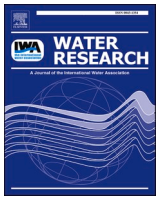
Water Research 286 (2025) 124176

Contents lists available at ScienceDirect

Water Research

journal homepage: www.elsevier.com/locate/watres



Using a hybrid modelling approach for high time-resolution prediction of influent orthophosphate load in a water resource recovery facility

Christoffer Wärff [a,b,*], Bengt Carlsson [c], Magnus Arnell [a,b], Federico Micolucci [d], Oscar Samuelsson [e], Ulf Jeppsson [a]

[a] Division of Industrial Electrical Engineering and Automation (IEA), Department of Biomedical Engineering, Lund University, P.O. Box 118, SE-22100, Lund, Sweden
[b] RISE Research Institutes of Sweden AB, Gibraltargatan 35, SE-41279, Gothenburg, Sweden
[c] Division of Systems and Control, Department of Information Technology, Uppsala University, PO Box 337, SE-75105, Uppsala, Sweden
[d] Nordvästra Skånes Vatten och Avlopp, PO Box 2022, SE-25002, Helsingborg, Sweden
[e] IVL Swedish Environmental Research Institute, Valhallavägen 81, Stockholm, 114 28, Sweden

ARTICLE INFO

Keywords:
Digital twin
Soft sensor
Optimisation

ABSTRACT

Water resource recovery facilities face challenges with increasingly stringent effluent demands, complexity and demand for capacity increasing investments. Emerging technologies such as digital twins could alleviate these problems but require high frequency influent data. This work presents a method for utilising measurements in the primary clarifier effluent with a model of the processes between the influent and primary clarifier effluent to predict influent orthophosphate load for a plant with considerable internal load. Five functions for describing daily load variations were tested and compared for accuracy and computational time. All functions were shown to reproduce the measured primary effluent orthophosphate concentration with high accuracy, although the function based on four normal distributions was deemed the most suitable due to its short computational time, realistic influent concentration variations and accurate estimated primary effluent orthophosphate concentration. Validation of the optimised influent concentrations shows that it follows similar patterns but might over-predict the afternoon load, which could be due to deviating daily patterns by inhabitants during the COVID-19 pandemic (although this requires further investigation). The presented methodology can be extended also to estimate influent COD-fractions, automate plant calibration and optimise plant performance.

1. Introduction

Water resource recovery facilities (WRRFs) face ongoing challenges related to more stringent effluent demands, increasing complexity and demand for capacity increasing investments. These challenges can be alleviated by new technologies and digitalization, where one such emerging technology is digital twins. With digital twins, process simulation models are used with automated data transfer from the real plant to run simulations automatically at given intervals (in near real-time) with pre-defined objectives (Torfs et al., 2022). The use of digital twins at WRRFs does, however, come with considerable challenges of their own.

One challenge is obtaining high quality and high frequency (i.e., hourly, or higher) influent data. These are required to keep the digital twin up to date with the latest data and to construct influent generators or forecasting models for scenario analysis. While influent flow rate is usually measured at high frequency (minutes), concentrations of chemical constituents, such as chemical oxygen demand (COD), ammonium nitrogen and orthophosphate phosphorus, are rarely measured with such high frequency. Rather, they are measured by daily composite samples. These samples lack the temporal frequency required for accurate dynamic simulation of plant process models that must be known or estimated for use with digital twins. The required frequency depends on the objective of the digital twin, but often time resolutions of 15 to 120 min are used to evaluate controller performance and estimate peak values (Jeppsson et al., 2007; Alex, 2024) in activated sludge models.

Installing new sensors or analysers to measure, e.g., ammonium or orthophosphate, in the influent require both an investment cost and considerable time for sensor maintenance (Rieger et al., 2010). For

* Corresponding author.
E-mail addresses: christoffer.warff@iea.lth.se (C. Wärff), bengt.carlsson@it.uu.se (B. Carlsson), magnus.arnell@iea.lth.se (M. Arnell), federico.micolucci@nsva.se (F. Micolucci), oscar.samuelsson@ivl.se (O. Samuelsson), ulf.jeppsson@iea.lth.se (U. Jeppsson).

https://doi.org/10.1016/j.watres.2025.124176
Received 28 November 2024; Received in revised form 2 July 2025; Accepted 6 July 2025
Available online 7 July 2025

influent measurements in particular, the measurement environment is challenging due to the characteristics of the raw sewage with many particles and colloids that quickly clog the pre-filters of chemical analysers. Furthermore, the high organic content of the wastewater lead to biofilm growth on sensors. Thus, methods for estimating the influent composition with high frequency that do not require installation of new sensors would reduce both investment and maintenance costs.

Several approaches have been suggested to estimate high time-resolution influent data without relying on (regular) direct measurements. As outlined by Martin and Vanrolleghem (2014), these approaches can be divided into three categories, which are based on: 1) typical load patterns such as daily, weekly or yearly influent profiles. This approach can also include correction factors for wet weather events. 2) harmonic functions (e.g., Fourier series). These are used to describe the periodicity of different phenomena (such as influent load variations), but unlike approach 1, where measured data is used directly to obtain the patterns, they are described by a mathematical function. Both approach 1 and 2 can be used to create a higher time frequency resolution from low resolution measurements. 3) phenomenological models. These models aim to capture the observed influent behaviour by describing the driving mechanisms, i.e., generation of wastewater at the source (households/industries), transport in the sewer system, and impact from precipitation. Phenomenological models often use a mix of mechanistic and empirical sub-models as well as typical daily profiles. Another category can be defined, which is not mentioned by Martin and Vanrolleghem (2014): 4) data driven or hybrid (data driven/mechanistic) methods, such as machine learning models.

An example of a Category 1 influent generator can be found in Langeveld et al. (2017). This influent generator used measured influent flow rate and typical dry weather concentration profiles to estimate influent concentrations, incorporating dilution, first flush and recovery processes during wet weather conditions (with separate processes for different sizes of rain events). It was used by Daneshgar et al. (2024) during development of a digital twin for the WRRF in Eindhoven, The Netherlands, where influent data for the twin was generated with 2-hour time-resolution. Langeveld et al. (2017) showed that this approach can be effective to augment measurements and replace faulty sensor data, but it still requires extensive influent measurements during dry and wet weather events to capture the relevant dynamics.

Alex (2024) developed a Category 2 method (extending the work of Langergraber et al. (2008)) to estimate influent concentration variations using only high time-resolution flow measurements and typical routine WRRF data (i.e., influent daily composite samples). The method divides the influent wastewater into categories with different characteristics and use Fourier series to describe periodic patterns. The method was shown to accurately reproduce high frequency influent data by combining the calibrated influent model with daily composite samples, for both dry weather and wet weather events. It does, however, require daily composite samples for all days where the high-resolution data should be generated. For days where this is not available (most plants do not do this every day), the method requires assumptions of the load.

Gernaey et al. (2011) presented a phenomenological model-based method (Category 3) using generation of wastewater flow and pollutant loads from households and industries and incorporating rainfall and sewer transport processes to generate realistic influent concentration variations. The model was used to generate influent data for the Benchmark Simulation Model No 2 (Jeppsson et al., 2007) but could also be adapted for use in a real plant. Compared to the other methods to generate influent data, this approach potentially requires a more substantial effort due to the need to model the sewer system in greater detail. The lack of detailed information of household and industry wastewater discharges also limits the accuracy of the results.

For Category 4, pure data driven methods, such as different machine learning models (e.g., long short-term memory (LSTM) models), have also been used to generate influent data with high frequency (Li and Vanrolleghem, 2022a, 2022b). However, they require long time-series of training data, which is not always required by the other mentioned methods. To overcome this data requirement, recent publications have showcased a hybrid modelling approach to generate influent data, where a mechanistic model of a process is combined with measurements of the effluent from the process. The influent concentration is then iterated and optimised until the model output matches the measured data. Johnson et al. (2021) used this approach to estimate high frequency influent ammonium nitrogen data based on a process model of a primary clarifier and measured data from the primary effluent. By combining the optimised concentration data with typical ratios between measured influent compounds, a full influent dataset could be constructed for a digital twin model. This concept was further developed by Johnson et al. (2024) to include also the activated sludge reactors in the process model and optimising the influent Total Kjeldahl Nitrogen (TKN) concentration to fit the measured air flow rate to the process. Other variables were calculated through typical influent ratios. The estimated daily flow weighted average error (normalised RMSE) was reported between 10 and 25 % for ammonium, TKN, orthophosphate, total phosphorus and filtered and total chemical oxygen demand (COD) for a full year of estimation for three WRRFs (note that not all the listed parameters were reported for all plants). Zorrilla et al. (2024) used a similar concept to estimate substrate characteristics for a (virtual) biogas plant using the Anaerobic Digestion Model No 1, although with much slower dynamics (daily or weekly averages) due to the slower processes of anaerobic digestion. Zorilla et al. could not accurately predict the complete influent characteristics, but they report that the results were sensitive to the variables they wanted to identify.

