Journal of Water Process Engineering 61 (2024) 105231

Contents lists available at ScienceDirect

Journal of Water Process Engineering

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

Time-series analysis approach to the characteristics and correlations of wastewater variables measured in paper industry

Esko Toivonen *, Esa Räsänen

Computational Physics Laboratory, Tampere University, P.O. Box 692, Tampere FI-33014, Finland

ARTICLE INFO

Editor: Ludovic F. Dumée

Keywords:
Wastewater treatment
Time-series analysis
Chemical oxygen demand
Biochemical oxygen demand
Optimization

ABSTRACT

Advanced wastewater treatment technologies and protocols are required to minimize the overall water footprint and environmental cost of paper industries. Time-series analysis provides powerful methods to analyze industrial wastewater data with the aim of understanding and predicting the behavior of relevant wastewater variables such as chemical and biochemical oxygen demand. In this study, we introduce a variety of novel computational methods for complex systems and demonstrate their applicability to data obtained from a wastewater treatment plant, including the preprocessing of the raw data, time-dependent characteristics of individual measured parameters, as well as the mutual correlations between multiple parameters. These methods include the Potts model for preprocessing, empirical mode decomposition for periodic component extraction, detrended fluctuation analysis for noise characterization of the data and finally time-lagged windowed cross-correlation, which is utilized to obtain time-dependent couplings between the measured parameters. The results provide valuable insights into the underlying processes in the wastewater treatment plant. The Potts model effectively processes noisy data, from which periodic variations can be successfully removed with empirical mode decomposition, while preserving the relevant characteristics. Further, detrended fluctuation analysis shows prospects for indicating periods of abnormal behavior in the data. Finally, the correlation analysis reveals the characteristic time delays between influent and effluent chemical oxygen demand, giving an average delay of one day, which has implications in plant control. In conclusion, our methods show promising prospects for further applicability in wastewater analysis.

1. Introduction

Water plays a vital role in various industries and is utilized for a wide range of purposes. One of the most water-intensive industries over the years and still today is the pulp and paper industry. Fortunately, efficient water management is increasingly adopted to minimize water usage, reduce environmental impact, and optimize resource utilization. Paper mills strive to implement water conservation measures, improve process efficiency, and employ advanced treatment technologies to minimize the overall water footprint of their operations [1].

In pulp and paper industries, water is used in large amounts during, e.g., pulp production, debarking and papermaking. These processes produce wastewater, which needs to be cleaned before discharging to protect water bodies from pollution and eutrophication. As the regulations concerning wastewater discharges become more and more stringent, new techniques are needed to overcome the restrictions imposed to discharged water quality [1].

Various models and approaches can be employed to optimize wastewater treatment processes in the paper industry. *Process models* employ mathematical characterization of treatment operations using the wastewater properties and operating parameters as inputs [2]. These models can predict the effectiveness of treatment processes and thus aid the optimization and design. On the other hand, *mass balance models* can be applied to calculate mass flow rates and concentrations of pollutants at different stages of the treatment process [3]. By analyzing the mass balance, potential inefficiencies or areas of improvement can be identified, allowing for optimization of process parameters. Further, *biological models* simulate the biological reactions occurring in treatment systems such as activated sludge processes. The aim is improved removal of biological phosphorus (P), nitrogen (N) and organic carbon compounds [4].

Besides models, data-based approaches can be utilized to optimize wastewater treatment in paper industries. These approaches utilize historical process data, real-time monitoring data, and sensor

* Corresponding author.
E-mail addresses: esko.toivonen@tuni.fi (E. Toivonen), esa.rasanen@tuni.fi (E. Räsänen).

https://doi.org/10.1016/j.jwpe.2024.105231
Received 21 December 2023; Received in revised form 25 March 2024; Accepted 29 March 2024
Available online 8 April 2024

information to develop predictive models and control strategies. Modern wastewater treatment plants (WWTPs) produce large amounts of data, which can be analyzed to produce valuable information about the treatment process. By analyzing patterns, correlations, and system dynamics, machine learning algorithms can optimize operational parameters, control actions, and dosing strategies [5]. However, further data analysis methods are needed to (i) capture detailed understanding of the system dynamics and correlations with respect to the underlying WWTP processes, and to (ii) analyze *nonlinear* and *non-stationary* data which poses a significant methodological challenge to conventional time-series analysis. For example, it has been recently shown that incorporating time lags improves the performance of machine learning models [6,7] and soft sensors [8], making the analysis of time-dependent characteristics between wastewater variables valuable. Moreover, advanced analysis methods are essential in reducing energy consumption and achieving energy neutrality in WWTPs [9].

In this study, we analyze data generated from a paper mill WWTP with advanced time-series analysis methods stemming from computational physics. First, we introduce the Potts model for effective preprocessing of chemical oxygen demand (COD) and P data. This is followed by the introduction of detrended fluctuation analysis (DFA), dynamical DFA (DDFA) and empirical mode decomposition (EMD) to analyze the dynamical correlations and frequency contents in the real-time COD data. Further, we introduce time-lagged windowed cross correlation (TLWCC) as a powerful tool to capture the real-time cross correlations and time delays between relevant measured parameters such as the influent and effluent COD, biochemical oxygen demand (BOD) and dissolved oxygen (DO) concentration. The time delays can be further optimized by employing simulated annealing, which is introduced in detail. These methods – previously used in various different contexts in, e.g., nonlinear physics and physiology – are shown with our case example to be valid additions to the present ensemble of methods used in water treatment processes. Thus, the main novelty of the present work is the establishment of additional computational tools to utilize the continuously increasing amount of data available from industries producing wastewater, and hence to further optimize the treatment processes in terms of water footprint and quality monitoring.

