, bar screening, grit removal, pre-aeration, primary sedimentation, activated sludge reactors, secondary sedimentation and biological post-filtration. Over 90 % nitrogen removal is achieved as yearly average. The plant is constructed in underground bedrock and equipped with advanced instrumentation including gas and liquid phase $N_2O$. Earlier studies have shown that the ASP is the main source of $N_2O$ emissions in the plant (Mikola et al., 2014; Heinonen et al., 2014).

Historical process (12 variables as hourly averages), laboratory (239 variables, analyses twice a week), and $N_2O$ emission data (from exhaust air pipe and liquid phase in the ASP as hourly averages) were collected from the period of 2021–23 for the purposes of the DT development. In addition, data from an intensive two-week wastewater characterization campaign with 2h- to 4h-composite samples conducted in June 2024 was available.

## METHODS

The soft-sensor development procedure included the following steps: (1) graphical inspection and statistical study of the raw data, (2) input variable selection for soft-sensor models and data reconciliation of the selected variables, (3) modelling approach selection for soft-sensor design, and (4) training and testing the soft-sensor models. Three methods for soft-sensor development are applied: (a) Ordinary least squares (OLS) regression, (b) autoregressive model with exogeneous variables (ARX), and (c) ARX with separate models for dry-weather and wet-weather.

Interquartile range (IQR) method was used for labelling outliers in the raw data, with the bound defined as $Q1 - 1.5 \times IQR$ and the upper bound as $Q3 + 1.5 \times IQR$, where Q1 is the first quartile, Q3 the third quartile and $IQR = Q_3 - Q_1$. After discarding the outliers, missing data were reconstructed with Principal Component Analysis.

OLS is a basic technique to estimate the parameters of the linear regression model (Hastie et al., 2009). The dependent variables (*y)* were $NH_4$-N and $COD_{Cr}$. Lagged values of the predicted variables were used as inputs *x* in OLS models.

ARX is a time-series forecasting method (Box et al., 2015). The term autoregressive refers to the model's ability to use the lagged versions of the dependent variable itself. The exogenous component of the model allows for adding impact of external variables, referred to as exogeneous variables. ARX model is expressed as following: $y_t = c + \sum_{i=1}^{p} \alpha_i y_{t-i} + \sum_{j=1}^{q} \beta_j x_{t-j} + \epsilon_t$, where $y_t$ is the dependent variable at time *t*, $y_{t-i}$ are the autoregressive terms; $x_{t-j}$ are the lagged exogeneous variable at time *t*; $\alpha_i$ are the weights of the autoregressive terms; $\beta_j$ are the weights of the exogeneous terms; $\epsilon_t$ is the error term at time *t*; and *p* and *q* are the total numbers of observations. We used wastewater flowrate to ASP as an external variable for the model to predict $NH_4$-N and primary effluent suspended solids for the model to predict $COD_{Cr}$. The external variables were selected based on correlation analysis and inspection of the data quality.

The procedure for ARX with separate models for different weather conditions was as follows. First, the training data were divided to two different sets based on a threshold for wastewater flowrate to ASP. Next, separate ARX models were trained for high flowrate and low flowrate. Finally, in the testing phase the model applied at time *t* was selected based on the same flowrate threshold as in training. The threshold was defined with equation mean_flowrate + *k* * std_flowrate, where *k* is a tuning parameter. The same external variable was used for the basic ARX models.

The data set was split in ratio of 80/20 for model training and testing data. The goodness of the model predictions is evaluated by computing root mean square error (RMSE), and RMSE normalized with the range (maximum – minimum) and the mean of the testing data set.