A benefit with the presented hybrid approach, compared to the other described categories, is that current, already available, high-frequency measurements can be utilised without the need to install new sensors at the influent (or through installing sensors downstream the influent at a more favourable location). However, an important choice must be made regarding which type of mathematical function that should be used to describe the diurnal influent variations during the optimisation period. Such functions have been presented in several studies. The previously mentioned Fourier series have been used by several authors with different types of models (Langergraber et al., 2008; Mannina et al., 2011; Li and Vanrolleghem, 2022a; Alex, 2024). Another approach was used by Sitzenfrei et al. (2017), who used a combination of three normal distributions to describe the probability density function for water use in households over the course of a day. This concept was extended to four normal distributions by Wärff et al. (2020) to accurate predict nighttime water use. Other simple methods such as polynomials could also potentially be used to describe the variations, or the load value at each time step could be optimised directly. Both the number of function parameters and the resulting shape of the output will affect the flexibility of the function to adapt to different influent profiles and the required optimisation time.

While the presented mathematical functions are good candidates for describing the influent load variations, the choice of which one to use impacts both the optimisation time and prediction accuracy (both vital components when used in an automated setting with digital twins). There is a lack of research and understanding of the impact of different functions on these metrics, which makes the method difficult to implement in practice. Furthermore, the previous studies by Johnson et al. (2021) and Johnson et al. (2024) have focused on primarily predicting nitrogen variables (ammonium and TKN), then calculating other variables such as phosphorus and COD through ratios to the predicted nitrogen variable. Direct prediction of other variables, such as orthophosphate, has not yet been demonstrated in the literature. This is important for wider validation of the methodology, as phosphorus behave differently than nitrogen in WRRFs and are affected by both biological and chemical processes (which makes modelling challenging). At the same time, the effluent requirements can be low, with new permits in Sweden often requiring annual average effluent concentrations in the proximity of 0.20 mg P/L as total phosphorus. This

demands accurate influent data for model-based decision-making in digital twins.

In this study, we aim to address two research gaps for using a Category 4 hybrid model-based software (soft) sensor as part of WRRF influent generation: 1) Validate the methodology for direct prediction of orthophosphate, based on primary clarifier effluent measurements and a model of upstream processes. 2) Compare mathematical functions for describing the daily influent variations and evaluate them on predicted primary effluent concentration accuracy and optimisation time. We address these questions by implementing a soft sensor at the Öresundsverket WRRF, Helsingborg, Sweden, to find the function best suited for use in the soft sensor.

2. Material and methods

The method section is structured as follows. The first part (2.1) is dedicated to the question of validating the methodology for orthophosphate prediction, with details of the WRRF under study, data requirements and model structure, model construction and data collection and treatment. The second part (2.2) is dedicated to the question of comparing functions for describing the influent load variations, with description of mathematical function tested, the optimisation algorithm used and a description of evaluation criteria for comparing of the results.

2.1. Construction of a model-based optimisation

2.1.1. Implementation case study: Öresundsverket water resource recovery facility

The Öresundsverket water resource recovery facility in Helsingborg, Sweden, treats wastewater from approximately 180 000 person equivalents (with a mean daily flow rate of 54 000 m^3/d during 2024) before it is discharged in the Öresund strait. The water train processes include grit chambers, wet weather detention basin, primary clarifiers, activated sludge process, secondary clarifiers, and sand filters for final polishing. The sludge train consists of gravity thickeners (separate for primary and secondary sludge), mesophilic anaerobic digestion and sludge dewatering. The activated sludge process is operated with enhanced biological phosphorus removal (EBPR), (occasional) dosing of ferric chloride for simultaneous precipitation of phosphorus, and predenitrification and nitrification. Ferric chloride is also dosed to the primary sludge before thickening. The influent contains low amounts of volatile fatty acids (VFA), so to produce such for the EBPR process, hydrolysis and fermentation of primary sludge is performed in the sludge pockets of the primary clarifiers. The sludge is then resuspended through a pump and settled again, while the VFA is flushed with the water to the activated sludge process. The primary clarifiers and the activated sludge process consist of four parallel trains with separate sludge settling systems. In each train, two primary clarifiers are operating in parallel. Due to the process configuration, a substantial internal orthophosphate load is generated in the plant, (mainly) through the anaerobic digestion process and the subsequent sludge dewatering reject water that is led back to the inlet. Orthophosphate is also released through the hydrolysis process in the primary clarifiers. The orthophosphate concentration in the primary effluent is therefore considerably higher than in the influent.

2.1.2. Structure and data requirements for model-based optimisation

When constructing a local model over the primary clarifier for estimating the influent concentration, two input data streams are required: raw influent and the sum of internal load (from recycle streams such as sludge dewater reject). The internal streams can be either measured directly or estimated from a model (or a combination of both). To make the setup as readily available as possible, it is desirable to use as much real plant data as possible. This allows setting up such a model-based optimisation without the need of a plantwide model (PWM), i.e., a whole-plant model including both liquid and solids processes, for generating data for certain streams. After initial tests, however, it was concluded that although the sludge dewatering reject can contain most of the internal P load in dissolved form, other internal loads and processes (such as primary sludge hydrolysis) can still have an impact on the primary effluent orthophosphate concentration. Therefore, it was decided to use three influent sources for the model-based optimisation:

1. Characterised raw influent concentrations and flow rate.
2. Sludge dewatering reject flow rate and orthophosphate concentration.
3. Model prediction of remaining internal load.

This means that a plantwide model initially calibrated with an estimated influent orthophosphate concentration is required. The combined output of the internal loads from this model can then be used as input data for the optimisation. If measurements of the sludge dewatering reject water flow rate and relevant concentrations are not available, although it often is at medium to large size plants in Sweden (on the contrary to other internal loads, which are usually not measured), model output of those can also be used.

A daily optimisation routine was chosen, where the influent orthophosphate load was optimised over 24 h at a time (between 00.00–23.59). This allowed using several previously developed mathematical functions describing the shape of the diurnal variation and lines up with the time during which daily composite samples are often collected. The load was chosen as target variable rather than the concentration, due to the effect of dilution events on the shape of the diurnal patterns. The temporal resolution was set to hourly values, although a higher resolution can be achieved through interpolation.

For model simulations, the Sumo simulation platform (version 22.1.0, Dynamita, France) was chosen along with the Digital Twin toolkit. Data handling and optimisation simulations were run through Python (version 3.9.13).

2.1.3. Plantwide model construction

For the initial evaluation of the plant, a PWM was developed in Sumo. Influent wastewater characterization was performed according to the STOWA protocol (Roeleveld and van Loosdrecht, 2002) for raw and filtered (1.2 μm glass fiber filter) wastewater to determine influent fractions. Respirometry was used to determine the fraction of heterotrophic biomass in the influent, according to the method by Wentzel et al. (1995). Dynamic daily profiles for influent load of total chemical oxygen demand (COD), dissolved COD and ammonium nitrogen were established from influent sensor data (WTW CarboVis 701 IQ TS and WTW AmmoLyt Plus SET/Comp). The model was calibrated and simulated with data from 2021 to 2022 to produce values on unmeasured internal loads.

2.1.4. Simplified model for optimisation

To increase simulation speed, a simplified model was constructed in Sumo, using the Sumo1 biokinetic model (Fig. 1). It includes the primary clarifier with a recycle flow for sludge resuspension, the processes upstream the primary clarifier (grit chamber and detention basin) and inputs (raw influent, sludge dewatering reject and sum of remaining internal loads). The primary clarifier model was based on the Sumo 3-compartment model but modified so that no elutriation flow (which is included in the original model to account for VFA washout from the sludge blanket) occurs. Instead, a pumped flow for resuspending sludge to the inlet was built into the model through pumping from the sludge blanket to the inlet. The clear liquid zone was modified to use a plug-flow reactor unit instead of a CSTR, allowing the user to decide on the number of tanks in series to use to model the clear water zone. Although this allows for describing more complex hydraulic behaviour, in the end the number of tanks in series was chosen as one (equal to a single CSTR) as this was deemed to produce the best results for this application. This primary clarifier model was coded as a complex unit using the