The paper is organized as follows. In Section 2 we introduce the wastewater treatment plant used as our case example, the measured data and the preprocessing carried out with the Potts model. This is followed by a thorough review of DFA, DDFA, EMD, TLWCC, and simulated annealing. In Section 3 we present the results of the application of the methodology to our case example. The paper is summarized in Section 5.

2. Materials and methods

2.1. Wastewater treatment plant

We consider a Finnish paper mill wastewater treatment plant, which utilizes the well-established activated sludge process [10–12]. As a pretreatment step, the water is screened to remove the largest chunks of solid material. Primary clarification, which removes most of the solids in the wastewater, is performed in a single circular clarifier. After primary clarification, phosphoric acid and urea are added to the wastewater to keep the nutrient levels high enough for the activated sludge to clean the wastewater effectively [13]. Before entering the aeration pools, the wastewater is cooled with mechanical aerators inside the equalization basin, which is used to balance the COD load entering the secondary treatment.

The secondary treatment consists of two aeration basins operating in parallel. The COD content of the wastewater is measured before it enters the aeration basins. After aeration, the treatment plant has two secondary clarifiers, each corresponding to one aeration basin. The secondary clarifiers remove activated sludge from the cleaned water and return it to the aeration basins to keep the biomass concentration high enough for the secondary treatment to be effective [14]. After the secondary clarification, the wastewater passes through a flotation tank, which is used in case the solids do not settle properly in the secondary clarifiers. Finally, the water is discharged into a river.

2.2. Data

We examine wastewater variables obtained from the treatment plant, including both automatic and laboratory measurements. The automatic measurements have a resolution of one minute, whereas the laboratory measurements are taken at most once a day. In this study, we focus on the analysis of essential parameters, which were chosen to be the influent and effluent COD, the effluent seven-day BOD (BOD_7), the temperature measurements, and the DO concentration in the aeration basins. Both basins measure the DO concentration at both the beginning and the end of the aeration basin, thus the total number of DO measurements is four. For analysis, we received two datasets, and mainly concentrated on the first one due to its longer timespan. The first dataset covered the year 2021, containing 525,601 one-minute samples, whereas the second dataset spanned from May to September 2022, containing 198,721 one-minute samples.

2.3. Data preprocessing

The first step in the preprocessing of the WWTP data is the extraction of spurious step-wise segments resulting from the incompatibility of the sampling frequency and the data update frequency – see Sec. 3 and Fig. 1 for details. These piecewise-constant functions u are extracted from a noisy signal f by minimizing the Potts functional with an efficient algorithm – a problem present in various signal processing applications [15]. The Potts functional is named after the physical Potts model [16] developed as a generalization of the Ising model. The Potts functional, as well as many other segmentation algorithms, consists of a regularity term denoted as $K(u)$ and a fitness term denoted as $D(u,f)$. It is formally defined as

$$P_\gamma(u) = K(u) + D(u,f) = \gamma J(u) + \sum_i |u_i - f_i|, \tag{1}$$

where the regularity term $J(u)$ counts the number of jumps in u and the summation term measures the fitness of u to the noisy signal f. The coefficient γ can be adjusted to control the balance between data fidelity and the sparsity of the jumps. The regularity term is formally equal to the Hamiltonian in the physical Potts model [17].

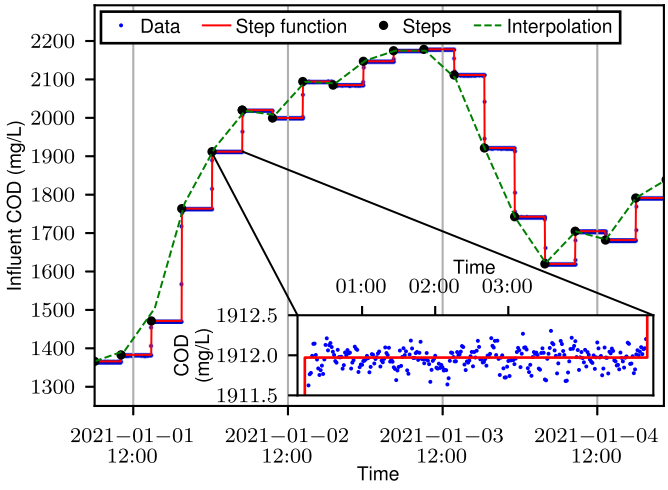

Fig. 1. Piecewise-constant function extraction with the Potts functional (red) and interpolation (green) based on the starting points of the steps. The minimization of the Potts functional successfully identifies and extracts the piecewise-constant function present in the noisy data.

The second step in the preprocessing is the removal of outliers in the data. This is carried out with a median-based scheme: for each point, the running median is calculated in a window of size w centered at the point, and if the absolute difference between the point and the median is higher than a specific threshold value Δ, the point is removed. The window sizes and threshold values are chosen by visual inspection, and the removed values are replaced with linear interpolation.

2.4. Detrended fluctuation analysis

Detrended fluctuation analysis [18] (DFA) and its recent extensions [19] provide effective tools to characterize short- and long-range fluctuations and noise in nonstationary data. DFA was originally developed for analyzing long-time correlation in DNA strands [18], and it has many applications, ranging from heart rate variability analysis [20] to financial time series [21].