## RESULTS AND DISCUSSION

Time series plotting and statistical analysis of the raw process data revealed that the primary effluent $NH_4$-N sensor data quality was not applicable for dynamic process model used in DT development. Therefore, soft-sensing $NH_4$-N was a preferable option, likewise for primary effluent $COD_{Cr}$, that is not analyzed online. Time series of the preliminary soft-sensor estimates are presented in Figure 1.2. Weather-specific ARX models were tested with *k* values ranging from 0 to 1. For $NH_4$-N *k*=0,75 and for COD *k*=1,00 showed the best prediction performances. Training data included two exceptionally high COD peaks for which the model estimates were poor.

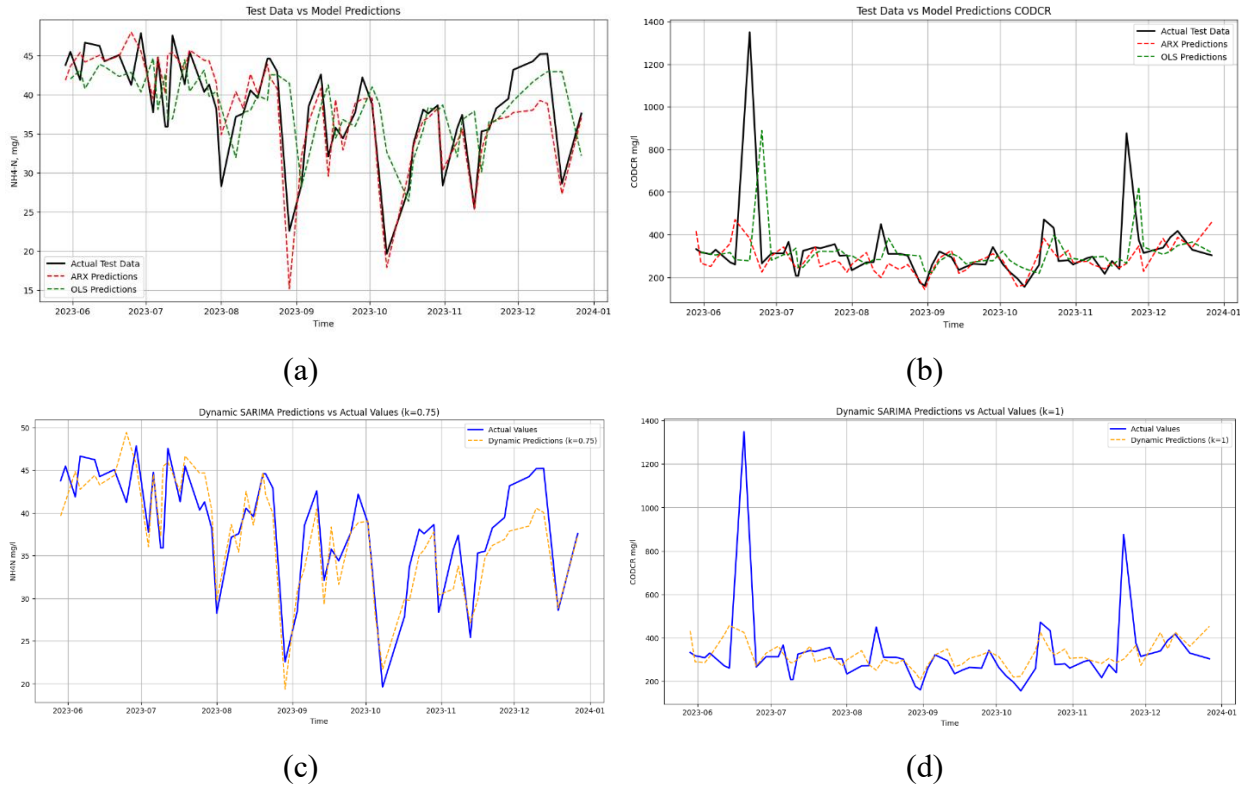


**Figure 1.2** Time series of $NH_4$-N and $COD_{Cr}$ soft-sensor primary effluent concentration estimates with OLS and ARX (a-b), and weather-specific ARX (c-d) models.