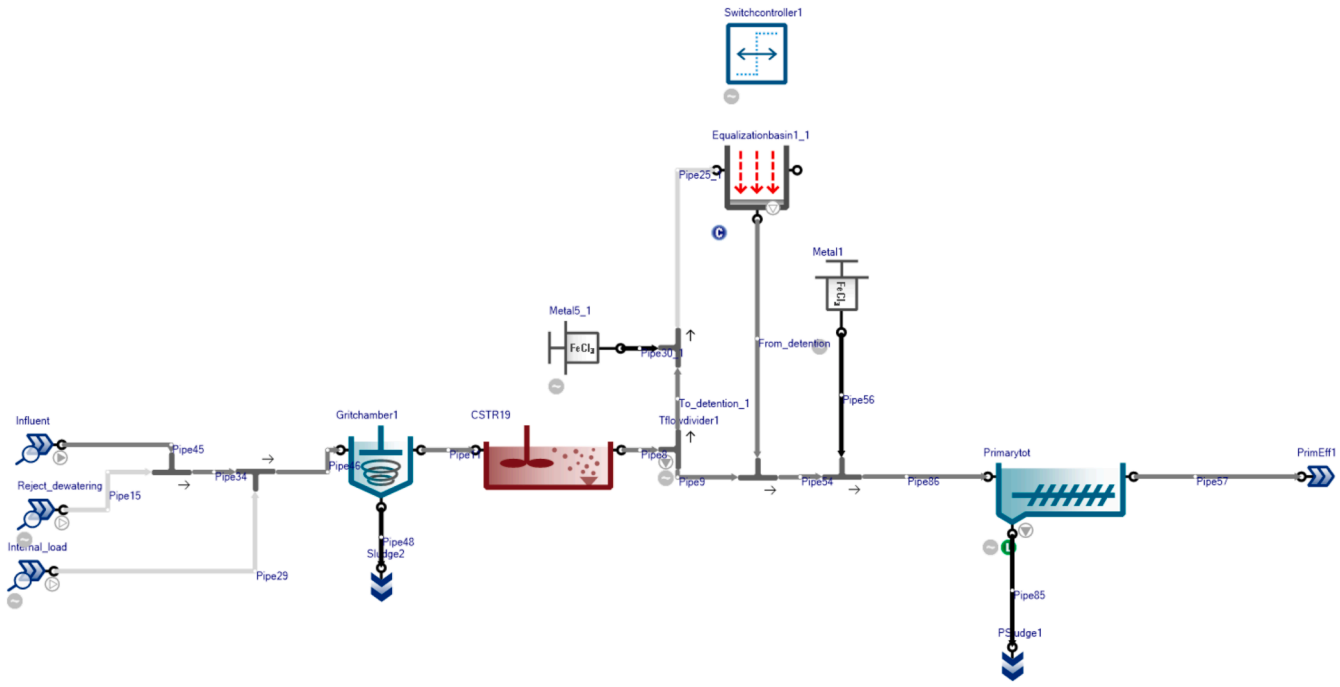


Fig. 1. Model of the primary treatment at Öresundsverket, Helsingborg, Sweden, used in the optimisation.

SumoSlang language.

2.1.5. *Dewatering reject flow rate estimation*

The sludge dewatering reject water flow rate is not measured directly at the plant. The sludge flow rate to the dewatering equipment and the measured weight of the dewatered sludge is, however, measured. Based on this, the dewatering reject flow rate could be estimated based on a mass balance over the dewatering equipment. For this calculation, the dewatered sludge was assumed to have a density equal to water. In reality, the density of the dewatered sludge will probably be slightly higher due to higher density of the solids (O'Kelly, 2005), but since the actual density was not measured and the difference on the calculated reject flow rate is estimated to be about 1 percent, the error was considered negligible. The full calculation is shown in the Supplementary materials.

The orthophosphate concentration in the anaerobic digester is analysed once per week at the laboratory at the plant. However, while the concentration of a dissolved constituent in the water phase should not change between the inlet to the dewatering equipment and the reject water, changes in factors affecting the chemistry (such as decreasing water temperature between the anaerobic digestor and dewatering) can create struvite precipitation during dewatering (Achilleos et al., 2022). Orthophosphate is not measured directly in the dewatering reject water at the plant, so instead measurements of filtered (filter with pore size 6–10 μm) total phosphorus in the reject water was assumed to equal the orthophosphate concentration in this stream. This might cause a slight overestimation of the internal orthophosphate load if there is colloidal phosphorus that is not contained in the filter. It was further assumed that this concentration changes slowly and remains constant over a given day and that the values between the measured data points could be calculated through interpolation. From this, the calculated or measured daily values could be combined with the estimated dewatering reject water flow rate to obtain the orthophosphate load.

2.1.6. *Primary effluent orthophosphate data and pre-treatment*

Orthophosphate concentration was measured by an online analyser (2029 process photometer, Metrohm) in the primary effluent from each of the four parallel trains at the plant. Flow rate to the inlet of each treatment train (before primary clarifiers) was measured by Parshall flumes. Hourly average values over 40 days (May 1st – June 9th, 2021) were extracted for further processing.

A single primary effluent orthophosphate time series was created through flow proportional averaging of the different primary effluent concentration time series. Since the primary effluent or sludge flow rate were not measured individually for each primary clarifier, the primary influent flow rate was used for the averaging. The primary sludge flow rate is usually only about 0.4 percent of the primary influent and the difference between influent and effluent flow rate is thus small.

As the measurements are grab samples, they appear noisy at times. Initial results with the direct method (not shown) also indicated problems with noise in the optimised influent values. Therefore, the combined primary effluent data were pre-treated by filtering through a Savitzky-Golay filter (Savitzky and Golay, 1964), using the SciPy package (Virtanen et al., 2020) and the signal.savgol_filter method in Python. A second-order filter and a time window of 7 samples were used.

2.1.7. *Validation data*

Validation data for the optimised influent concentrations were collected through flow proportional sampling in the influent for three days. For two of the days, orthophosphate was analysed (Hach LCK348) in hourly samples. For the third day, the samples were filtered and sent to an external accredited laboratory for analysis of orthophosphate. For all measured concentration values, the corresponding hourly load values [kg P/d] over 24 h were calculated and compared against the load variations estimated through the optimisation. The validation data was collected during 2023, but due to bad data quality from the orthophosphate analyser during those days the influent estimation could not be done with this data. Instead, the data from 2021 had to be used for the optimisations. The validation data was therefore used to compare trends in diurnal load variations.

2.2. *Modelling the diurnal load variations*

2.2.1. *Methods for mathematical description of the influent load*

Five different mathematical functions from the literature for describing the daily load variations were selected for evaluation. In

addition, a baseline method was included for comparison. Details of each function is described below. The time t is set between 0 and 23 for one full day, in all equations.

Function 1 – Sum of 3 normal distributions

Function 1 is based on the work from Sitzenfrei et al. (2017) and includes 7 parameters: 3 mean values, $\boldsymbol{\mu_n}$, 3 standard deviations, $\boldsymbol{\sigma_n}$, and 1 scaling parameter, $\boldsymbol{c}$. It is described mathematically as):

$$f(\boldsymbol{t}) = \sum_{\boldsymbol{n}=1}^{3} \frac{\boldsymbol{c}}{\boldsymbol{\sigma_n}\sqrt{2\boldsymbol{\pi}}} \boldsymbol{e}^{-\frac{(\boldsymbol{t}-\boldsymbol{\mu_n})^2}{2\boldsymbol{\sigma_n^2}}} \tag{1}$$

Function 2 – Sum of 4 normal distributions

Function 2 is identical to function 1 but with the load described by a sum of four normal distributions instead of three, based on Wärff et al. (2020). It has 9 parameters: 4 mean values, $\boldsymbol{\mu_n}$, 4 standard deviations, $\boldsymbol{\sigma_n}$, and 1 scaling parameter, $\boldsymbol{c}$. It is described mathematically as

$$f(\boldsymbol{t}) = \sum_{\boldsymbol{n}=1}^{4} \frac{\boldsymbol{c}}{\boldsymbol{\sigma_n}\sqrt{2\boldsymbol{\pi}}} \boldsymbol{e}^{-\frac{(\boldsymbol{t}-\boldsymbol{\mu_n})^2}{2\boldsymbol{\sigma_n^2}}} \tag{2}$$

Function 3 – Direct method

Function 3 has 24 parameters, where each parameter is the load value [g P/d] during one of the 24 h. Thus, the parameters for this method are also the output of the function.