The DFA algorithm proceeds as follows. The first step is to integrate the time series $x(t)$ to obtain the so-called signal profile $y(t)$ as a cumulative sum

$$y(t) = \sum_{i=1}^{t} [x(i) - \langle x \rangle], \tag{2}$$

where $\langle x \rangle$ denotes the mean of the original time series. The summing ensures that no assumptions about the stationarity of the original time series are made [22].

Next, the signal profile is split into non-overlapping windows of length s. The root-mean-square fluctuations of the data against least-squares polynomial fits in these windows are computed (detrending), yielding the result for the fluctuation function $F(s)$ for this scale [18]. The procedure is repeated for different scales s. The final step is to obtain the scaling exponent α by fitting a line to a log-log plot of $F(s)$ versus s. In other words, α gives the best fit to the dependency $F(s) \sim s^{\alpha}$. The value of α indicates the characteristics of the noise present in the data [18] as shown in Table 1.

The conventional DFA method produces a single scaling exponent for the entire time series. The behavior of a time series, however, might change as a function of scale or as a function of other parameters. This is why we also employ the dynamical DFA (DDFA) [19] to obtain scale- and time-dependent scaling exponents.

The first step in DDFA is to perform dynamical segmentation. The lengths of these segments are chosen as a function of scale by a function $l(s)$, which we choose to be a simple linear function $l(s) = as$, where a is a constant. The value of a can be adjusted to increase temporal resolution and to reduce noise. A typical value is $a = 5$, which we also use. After choosing the function, the time series is divided into overlapping segments of length $l(s)$ for each scale. These segments are identified with their temporal indices, which may indicate the mean time in the segment, but can also represent other parameters related to the segment.

Table 1
Interpretation of the scaling exponent α in detrended fluctuation analysis [19].

Scaling exponent α	Interpretation	Stationarity
$0 < \alpha < 1/2$	Anti-correlated	Stationary
$\alpha = 1/2$	White noise	
$1/2 < \alpha < 1$	Correlated	
$\alpha = 1$	$1/f$ (pink) noise	
$1 < \alpha < 1\,1/2$	Anti-correlated Increments	Non-stationary, stationary increments
$\alpha = 1\,1/2$	Brownian noise	
$1\,1/2 < \alpha < 2$	Correlated Increments	

The segments can then be denoted as $\mathscr{S}_{s,t}$, where t is the temporal index [19].

After dynamical segmentation, the DFA fluctuation function is calculated in each segment $\mathscr{S}_{s,t}$ at scales $\{s-1, s, s+1\}$ with overlapping windows. The logarithms of these fluctuation functions are denoted as $\widetilde{F}_t(s-1)$, $\widetilde{F}_t(s)$ and $\widetilde{F}_t(s+1)$, respectively. Based on these fluctuation functions, the dynamic scaling exponent $\alpha(t,s)$ can be calculated with the finite difference approximation as

$$\alpha(t,s) \approx \frac{h_-^2 \widetilde{F}_t(s+1) + \left(h_+^2 - h_-^2\right)\widetilde{F}_t(s) - h_+^2 \widetilde{F}_t(s-1)}{h_- h_+ (h_+ + h_-)}, \tag{3}$$

where $h_- = ln(s) - ln(s-1)$ and $h_+ = ln(s+1) - ln(s)$. The approximation is justified as the fluctuation functions calculated in overlapping segments have been empirically found to be sufficiently smooth [19].

2.5. Empirical mode decomposition

Empirical mode decomposition (EMD), first proposed by Huang et al. in 1998, is a powerful method for processing the frequency content of time series and signals. Compared to the conventional Fourier transform, EMD does not require the data be linear or stationary [23]. The EMD method is based on decomposing the signal into a relatively small number of periodic components known as Intrinsic Mode Functions (IMF) meeting the following criteria:

1. The number of extrema (local minima and maxima) and the number of zero crossings must either be equal or differ by at most one for the whole IMF.
2. At any point in the IMF, the local mean of the envelopes defined by the local minima and maxima must be zero.

In EMD, the IMFs are extracted from a time series $x(t)$ through a process called *sifting*. It proceeds as follows [23]:

1. The local maxima and minima of $x(t)$ are identified. The algorithm for extracting these points can be chosen freely.
2. The maxima and minima are separately connected by a cubic spline, and the mean of these splines m_1 is calculated.
3. The first component h_1 is defined as

$$h_1(t) = x(t) - m_1(t). \tag{4}$$

4. If this component does not fulfill the requirements for an IMF, the steps 1–3 are repeated by using the component in place of the original time series, producing the second component

$$h_{11} = h_1 - m_{11}, \tag{5}$$

where m_{11} is the mean of the cubic splines connecting the maxima and minima of h_1.

5. The preceding steps are repeated k times until a component h_{1k} fulfilling the IMF criteria is found. This component is defined as the first intrinsic mode function

$$c_1(t) = h_{1k}(t). \tag{6}$$

6. The first residual r_1 is calculated as

$$r_1(t) = x(t) - c_1(t). \tag{7}$$

7. The residual is subjected to the same sifting process as the original time series.
8. When an IMF c_n is small enough or it becomes a monotonic function, the process is stopped. The original time series can be expressed as a sum of its IMFs as

$$x(t) = \sum_{i=1}^{n} c_i + r_n, \tag{8}$$

where r_n is the last residual. Thus, the decomposition is complete, i.e., the original time series can be exactly reconstructed with the IMFs and the last residual.