The estimation accuracies in the model training sets are provided in Table 1.1. The performance statistics for COD prediction excluding the two high, anomalous peaks are provided in brackets. The most accurate predictions were achieved using ARX model for $NH_4$-N and weather-specific ARX model for COD.

**Table 1.1** Training estimation accuracies for different soft-sensor models.

| **Method** | **$NH_4$-N** | | | **$COD_{Cr}$** | | |
|---|---|---|---|---|---|---|
| | **RMSE, mg/l** | **RMSE/ range, %** | **RMSE/ mean, %** | **RMSE, mg/l** | **RMSE/ range, %** | **RMSE/ mean, %** |
| OLS | 5,9 | 20,1 | 15,4 | 192 (60) | 16,1 (19,0) | 59,7 (20,3) |
| ARX | 3,2 | 11,3 | 8,3 | 161 (66) | 13,5 (21,0) | 50,3 (22,5) |
| Weather-specific ARX | 3,2 | 11,5 | 8,5 | 153 (53) | 12,9 (17,0) | 47,9 (18,2) |

The prediction accuracy especially for COD is not yet sufficient. We continue the soft-sensor development by testing next these alternatives to improve the estimation performance:

1. Systematic optimization of the flow-rate threshold for dry-weather and wet-weather models will be performed to minimize the prediction errors. In addition, another approach of three flowrate ranges for soft-sensor models will be tested.
2. Analysis of the laboratory data indicated that influent wastewater quality differs between weekdays and weekends, presumably due to the share of industrial wastewaters. The influent COD loads are approximately 5 % lower during weekends whereas the influent nitrogen load is on the same level on weekdays and weekends. The typical weekly influent quality pattern will be used to weight the soft-sensor estimates.
3. Influent tunnel water level needs to be lowered each day due to practical reasons. Negative correlation of approximately 0,4 between influent tunnel water level and influent COD was observed during the wastewater characterization campaign. We will test integrating the influent tunnel level data to fine-tune the soft-sensor models.

Detailed results of the afore-mentioned soft-sensor development steps will be presented. In addition, soft-sensors' prediction data will be tested in calibration procedure of the SUMO process model to investigate their applicability for the Viikinmäki DT concept.

**REFERENCES**

Box, G.E.P., Jenkins, G.M., Reinsel, G.C. & Ljung, G.M. 2015 *Time Series Analysis: Forecasting and Control* (5th Edition). John Wiley & Sons, New Jersey.

Daneshgar, S., Amerlinck, Y., Amaral, A., De Mulder, C., Di Nisio, A., Bellandi, G., Gori, R., Caretti, C., Ducci, I., Rehman, U., Porro, J., Nopens, I. & Torfs, E. 2024 An innovative model-based protocol for minimisation of greenhouse gas (GHG) emissions in WRRFs. *Chem. Eng. J.* **483**, 148327.

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

Hastie, T., Tibshirani, R. & Friedman J. 2009 *Elements of Statistical Learning* (2nd edition). Springer, New York.

Heinonen, M., Kosonen, H., Mikola, A. & Fred, T. 2014 Nitrous oxide production in denitrifying post-filtration at the Viikinmäki wastewater treatment plant. In: Proceedings of IWA World Water Congress & Exhibition 2014, 21–26 September, Lisbon, pp. 21–26.

Mikola, A., Heinonen, M., Kosonen, H., Leppänen, M., Rantanen, P. & Vahala, R. 2014 $N_2O$ emissions from secondary clarifiers and their contribution to the total emissions of the WWTP. *Water Sci. Technol.* **70**, 720−728.

Torfs, E., Nicolaï, N., Daneshgar, S., Copp, J.B., Haimi, H., Ikumi, D., Johnson, B., Plosz, B.B., Snowling, S., Townley, L.R., Valverde-Pérez, B., Vanrolleghem, P.A., Vezzaro, L. & Nopens, I. 2022 The transition of WRRF models to digital twin applications. *Water Sci. Technol.* **85**, 2840–2853.