Function 4 – 6th order polynomial

Function 4 is a 6th order polynomial, meaning that the function has 7 parameters ($\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{c}$, $\boldsymbol{d}$, $\boldsymbol{e}$, $\boldsymbol{f}$ and $\boldsymbol{g}$) and is described as

$$f(\boldsymbol{t}) = a\frac{\boldsymbol{t}}{24}^6 + b\frac{\boldsymbol{t}}{24}^5 + c\frac{\boldsymbol{t}}{24}^4 + d\frac{\boldsymbol{t}}{24}^3 + e\frac{\boldsymbol{t}}{24}^2 + f\frac{\boldsymbol{t}}{24} + g \tag{3}$$

Function 5 – Fourier series

Function 5 is based on a Fourier series as defined by Mannina et al. (2011). The function has 10 parameters ($\boldsymbol{\mu}$, $\boldsymbol{\beta}_1$, $\boldsymbol{\omega}_1$, ϕ_1, $\boldsymbol{\beta}_2$, $\boldsymbol{\omega}_2$, ϕ_2, $\boldsymbol{\beta}_3$, $\boldsymbol{\omega}_3$ and ϕ_3) and is described as (4):

$$f(\boldsymbol{t}) = \mu\left(1 - \left(\boldsymbol{\beta}_1 \sin\left(\boldsymbol{\omega}_1\frac{\boldsymbol{t}}{24} + \phi_1\right) + \boldsymbol{\beta}_2 \sin\left(2\boldsymbol{\omega}_2\frac{\boldsymbol{t}}{24} + \phi_2\right) + \boldsymbol{\beta}_3 \sin\left(3\boldsymbol{\omega}_3\frac{\boldsymbol{t}}{24} + \phi_3\right)\right)\right) \tag{4}$$

Baseline

A baseline method was included to compare the output when using the optimisation with different functions and determine if and how much they improve predictions of the primary effluent orthophosphate concentration. The baseline scenario was defined as a fixed daily orthophosphate load profile based on the hourly mean values from the collected validation data, where each load value was divided by the measured flow rate to obtain the influent concentration values. This scenario is a simple alternative to detailed influent measurements to still utilise realistic dynamic variations at the plant, which could be useful for example for less detailed modelling studies. It was therefore deemed suitable for use as a baseline case.

2.2.2. Optimisation algorithm

The Nelder-Mead (Nelder and Mead, 1965) algorithm (in Python, using the SciPy-package) was used for optimisation. To initialise the parameters for the mathematical load descriptions, the baseline influent orthophosphate load profile [kg P/d] over 24 h was used. The baseline load profile was used directly as initial parameter values for function 3. For the other functions, the parameters for each function were optimised to fit the calculated load profile to the averaged measured load profile. The optimised parameters were then used as starting point parameters for each optimisation day.

When continuous data are optimised one day at a time, there is a risk that there is an artificial "jump" at the transition between days. This is showcased for function 4 (although it occurs for all parametric methods to some extent) in Figure S.2 in Supplementary materials, where the rate of change between data points often are often 2–6 times higher than normal at the transitions between days. One way to minimise this "jump" is by including datapoints before and/or after the day to be optimised in the objective function. After initial tests (data not shown) with either including four data points before the day to be optimised, two data points before and after, or four data points before and after, it was concluded that using four data points before and after gave the best results and was used for further analysis. As the four datapoints after the day to be optimised are not known, the load during these hours was assumed to be equal to the load during 00.00 – 04.00 the current day. The four data point for before the day to be optimised were saved from the final values from 20.00 – 23.59 from the last optimisation day. For the first optimisation day of the series, the baseline load was used together with the data from the first day in the series to be optimised. This day was then looped for 10 days to reach a pseudo steady state before starting the optimisation of the first day. When the optimisation of each day was finished, the final parameter set was used to run the simulated day once more. The simulation was paused at 20.00 and the state variables were saved for use as initialisation for the next day of optimisation.

For the evaluation, the simulated orthophosphate output for the 32-hourly values of the current optimisation (24 h + 4 h before the current day + 4 h after the current day) were extracted. The measured and pre-treated primary effluent orthophosphate data were extracted for the corresponding time stamps. An objective function was defined as the root mean square error (RMSE) between the simulated and measured primary effluent orthophosphate over the evaluation period.

The optimisation was done in the following steps (schematically shown in Fig. 2):

1. the Sumo model was loaded and initialised;
2. the parameters to describe the influent orthophosphate load were chosen;
3. a simulation was run, and the objective function calculated.

The optimisation procedure repeats the steps above until convergence. To assure accurate results, the maximum number of iterations were set high enough so that the limit was never reached for any of the methods (400 times the number of parameters). To assure that no negative values occurred, a final check of the load profile was done, and negative values set to 10^{-3} g P/d. For the direct method, non-negative bounds were set for the parameters as they are also the output (for the other methods negative values of the parameters could still produce positive values for the output). No other constraints were set for the parameters in the optimisation. The optimisation was run for the data spanning 40 days of operation of the plant.

All optimisations were run on a Windows PC (AMD Ryzen 7 3800X, 8 core 3.89 GHz, 32.0 GM RAM).

2.2.3. Comparison and evaluation of results using different mathematical functions

For a method to be useful and usable in a near real time digital twin soft sensor, it must provide accurate results and be reasonably fast. To determine the most suitable function for this application, we evaluated

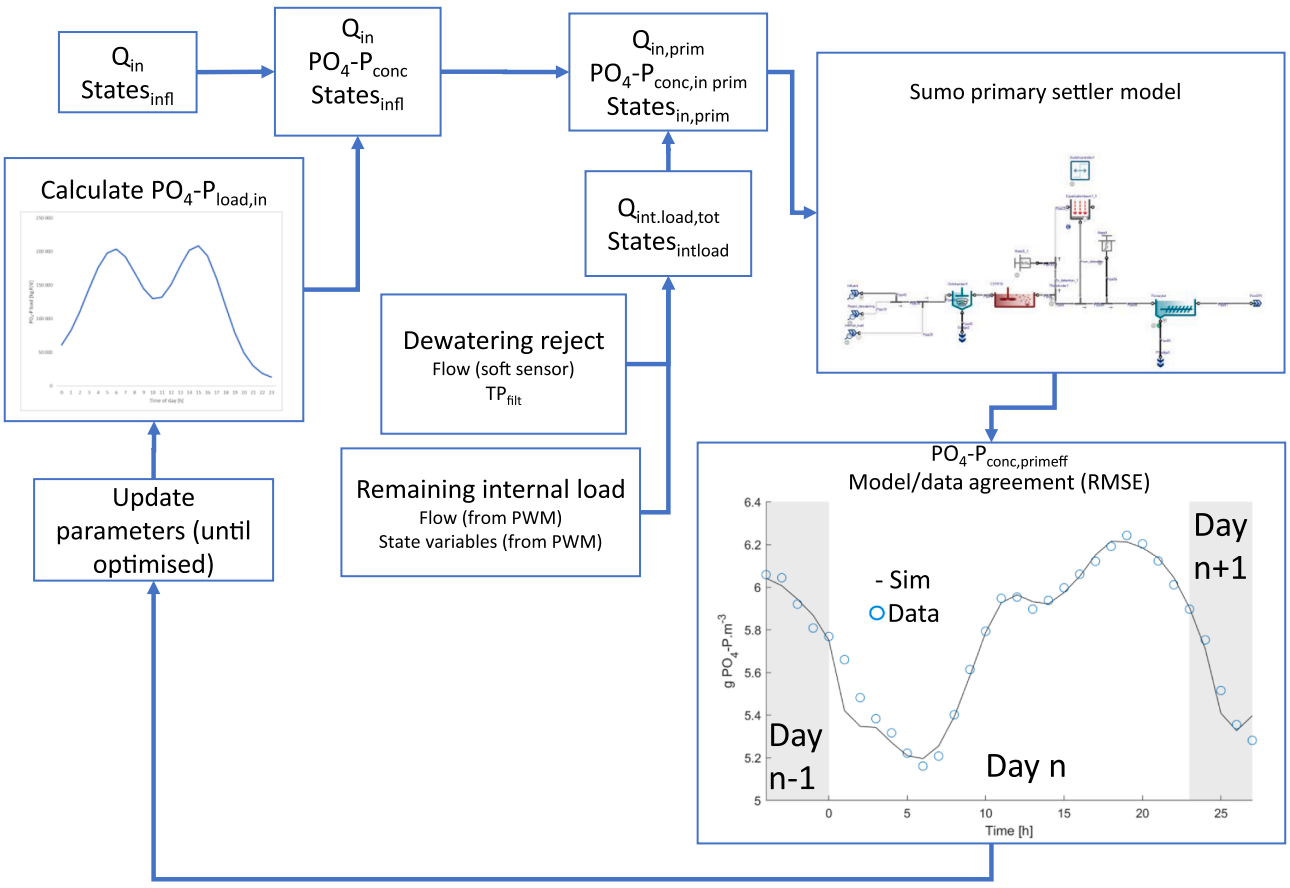


Fig. 2. Schematic overview of the influent orthophosphate soft sensor optimisation setup. PWM = plantwide model. The plot in the down right corner displays how the RMSE is calculated during optimisation, where four datapoints before and after the current day *n* are used. For the comparison of results later in the paper, only the values from day *n* is used.

the results using the following criteria: 1) if the resulting concentration variations appear realistic based on expert knowledge, i.e., with typical diurnal variation patterns and without artificial spikes; 2) by calculating the RMSE for each day in the time series (from the hourly values), then comparing the median daily RMSE between the functions (the median was selected due to the statistical methods used, see below for details); 3) by calculating the average time for the optimisation to complete (normalised with the time for the function with lowest RMSE). Finally, the output from the most suitable method is compared to the validation data.