EMD has been developed further to avoid particular numerical problems such as mode mixing [24]. This phenomenon occurs when either a single IMF contains signals of vastly different time scales, or when a signal of a certain time scale is split between several IMFs. To mitigate this problem, ensemble EMD (EEMD) was developed [25]. In EEMD, several instances of finite white noise are summed to the original time series. These sums are then decomposed independently and finally, the resulting IMFs are averaged over all instances of noise, which, given enough noise instances, will cancel the effect of the noise. As the average is taken only after all IMFs have been extracted, the resulting decomposition is not exact, which is why a complete EEMD with adaptive noise (CEEMDAN) was developed [26]. In CEEMDAN, the averaging is done after each IMF, and the residuals are obtained by subtracting this averaged IMF from the original time series, which makes the decomposition complete.

The resulting IMFs and especially their frequency content can be analyzed by utilizing the Hilbert transform [23]. As the instantaneous frequency and amplitude of an IMF may change over time, the Fourier transform is not suitable for this analysis, as it assumes that the time series is stationary and the frequencies and that amplitudes are constant over time. The Hilbert transform $\widetilde{x}(t)$ of a time series $x(t)$ is defined as

$$\widetilde{x}(t) = \mathscr{H}[x(t)] = \frac{1}{\pi}\mathrm{PV}\int_{-\infty}^{\infty}\frac{x(t')}{t-t'}\mathrm{d}t, \tag{9}$$

where the integral is defined in the sense of the Cauchy principal value (PV). The Hilbert transformed time series is the complex conjugate of the original time series, and the analytical signal $z(t) = x(t) + \widetilde{x}(t)\mathrm{i}$ can be formed. In a polar form, the analytical signal is expressed as $z(t) = a(t)exp[\mathrm{i}\theta(t)]$, where $a(t)$ is the instantaneous amplitude and $\theta(t)$ the instantaneous phase angle. By utilizing this phase angle, the instantaneous frequency $\omega(t)$ of a time series can be defined as

$$\omega(t) = \frac{\mathrm{d}\theta(t)}{\mathrm{d}t}. \tag{10}$$

In this study, we utilize the CEEMDAN algorithm to obtain the IMFs of several wastewater-related variables. The CEEMDAN calculations are performed with the libeemd library [27]. The Hilbert transforms and instantaneous frequencies are obtained with the emd Python library [28].

2.6. Time-lagged windowed cross-correlation

The Pearson correlation coefficient is one of the most used parameters to study the degree of a linear relationship between two variables [29]. However, calculating a single Pearson correlation coefficient between two variables assumes the variables to be stationary. In order to assess the correlations between two possibly non-stationary variables, we utilize the time-lagged windowed cross-correlation (TLWCC) algorithm [30]. In this algorithm, the variables are divided into short, overlapping windows of size w and the correlation coefficients are calculated between these windows. Now, only the windows are assumed to be stationary, not the whole time series. Our method differs slightly from the original algorithm [30]. In particular, the definition of the lag is opposite, and all the possible correlation coefficients are calculated.

The TLWCC algorithm proceeds as follows. First, two time series, both with N observations, are denoted with $x = \{x_1, x_2, \ldots, x_N\}$ and $y = \{y_1, y_2, \ldots, y_N\}$, respectively. The intervals between subsequent samples are assumed to be equal and constant between the time series. Next, a window size w and a lag τ are chosen so that $-\tau_{\max} \le \tau \le \tau_{\max}$, where $\tau_{\max}$ is the chosen maximum time lag. Additionally, a time index i is chosen satisfying $i \in \{1, 2, \ldots, N-w\}$. Note that since we let the time index i to be smaller than $\tau_{\max}$, the available range of lags is smaller while $i \in \{1, \ldots, \tau_{\max}\}$. Note that since we let the time index i to be smaller than $\tau_{\max}$, the available range of lags is smaller while $i \in \{1, \ldots, \tau_{\max}\}$. For this combination of τ and i, a pair of windows W_x, W_y can be constructed from the original time series as

$$W_x = \begin{cases} \{x_{i+\tau}, x_{i+\tau+1}, \ldots, x_{i+\tau+w}\} & \text{if } \tau \le 0 \\ \{x_i, x_{i+1}, \ldots, x_{i+w}\} & \text{if } \tau > 0 \end{cases} \tag{11}$$

$$W_y = \begin{cases} \{y_i, y_{i+1}, \ldots, y_{i+w}\} & \text{if } \tau \le 0 \\ \{y_{i-\tau}, y_{i-\tau+1}, \ldots, y_{i-\tau+w}\} & \text{if } \tau > 0 \end{cases}. \tag{12}$$

After forming the windows, the Pearson correlation coefficient between them is defined as

$$r(W_x, W_y) = \frac{1}{w}\sum_{j=1}^{w}\frac{(W_{x,j} - \langle W_x\rangle)(W_{y,j} - \langle W_y\rangle)}{\sigma(W_x)\sigma(W_y)}, \tag{13}$$

where $\langle W_x\rangle, \langle W_y\rangle$ are the means, $\sigma(W_x), \sigma(W_y)$ the standard deviations and $W_{x,j}, W_{y,j}$ the jth elements of windows W_x, W_y, respectively.

After computing the correlation coefficients for all the possible combinations of lag τ and time index i, the results can be collected into a matrix form. We have chosen to utilize a matrix where each row represents one lag, and each column one time index i. To further assess the overall behavior of the two correlated time series, the resulting matrix can be aggregated over time indices to obtain per-lag correlation coefficients. However, the correlation coefficients may not be simply averaged due to their skewed distribution. To alleviate this, we use the Fisher z-transform and average the z-transformed quantities instead of the original correlation coefficients [31].