A statistical analysis was used to better be able to distinguish between the RMSE obtained using the different functions and evaluate if the differences were statistically significant. See Supplementary materials för more details.

3. Results and discussion

3.1. Differences between mathematical functions

Fig. 3 shows the results from the optimisation of the influent orthophosphate load, including the measured and model predicted primary clarifier effluent concentration as well as the optimised plant influent concentration. A zoomed in version is shown in Fig. 4, to better visualise the difference between the functions. Table 1 shows the results from the evaluation criteria for each method, with simulation time shown relative to function 2 since this function had the lowest median RMSE. All functions using optimisation (Fig. 3(a)-(e)) could reproduce the primary effluent concentration with high accuracy, as evident by the highest median daily RMSE at 0.24 and the target primary effluent orthophosphate concentration statistics according to Table 2.

Function 2 (4 normal distributions) resulted in the lowest median daily RMSE, only slightly lower than function 3 (direct). Functions 1 (3 normal distributions) and 5 (Fourier series) had similar RMSE. with slightly higher standard deviation for function 1. Function 4 (polynomial) had the highest RMSE. The R^2 between model and data indicate the same pattern (with a negative value for the baseline scenario, meaning that the prediction is worse than using the mean of the data during the period). When analysing the generated influent time series, the results from most functions appear realistic in shape from what is expected, i.e., a distinct diurnal pattern. All parametric functions to some degree, but in particular function 4, exhibit occasional spikes in the data that seem to appear mainly at the transition between days (even with the efforts to minimise such effects). Further treatment of the data to remove such spikes can of course be done, for example by using a noise removing filter such as the Sovitzky-Golay filter, or through a rate of change analysis and removal of suspected erroneous data points followed by linear interpolation to fill the gaps. Since this system has a high internal recycle rate, one hourly load value will affect the output during several subsequent hours afterwards. If a high spike is removed from the optimised dataset, the following values would probably need to be increased to avoid too low primary effluent concentration. The same effect can probably explain why the resulting influent data from function 3 becomes very noisy, since a too high load value in one hour can be compensated for by a low load value in the next hour. In addition, a higher order model as in function 3 (as many parameters as data points) will result in less dampening of noise in the measured data. This is likely avoidable with the parametric functions since they are forced to adhere to a certain shape of the load profile (although the problem with

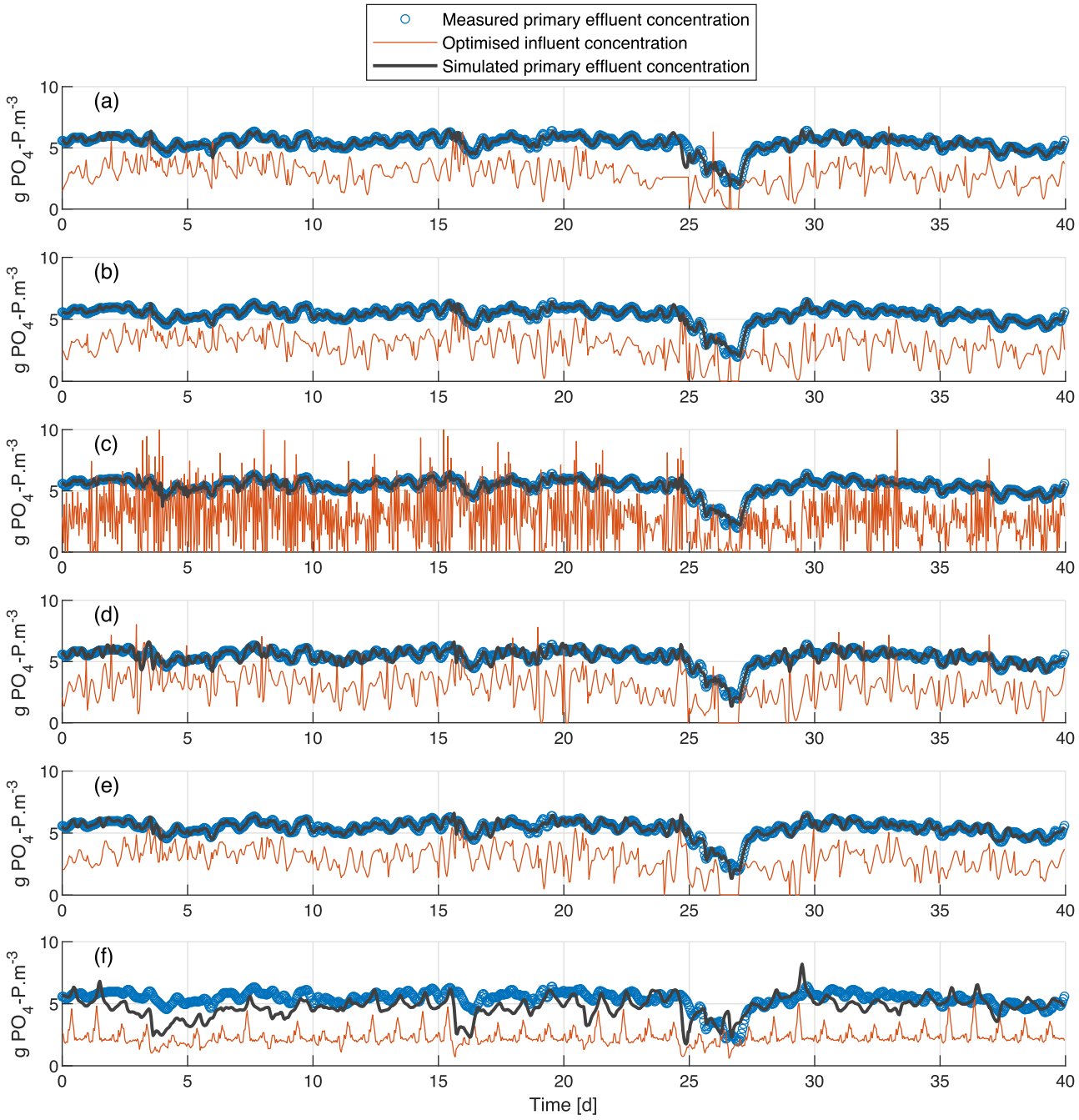


Fig. 3. Optimised influent orthophosphate concentrations with simulated and measured primary effluent concentrations for five different functions: (a) function 1: 3 normal distributions; (b) function 2: 4 normal distributions; (c) function 3: direct function; (d) function 4: 6th order polynomial; (e) function 5: Fourier series; and (f) baseline.

transition between days can remain). A way to possibly avoid the aforementioned problems while still using the direct function is to estimate the load values one hour at a time sequentially. In fact, this is a plausible scenario for use in a digital twin, where the value is updated once per hour. However, preliminary tests (data not shown) with the dataset in this paper showed this option to be unstable with optimised hourly load values being very high in one timestep and in the next timestep zero. This option requires more research to be explored in detail. The methods in this paper can still be used to generate time series data, which in turn can be used to fit an influent generator model for use with higher frequency simulations.

The statistical analysis (see Supplementary materials) showed that the RMSE medians of functions 2 and 3, as well as functions 1 and 5, were not statistically significant. We can thus not say that function 2 performed better than function 3, or that function 1 performed better than function 5, in terms of RMSE. The rest of the comparisons between the functions showed statistically significant differences.

3.2. *Wet weather performance*

The simulated period contains several wet weather events (see Supplementary material, Figure S.3). During many of these, the concentration is diluted substantially in the baseline simulations but, counter intuitively, this is not appearing to occur to the same degree in the observed data. This results in substantial load peaks in during several of the rain events when using the optimisation methods. However, not in

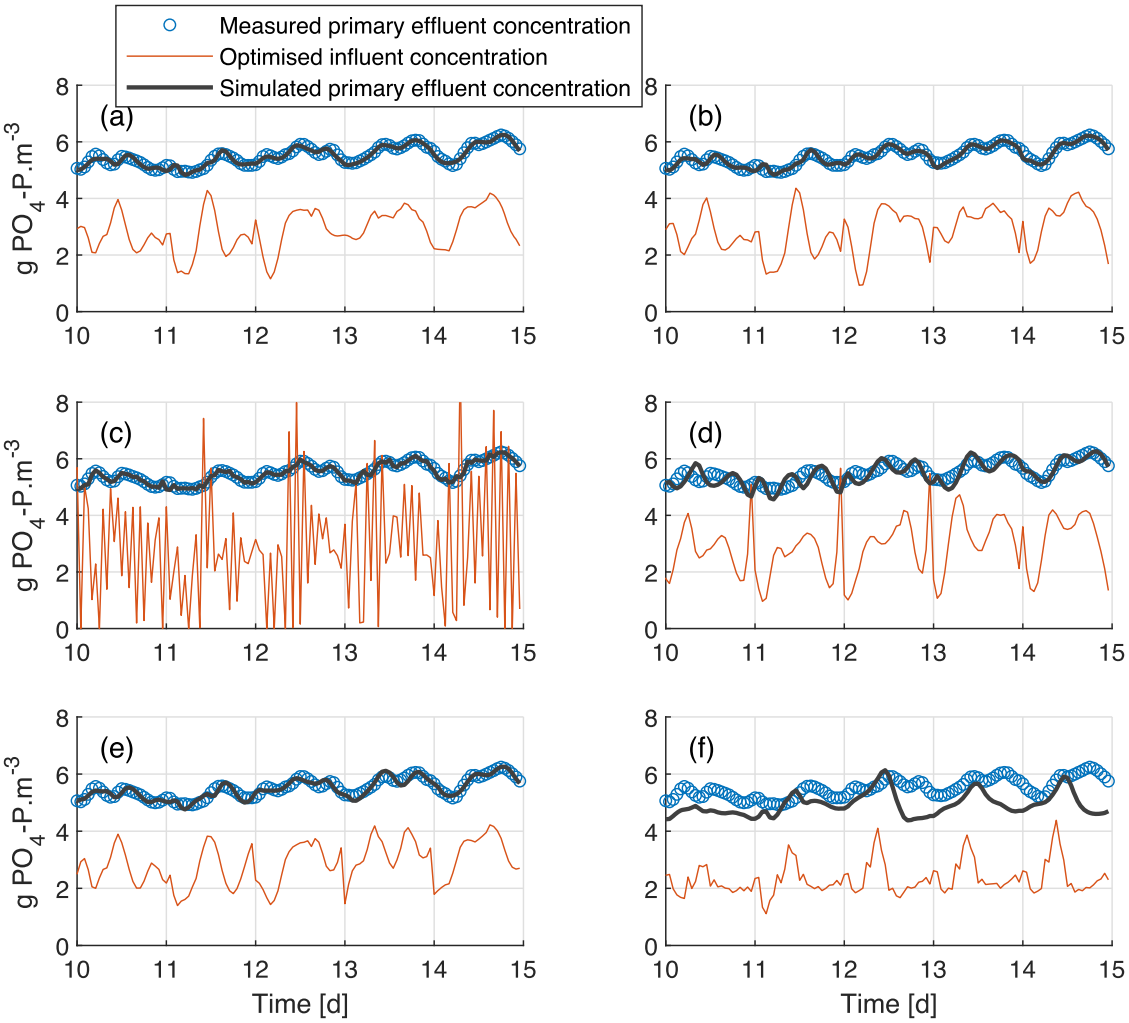


Fig. 4. Zoomed in version of Fig. 3 (days 10–15). Optimised influent orthophosphate concentrations with simulated and measured primary effluent concentrations for five different functions: (a) function 1: 3 normal distributions; (b) function 2: 4 normal distributions; (c) function 3: direct function; (d) function 4: 6th order polynomial; (e) function 5: Fourier series; and (f) baseline.

Table 1
Statistics of measured primary effluent orthophosphate concentration variations (g P/m^3), with mean, standard deviation and percentiles.

Mean	Std.dev.	10th perc.	25th perc.	75th perc.	90th perc.
5.36	0.69	4.69	5.14	5.80	5.99

all, as during day 26 there is a storm event with flow peaks considerably higher (178 000 m^3/d) than the daily average flows (around 50 000 m^3/d). During this event, all optimisation functions result in load values of zero during parts of or the whole day.

The high load peaks can possibly be explained by two mechanisms related to a first flush effect for dissolved compounds: 1) due to the fact that water travels as a wave in the sewer system (Krebs et al., 1999). This means that as rainwater enters the sewer system downstream, a wave is created in the sewer, increasing the flow rate downstream as the wave travels. The water that is diluted by the rainwater does not travel with the wave, but is transported with the regular water flow. Thus, as the wastewater with increased flowrate that reaches the WRRF first will be undiluted, creating a load peak. 2) recent discussions during a workshop at the WRRmod 2024 conference (Copp, 2024) and subsequent webinar presentation (Copp, 2025) displayed how deposition of particulates in the sewer system, followed by hydrolysis, caused dissolved compounds such as ammonium and orthophosphate to accumulate in the sediments and be flushed out during rain events. This could explain the variable influent load also during dry days with equal flow rates, and why the behaviour is different when there are several rain events in a row (as for day 24–26 in this work). This is important to follow up in future research.

After discussion with the plant sensor technician, a likely explanation for the predicted zero concentration during the storm on day 26 is due to partial clogging of the pre-filter in the analyser due to high TSS load. This leads to a too low sample volume being extracted, with erroneous data as a result as the calculated concentration is too low (which would lead to an estimated influent concentration that is too low). The underestimated measured primary effluent concentration will in turn lead to a too low influent concentration, as the measurement error propagates to the influent concentration in the optimisation. Although this remains to be confirmed, it shows the importance of regular sensor maintenance when using the data for control or for input in automatic calculations.

3.3. *Optimisation speed*

In general, the functions with fewer parameters are faster than the ones with more. However, the type of equation that is used also appear to affect the optimisation time. The optimisation time was similar between functions 2 and 5, while function 1 was 17 % faster and function 4

Table 2
Evaluation of results for comparison between the different mathematical functions used to describe the daily influent load variations. RMSE is the median daily value (calculated from hourly values each day) over the 40 days of optimisation, with σ the standard deviation. Relative optimisation time is the average time normalised to the time for the function with the lowest median RMSE (46.7 min, for function 2).

Function #	Qualitative assessment of produced influent time series	RMSE (median ± σ)	Relative optimisation time	R^2
1	Mostly realistic. Occasionally constant over a day and problems with spikes in transition between days. Extended time with zero values during heavy rainfall.	0.12 ± 0.15	0.83	0.90
2	Mostly realistic. Extended time with zero values during heavy rainfall.	0.07 ± 0.07	1.0	0.97
3	Extremely noisy. Some zero values during heavy rainfall.	0.08 ± 0.08	4.4	0.96
4	Realistic but substantial problem with spikes in concentration during transition between days. Extended time with zero values during heavy rainfall.	0.24 ± 0.11	0.62	0.83
5	Mostly realistic. Extended time with zero values during heavy rainfall.	0.13 ± 0.10	0.99	0.91
Baseline	–	0.82 ± 0.43	–	−1.24

was 38 % faster. Function 3 took more than four times as long to run than function 2, with similar mean RMSE and with the previously mentioned highly noisy data. The longer optimisation time together with the noisy output, without increase in prediction capabilities, show that parametric optimisation functions are preferred over the direct function for this type of data and system. For primary clarifiers without this high internal recycle flow (about 20 % of the total dry weather influent), the function might be less prone to this type of problem, although the dilution effect can also affect the amount of time that an individual load data point affects the output (depending on the hydraulic characteristics of the clarifier). If the time resolution is increased from one hour to several minutes, some increase in optimisation time is expected for all methods since producing higher frequency simulation output usually requires the solver to take smaller time steps (i.e., each simulation in the optimisation will take longer). However, the parametric methods have the advantage that the number of parameters does not increase with higher time resolution. The direct method (Function 3), however, will increase its number of parameters according to the new time resolution relative to the total number of minutes in a day. Consequently, for a time resolution of 10 min the number of parameters will be $24 \times 60/10 = 144$, likely resulting in a substantially longer optimisation time.

For practical purposes, the optimisation time does not appear to be limiting. For function 2, the mean optimisation time was 46.7 min with the default Nelder-Mead parameters on the PC used. This can likely be shortened by adjusting the solver tolerances of the Nelder-Mean algorithm, with minor loss in accuracy. It is therefore deemed suitable to run in a digital twin system once per day. For function 3, the mean optimisation time was 206.9 min (nearly four hours), which makes it less practically useful (although this optimisation time could probably also be shortened by adjusting the solver tolerances).