2.7. Optimization with simulated annealing

Generally, an optimization problem considers some target function f and attempts to find the parameter combination with which the function reaches either its maximum or minimum value. The parameters usually have some boundary conditions and, in general, these types of problems can be divided into two categories, with either discrete or continuous parameters [32]. Numerous methods have been developed for solving these optimization problems [33]. Many of these algorithms, however, require certain properties from the target function f, they are only suitable for one problem type, or the extrema they find might only be local. To solve these problems, metaheuristic algorithms such as simulated annealing [34] have been proposed. In this study, we utilize simulated annealing for the optimization of discrete time lags between relevant WWTP parameters, e.g., between COD removal fraction and DO concentration.

Simulated annealing is a metaheuristical global optimization algorithm based on statistical physics. The name is based on metallurgy, where annealing is used to control the crystallization of materials by adjusting their temperature. Simulated annealing is suitable for both continuous and discrete problems, and it avoids local extrema by also allowing some parameter changes where the value of the target function f increases [34].

Simulated annealing employs the widely applied Metropolis algorithm [35], where the system under study (i.e. the target function f) is perturbed slightly (the parameters are varied) and the resulting energy of the system (value of the target function f) is calculated. If the perturbation lowered the energy the change is accepted unconditionally,

and if it raised the energy the change is accepted with a probability

$$P(\Delta E) = exp\left(-\frac{\Delta E}{T}\right), \tag{14}$$

where ΔE is the change in energy (change in function value) and T is a temperature. By varying T, the fraction of accepted changes can be adjusted. At a low temperature, the acceptance probability is low, but if the temperature is high, nearly all changes are accepted.

In this study, we utilize a temperature-adjustment scheme based on two nested loops [34]. In the outer loop, the temperature is adjusted and in the inner loop, the system state is changed. For controlling the temperature, we use the rule

$$T_{k+1} = \frac{T_k}{1 + T_k \frac{ln(1+\varepsilon)}{3\sigma_k}}, \tag{15}$$

where T_k is the temperature on the kth iteration of the outer loop, ε is 0.1 and σ_k is the standard deviation of the values of the target function f that were observed during the kth iteration of the outer loop. This scheme was chosen because it produced relatively fast convergence. We determine the initial temperature with the relation

$$\tau_0 = exp\left(-\frac{\langle \Delta E \rangle}{T_0}\right), \tag{16}$$

where τ_0 is the initial acceptance probability, T_0 the initial temperature, and $\langle \Delta E \rangle$ the average change in energy from initiating 100 disturbations to the initial position [32]. In our case, the initial acceptance probability is set to $\tau_0 = 0.5$.

3. Results

In this section we demonstrate the applicability of the analysis methods to data obtained from a Finnish WWTP that utilizes the well-established activated sludge process [10,11,36].

3.1. Preprocessing

The original time series describing the COD and P concentrations in the wastewater resemble a step function with added noise. This is due to the incompatibility of the sampler and data recording frequencies: the automatic sampler takes a sample every 4 h and 40 min, whereas the concentrations are recorded every minute into the database. Therefore, the time series forms a step function shown in Fig. 1, and the steps itself also contain noise. This noise is most likely caused by the sampler used and its AD-converter. As seen in Fig. 1, our Potts functional -based preprocessing method successfully identifies the step function present in the data.

To make the time series obtained with the Potts model uniformly spaced in time, linear interpolation was utilized. The starting points of the steps were taken as the basis of the linear interpolation, and the sampling period was chosen to be 4 h and 40 min, i.e., the interval between water samples taken by the automatic sampler. The starting points were chosen because they are the closest points to the moment in time when the sample was actually taken. The result of this linear interpolation is presented in Fig. 1 as a dashed green line. This preprocessing step lowered the number of samples significantly, to 1884 samples for the one-year dataset.

3.2. Dynamical detrended fluctuation analysis

Next we demonstrate the capability of DDFA to extract the characteristics of the measured data in different time scales. In particular, we focus on the nature of the fluctuations (e.g., white noise or long-range correlations) observed in the COD data as a function of both time (date of measurement) and scale (from short to long term). Fig. 2 (a) shows the DDFA results for the influent COD measured from February to December in 2021, together with a zoomed section in Fig. 2 (b). The (preprocessed) COD data is shown as a gray line, whereas the DDFA scaling exponent α is portrayed in colorscale as a function of time (x-axis) and scale (y-axis). The α landscape is dominated by yellow-red regions, indicating that the COD data is characterized by Brownian noise. This behavior is the most prominent in the smallest scales ($\sim$ 1 day) and in the range of 4–16 days. The results reflect the smooth development of COD during the day (see Fig. 1 for an example) as well as the longer trends clearly visible in Fig. 2 (b). In this zoomed Figure, the yellow-red regions at smaller scales are clearly visible, with a transition into lower α values between scales of 300 and 400 h. There are occasional abrupt fluctuations in COD in a scale of a few days, leading to decreased α values, e.g., in early April 2021. In addition, in the largest scales (>2 weeks), the COD behavior resembles pink noise with $\alpha \sim 1$. The dataset we analyzed, however, only spans one year, so we cannot assess, e.g., seasonal patterns that arise on a yearly basis. With further data, these seasonal patterns could be uncovered and DDFA could be possibly used to, e.g., identify periods of abnormal operation.