3.4. Comparison to the baseline method

All functions showed better RMSE compared to the baseline method. Some of the days the average output from the baseline method appears to be close to the average of the data, but for several days the daily average load appears to be too low in the baseline method as the primary effluent concentration is underestimated. This indicates that there are either daily variations in load substantial enough to create the observed difference, or that one or several of the internal loads are overestimated for parts of the dataset. For the optimisation functions, any such uncertainties and errors (influent and internal load input data and model structure) will be propagated to the optimised influent load. This highlights the risk of using such a function, as faulty data gives faulty influent data, but also a possibility to incorporate unknown phenomena in the influent file to still produce correct primary effluent data. This of course depends on the objective of the model, but one such scenario could be to use time series data generated with the function presented in this paper to develop an influent forecasting model. In any case, this highlights the need to properly validate the optimised influent data before use, to be aware of any discrepancies between model output and data and take decisions consciously.

3.5. Validation data

Fig. 5 shows a comparison between the influent orthophosphate load from the three measured days and optimised influent load from 26 simulated days with function 2 (weekend days and public holidays were excluded from the load analysis since no sampling was done on weekends and the load pattern is usually different). The mean absolute error between measured median hourly load values (based on three measurement days) and median simulated values was 35.8 kg P/d, which can be compared to the morning peak values near 200 kg P/d. The median optimised influent load is similar in shape to the observed variations and follows the diurnal trend, but three general observations can be done. Firstly, the morning peak is occurring around 11 in the simulated data and around 9–10 in the measured data. Secondly, the nighttime load appears to be very similar in the optimised influent and data. Thirdly, the load values after noon appears to be higher in the optimisation. When comparing the simulated load profile to the measured data, it should also be noted that we do not expect the daily load variation to be identical every day. Both the output from the baseline simulation, observed variations in the primary clarifier effluent and measured hourly influent NH_4–N load at the plant (with variations of 20–25 % around the median values, data not shown) supports this. An explanation to the difference between measured and optimised data could be due to a faulty description of the hydraulic behaviour in the primary clarifier. However, attempts to vary the structure of the 3-compartment model in Sumo (such as division of volume between the compartments or increasing the number of tanks in series in the clear water zone) did not have any meaningful impact on the results (data not shown) due to the short hydraulic residence time in the clarifier (on average around 4 h). Another possibility is that the high internal load causes small changes in the measured primary effluent to have large effects on the optimised influent, and the noise in the measured data could cause the deviation during the afternoon. This would not explain the good fit during nighttime and morning, though. Since this figure is comparing general trends between optimised days during 2021 and measurements from 2023, it is quite possible that the influent loads on the optimised days were slightly different than during the measurement days. In May/June 2021 the COVID-19 pandemic was still ongoing, likely affecting the behaviour related to the pollutant loads to the plant (i.e., people working at home and therefore waking up later and being home in the afternoon). This could explain why the nighttime load is comparing well to the measured load, if the nighttime routines were still similar at both points in time, while the morning peak occurs slightly later. Another possible explanation could be that the internal loads or

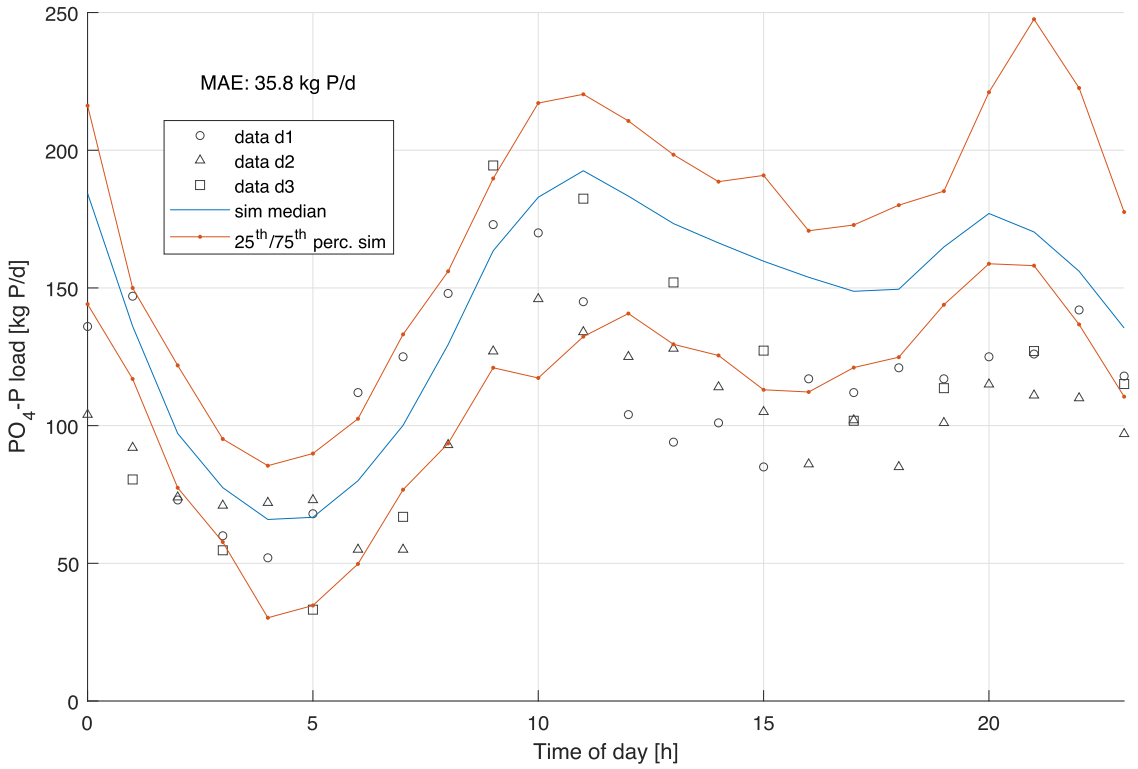


Fig. 5. Difference between validation measurements and predicted (optimised) influent orthophosphate load for the studied methods. The model output is calculated from 26 simulated days using function 2 (median, 25th precentile and 75th percentile values each hour), while the measured validation data are from three different days (data d1, data d2 and data d3). From the original 40 optimised days, weekend and holiday days were excluded as the measured data only includes workdays. The difference between simulated median hourly values and median hourly data is shown as mean absolute error, MAE.

processes contributing to primary effluent orthophosphate are underestimated, and that this error accumulates in the optimised influent load. The majority of the internal load is from the sludge dewatering reject, where the used flow rate is trusted (it is based on the digested sludge flow, which is validated by flow measurements at several locations showing near identical values). The orthophosphate concentration, however, is based on measurements of filtered total phosphorus and interpolation between the values. This is one great source of uncertainty identified in the data. Overfitting could also be a potential issue, which could probably be a partial explanation for the observed behaviour for function 3, but should be less of a problem with the parametric functions. In the end, the exact reason for the discrepancy between optimised output and measured data during the afternoon remains to be confirmed.

3.6. General discussion

In generalising the results, many of the potential issues presented (i.e., uncertainty in interval load and processes as well as time during individual load values impact effluent concentration) are probably due to the unusual process configuration at the specific plant. A primary clarifier with primary sludge hydrolysis occurring in the primary settler (i.e., an activated primary settler) is, to the authors' knowledge, not a common process choice. For the majority of cases, the uncertainties would therefore be lower compared to the presented case. For a case with prediction of influent ammonia nitrogen based on primary effluent measurements, without primary hydrolysis, it might be possible to set up a model without the need for using full plant data for the internal loads and only use measured sludge dewatering reject flow rate and ammonia concentration, greatly simplifying the effort and not requiring a plantwide model. This will have to be verified on a case-by-case basis, though. The presented optimisation functions are not expected to perform substantially different from the results presented here at other sites, as most functions have been used at different WRRFs.

Depending on the objectives and available data, the methodology of using downstream measurements and a model to predict upstream values could be extended to use more complex models and data further downstream. For example, using effluent measurements from an activated sludge process to estimate influent data. This concept is similar to what Johnson et al. (2024) did, where airflow rates in the activated sludge process were used to estimate influent concentration data. However, the more processes there are between the measurement point and estimation point, the higher the requirements will be on the quality of the data and the accuracy of the process model due to the accumulation of measurement and model errors in the influent prediction.

Compared to other soft sensor techniques, a benefit of the presented method is that it derives results from actual measurements of the variable of interest, although at another point in the process. It therefore already contains much of the information contained in the variable of interest. It is, however, potentially more complex than other methods and sensitive to errors in data and model. The work presented in this paper shows suitable mathematical functions when using this type of soft sensor and highlights areas of future research.