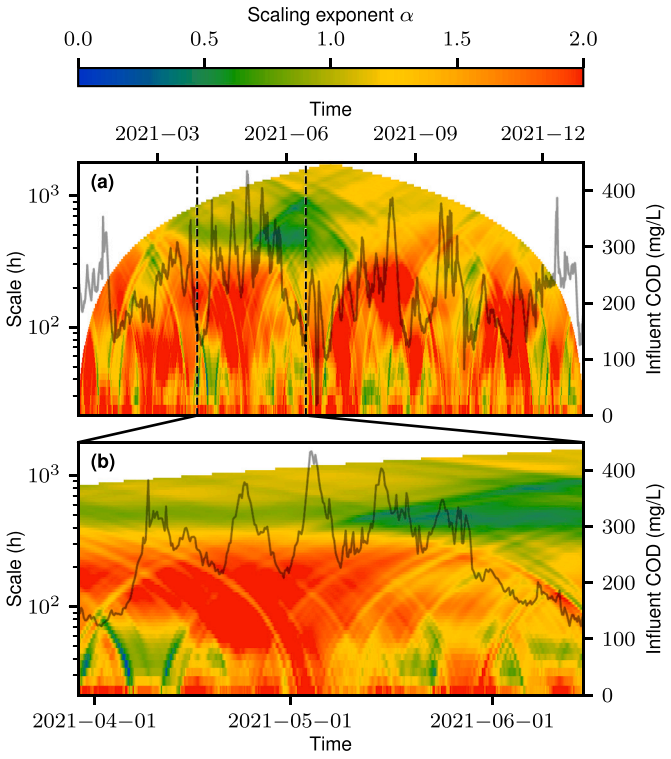

Fig. 2. Dynamical detrended fluctuation analysis (DDFA) of influent COD. (a) Influent COD (gray curve) and the scaling exponent of DDFA (colorscale). Note that the scaling exponent is a function of both time (x-axis) and scale (y-axis). (b) Zoomed section of the results in (a).

3.3. Empirical mode decomposition

We utilized the CEEMDAN algorithm to extract and remove the daily variations from temperature measurements taken from the influent water to aeration. The removal of the daily variation is important because these (trivial) variations may complicate the analysis of, e.g., cross correlations between the temperature and other variables. First, the sifting process was applied to obtain the IMFs. Then, the Hilbert transform was performed to obtain the instantaneous phases and frequencies present in these IMFs. As seen in Fig. 3 (a), the CEEMDAN algorithm successfully extracts the IMFs containing the daily variation, i.e., those with numbers 12 and 13, centered around a frequency of one per day. Removing these IMFs gives a time series free of daily variation

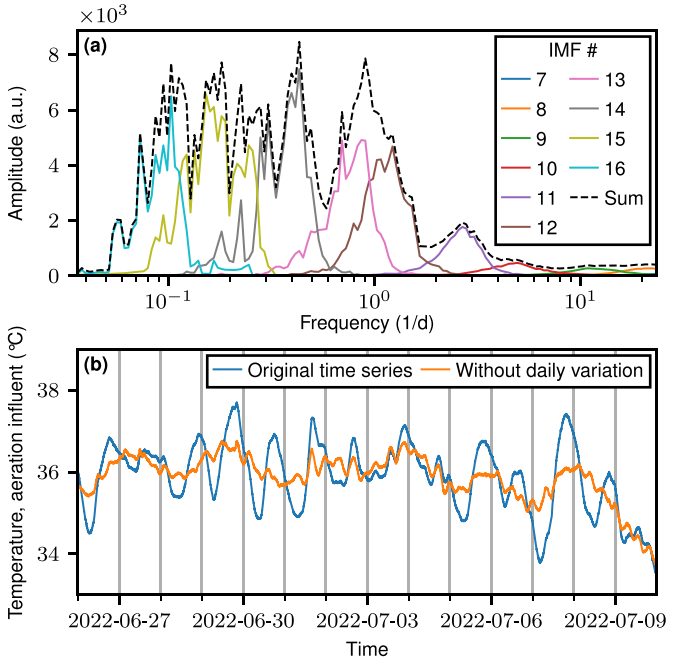

Fig. 3. Removal of daily variation from temperature data with CEEMDAN. (a) Intrinsic mode functions (IMFs) extracted with the CEEMDAN method from the temperature data in (b) measured in the influent water before aeration. The blue line in (b) shows the original temperature, whereas the orange line shows the temperature without daily variations, i.e., with IMFs 12 and 13 removed. The removal of these IMFs preserves trend and fluctuation characteristics of the original data.

as shown in Fig. 3 (b). Here the blue curve represents the original, unprocessed temperature data and the orange curve the data without daily variations. Our CEEMDAN-based algorithm successfully removes the daily variation while preserving (i) the general trend of the data and (ii) the fluctuations present in the data, which can then be analyzed with additional methods, if desired.

3.4. Cross-correlation

3.4.1. Influent and effluent COD

We applied TLWCC to analyze the time-dependent cross-correlation between influent and effluent COD. The result of this analysis is presented in Fig. 4 (a). We find a red area, indicating positive correlations, at negative lags. This suggests – as expected – that the influent COD precedes the effluent COD. This red area is present throughout the entire timespan. To better quantify the time delay between influent and effluent COD, we averaged the correlation matrix over time with Fisher's z-transform (see Section 2.6) and computed the time delay giving the maximal average correlation coefficient. Additionally, to assess the robustness of our method, we computed these averages for several different window lengths in the correlation matrix of the TLWCC method. These averaged results are presented in Fig. 5. We find that the lag of maximal correlation stays the same for several different values of