4. Conclusions

A method for constructing a soft sensor for estimating the influent orthophosphate load based on measured concentration in the primary clarifier effluent at a water resource recovery facility has been presented. The soft sensor considers effects of internal load and processes upstream of the sampling location through a combination of results from plantwide model simulation results and real plant data. The following conclusions could be drawn:

- All mathematical functions for describing the load variations could reproduce the primary clarifier effluent data with high accuracy and

improved primary effluent orthophosphate concentration predictions compared to the baseline scenario of a constant daily load profile. However, not all produced realistic influent time series.
- A combination of four normal distributions was deemed the most suitable function for describing the daily influent load variations, based on the low RMSE, low optimisation time and realistic influent concentration variations.
- Contrary to the baseline method, the optimisation methods predicted "first flush" effects of soluble compounds during rain events, which can have large effect on the process performance.
- This study supports the method with functions practical usefulness seen in other studies, and thus emphasise its transferability to other plants. For other plants without the specific process configuration with primary sludge hydrolysis, the implementation effort is expected to be lower.

The following future challenges remain:

- For most functions, the transitions between optimisation days still produces a "jump" in the values. Further research is required to find other ways of handling these issues.
- The validation of the model showed that the variations in optimised influent load were similar to the measured load variations, although slightly higher load values during the afternoon to evening requires further investigation. A long-term validation of the soft sensor is required for better understanding of its performance and is planned for a future publication.
- Some phenomena during wet weather that were seen in data could not be explained by the model and requires further investigation.

The presented methods are promising alternatives for utilities to both save on investment cost of additional sensors/analysers when sensors are already present downstream, although dynamic validation of the predicted influent concentration is recommended. It also allows placement of such equipment, e.g. after a primary clarifier, to avoid the potential problems of installing them at the plant inlet (such as clogging of pre-filters), although this problem might still exist to a lesser extent depending on the equipment used. Extending beyond the application presented in this work, the optimisation method presented can also be used for prediction of influent COD fractions or for control related tasks in a digital twin, such as optimisation of, for example, energy use, chemical use, greenhouse gas emissions or resource recovery.

CRediT authorship contribution statement

Christoffer Wärff: Writing – original draft, Visualization, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Bengt Carlsson:** Writing – review & editing, Methodology. **Magnus Arnell:** Writing – review & editing, Supervision. **Federico Micolucci:** Writing – review & editing, Investigation. **Oscar Samuelsson:** Writing – review & editing, Supervision, Formal analysis. **Ulf Jeppsson:** Writing – review & editing, Supervision.

Declaration of competing interest

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

Acknowledgements

The authors kindly acknowledge the funding provided by the Swedish Research Council Formas (2020–00222), Svenskt Vatten Utveckling and Nordvästra Skånes Vatten och Avlopp AB (NSVA). The authors also kindly acknowledge the input received on the subject during the WRRmod 2024 conference in Notre Dame, Indiana, USA, April 6–10, 2024. Especially the comments from Mariane Schneider, Bruce Johnson and Nina Gubser helped improve the presented work. Finally, the authors also thank the reviewers for their comments that helped improve the quality of the paper.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.watres.2025.124176.

Data availability

Data will be made available on request.

References

Achilleos, P., Roberts, K.R., Williams, I.D., 2022. Struvite precipitation within wastewater treatment: a problem or a circular economy opportunity? Heylion 8 (7), e09862.
Alex, J., 2024. Model-based construction of wastewater treatment plant influent data for simulation studies. Water 16 (4), 564.
Copp, J., 2024. Semi-mechanistic sewer model – tell me where you've been and I'll calculate a plant influent. In: *WRRF* Influent *Soft* Sensors – *Do* All Paths L*ead to Rome?* Workshop at the 11th IWA Water Resource Recovery Modelling Seminar (WRRmod2024). Notre Dame, USA.
Copp, J., 2025. All paths start in the sewer: variable influent characterisation through sewer modelling. In: WRRF Soft Sensors – It all Starts With Process knowledge, IWA Modelling and Integrated Assessment (MIA) Specialist Group Webinar. https://www.youtube.com/watch?v=8XlcnsLIg6Y.
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. Water Environ. Res. 96 (3).
Gernaey, K.V., Flores-Alsina, X., Rosen, C., Benedetti, L., Jeppsson, U., 2011. Dynamic influent pollutant disturbance scenario generation using a phenomenological modelling approach. Environ. Model. Softw. 26, 1255–1267.
Jeppsson, U., Pons, M.-N., Nopens, I., Alex, J., Copp, J.B., Gernaey, K.V., Rosen, C., Steyer, J.-P., Vanrolleghem, P.A., 2007. Benchmark simulation model no 2: general protocol and exploratory case studies. Water Sci. Technol. 56 (8), 67–78.
Johnson, B.R., Kadiyala, R., Owens, G., Ping, M.Y., Grace, P., Newbery, C., Sing, S., Sazena, A., & Green, J. (2021). Water reuse and recovery facility connected digital twin case study: singapore PUB's Changi WRP process, control and hydraulics digital twin. In: *Proceedings of WEFTEC 2021* 2021, Chicago, USA.
Johnson, B.R., Yang, C., Registre, J., Stewart, H., Rieger, L., Miletic, I., Pienta, D., Xiang, F., Rickermann, J., Menniti, A., Lesnik, K., Oristian, M., 2024. Development of Hybrid Digital Twins for Predictive Nutrient Control. Water Research Foundation, Alexandria, VA, USA. Project No. 5121.
Krebs, P., Holzer, P., Huisman, J.L., Rauch, W., 1999. First flush of dissolved compounds. Water Sci. Technol. 39 (9), 55–62.
Langergraber, G., Alex, J., Weissenbacher, N., Woerner, D., Ahnert, M., Frehmann, T., Halft, N., Hobus, I., Plattes, M., Spering, V., Winkler, S., 2008. Generation of diurnal variation for influent data for dynamic simulation. Water Sci. Technol. 57 (9), 1483–1486.
Langeveld, J., Van Daal, P., Schilperoort, R., Nopens, I., Flameling, T., Weijers, S., 2017. Empirical sewer water quality model for generating influent data for WWTP modelling. Water 9 (7), 491.
Li, F., Vanrolleghem, P.A., 2022a. Including snowmelt in influent generation for cold climate WRRFs: comparison of data-driven and phenomenological approaches. Environ. Sci. Water Res. Technol. 8, 2087–2098.
Li, F., Vanrolleghem, P.A., 2022b. An essential tool for WRRF modelling: a realistic and complete influent generator for flow rate and water quality based on data-driven methods. Water Sci. Technol. 85 (9), 2722–2736.
Mannina, G., Cosenza, A., Vanrolleghem, P.A., Viviani, G., 2011. A practical protocol for calibration of nutrient removal wastewater treatment models. J. Hydroinformatics 13 (4), 575–595.
Martin, C., Vanrolleghem, P.A., 2014. Analysing, completing, and generating influent data for WWTP modelling: a critical review. Environ. Model. Softw. 60, 188–201.
Nelder, J.A., Mead, R., 1965. A simplex method for function minimization. Comput. J. 7 (4), 308–313.
O'Kelly, B.C., 2005. Mechanical properties of dewatered sewage sludge. Waste Manag. 25 (1), 47–52.
Rieger, L., Takács, I., Villez, K., Siegrist, H., Lessard, P., Vanrolleghem, P.A., Comeau, Y., 2010. Data reconciliation for wastewater treatment plant simulation studies – planning for high quality data and typical sources of errors. Water Environ. Res. 82 (5), 426–433.
Roeleveld, P.J., van Loosdrech, M.C.M., 2002. Experience with wastewater characterization in The Netherlands. Water Sci. Technol. 45 (6), 77–87.
Savitzky, A., Golay, M.J.E., 1964. Smoothing and differentiation of data by simplified least squares procedures. Anal. Chem 36 (8), 1627–1639.

Sitzenfrei, R., Hillebrand, S., Rauch, W., 2017. Investigating the interactions of decentralized and centralized wastewater heat recovery systems. Water Sci. Technol. 75 (5), 1243–1250.

Torfs, E., Nicolai, N., Daneshgar, S., Copp, J.B., Haimi, H., Ikumi, D., Johnson, B., Plosz, 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 (10), 2840–2853.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C.J., Polat, I., Feng, Y., Moore, E.W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A. H., Pedrogosa, F., van Mulbregt, P., SciPy 1.0 Contributors, 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods 17 (3), 261–272.

Wentzel, M.C., Mbewe, A., Ekama, G.A., 1995. Batch test for measurement of readily biodegradable COD and active organism concentrations in municipial waste waters. Water SA 21 (2), 117–124.

Wärff, C., Arnell, M., Sehlén, R., Jeppsson, U., 2020. Modelling heat recovery potential form household wastewater. Water Sci. Technol. 81 (8), 1597–1605.

Zorrilla, F., Sadino-Riquelme, M.C., Hansen, F., Donoso-Bravo, A., 2024. Soft sensor for substrate characterization through the reverse application of the ADM1 model for anaerobic digestion plant operations. Water Sci. Technol. 90 (3), 721–730.