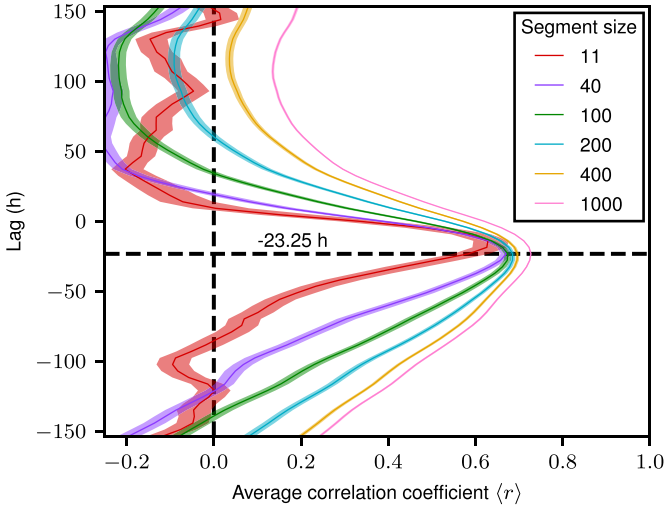

Fig. 5. Aggregated cross-correlation between influent and effluent COD for different segment sizes. The lag at which the correlation reaches its maximum converges fast as a function of the segment size. Only the smallest segment size of 11 reaches a different lag compared to the other segment sizes. The error estimate represents the 95 % confidence interval, obtained with bootstrapping.

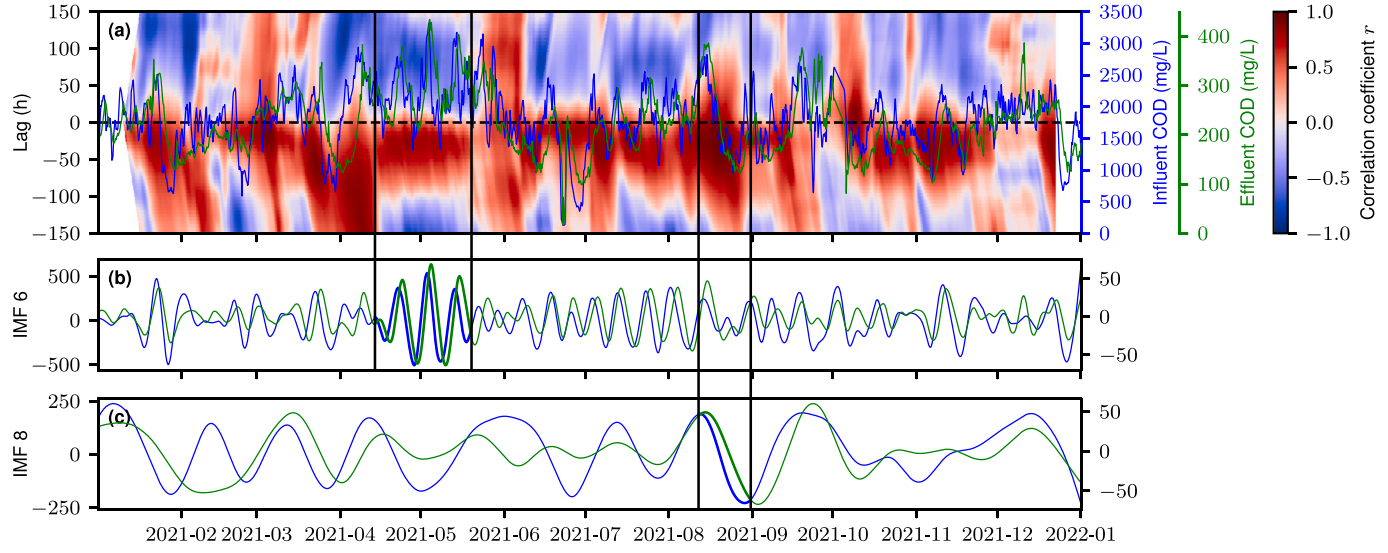

Fig. 4. Cross-correlation between influent and effluent COD. (a) Time-lagged windowed cross correlation (colorscale) between the influent and effluent COD shown in blue and green curves, respectively. (b-c) IMFs 6 and 8, respectively, computed with empirical mode decomposition. As emphasized, the IMF 6 captures oscillations whereas a large drop is present in IMF 8. See the text for details.

the segment size, which confirms the robustness of the method.

As seen in Fig. 5, the average correlation coefficient over time reaches its maximum at a lag of 23.25 h, which suggests that the influent COD precedes the effluent COD by this time interval, i.e., the time delay in the treatment process is approximately one day. This corresponds to the time it takes for the water to run through aeration and clarifier basins. The delay also affects the control of the wastewater treatment plant: there is approximately one day time to react to unexpected changes in the incoming water entering the aeration process.

The type of correlation between the influent and effluent COD changes over time, depending on different conditions and protocols at the WWTP. According to Fig. 4 (a), during some time intervals the correlation coefficient is relatively high for all the calculated lags, whereas sometimes the high correlation occurs at a limited span of lags. For example, between April and June 2021, the positive correlations are limited to a narrow span of lags, which implies that both influent and effluent COD values behave periodically with approximately similar frequencies. These periodic similarities can be verified with EMD. In Fig. 4 (b), we present the sixth IMFs of both influent and effluent COD. We find that in April–June 2021 the oscillation captured in this IMF is particularly strong. On the other hand, the large drop present in both effluent and influent COD in August 2021 is captured by the 8th IMF as shown in Fig. 4 (c). These IMFs also demonstrate the time lag present between the influent and effluent COD measurements, as there is clearly a phase difference between the IMFs for the influent and effluent COD.

3.4.2. DO concentration and COD removal

Typically, the most energy-consuming part of the activated sludge process is the pumping of air into aeration. This makes the analysis of the correlations between DO concentration in aeration and different WWTP performance parameters, such as COD removal fraction, relevant. The plant operators want to keep the pumps running at a low power level while also removing a high fraction of the COD from the wastewater. In this study, the WWTP has two DO concentration measurements in both aeration basins, at the beginning and end of each basin.

To obtain the COD removal fraction, we first apply the lag of 23.25 h to the influent COD value. Then, we calculate the removal fraction $\mathrm{COD}_{\mathrm{frac}}$ as

$$\mathrm{COD}_{\mathrm{frac}} = \frac{\mathrm{COD}_{\mathrm{in}} - \mathrm{COD}_{\mathrm{eff}}}{\mathrm{COD}_{\mathrm{in}}}, \tag{17}$$

where $\mathrm{COD}_{\mathrm{in}}$ represents the influent COD with the lag applied and $\mathrm{COD}_{\mathrm{eff}}$ the effluent COD.

To assess the collective effects of DO concentrations and still obtain individual time delays for each measurement, we sum the time-lagged DO measurements and then calculate the TLWCC matrix between this summed quantity and the effluent COD. As the time delays are discrete, and the resulting correlation coefficient does not necessarily form a smooth function, the most usual optimization methods are not suitable. Thus, we employ simulated annealing introduced in Section 2 to obtain the time delays corresponding to the maximal correlation between the sum of the DO concentrations and the COD removal fraction. Before performing the calculations, the DO measurements were preprocessed with our median-based scheme to remove outliers caused by the periodic surges of air coming from the pumping system.

In this case, we computed $N = 10000$ instances of the simulated annealing process. Several instances were needed due to the stochastic nature of the algorithm to ensure that the global optimum would be reached. Both the algorithmic temperature and the correlation coefficient averaged through Fisher's z-transform converge quickly, at around 300 annealing steps. The time delays for optimal lags for the aeration basins (A1 and A2) and DO measurements (O1 and O2) are shown in the table in Fig. 6. We find that all lags are quite high, suggesting that oxygen affects the COD removal fraction quite slowly.

The cross correlation map between the sum of DO concentrations and the COD removal fraction computed with the TLWCC method *without* the optimized lags is show in Fig. 7 (b). The averaged correlation coefficients for each lag are presented in Fig. 7 (a). For the most part of the measurement time, we find a strip of positive correlation at a lag of around 100 h, implying a significant delay between the DO concentration rising and it affecting the COD removal fraction. The result of positive correlation is plausible, since as the DO concentration increases, the micro-organisms can use more energy and break down pollutants and contaminants more effectively, leading to a reduction of effluent COD and thus also increasing the COD removal fraction. However, the time delay with which this positive correlation is achieved is relatively high, in the order of three days.

Measurement	Lag
A1O1	71:38
A1O2	79:00
A2O1	111:31
A2O2	61:01

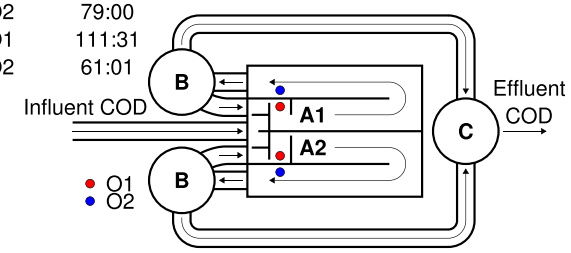

Fig. 6. Optimal lags between the oxygen measurements and COD removal fraction, obtained with simulated annealing. The table displays the lags, while the figure illustrates the WWTP structure. The notation is defined as follows: A1 = aeration basin 1, A2 = aeration basin 2, O1 = oxygen measurement 1, O2 = oxygen measurement 2, B = secondary clarifier, C = flotation basin.

Fig. 7 (d) shows the corresponding cross-correlation map *with* optimized lags listed in the table in Fig. 6. The largest difference compared to the case without optimization comes from the location of the positive correlations. Now, the positive correlations are concentrated around zero lag, and the averaged correlation coefficient shown in Fig. 7 (c) also peaks at zero lag. The value of this maximal correlation coefficient is, however, only slightly higher than in the case without optimized lags.

Here we analyzed the relationship between DO concentration in aeration and the COD removal fraction. The effluent COD, however, is a more readily available parameter to the operators, and therefore the relations between it and the DO concentration in aeration are also valuable. We have performed similar correlation analysis with these parameters and included it as a Supplementary material. The results suggest that there is a negative correlation between the effluent COD and DO concentration in aeration (see Supplementary Fig. 2), meaning that increasing the DO concentration lowers the effluent COD. The effect, however, diminishes as DO concentration rises, as demonstrated by Supplementary Fig. 3.

3.4.3. Effluent COD and BOD_7

BOD measures the amount of oxygen consumed when carbon-based organic substances in wastewater stabilize under aerobic conditions a result of biochemical processes. BOD correlates with the amount of organic matter in wastewater, so it can be used as an indirect indicator of pollution potential. The laboratory determination of BOD from a wastewater sample is performed as follows: first, the amount of DO in the sample is measured. Then, the sample is placed in a container defined by a standard, and a specified time, e.g., seven days, is allowed to pass before measuring the DO concentration again. The difference between the measured DO concentrations is used to calculate BOD_7, which represents the BOD over a seven-day period.

Fig. 8 shows the TLWCC result for the cross-correlation between effluent BOD_7 and effluent COD. The automatically measured COD values have been averaged over weekly periods to match with the BOD_7 measurement data. We find that the ratio of BOD_7/COD changes significantly over the measurement period, which is due to the frequent

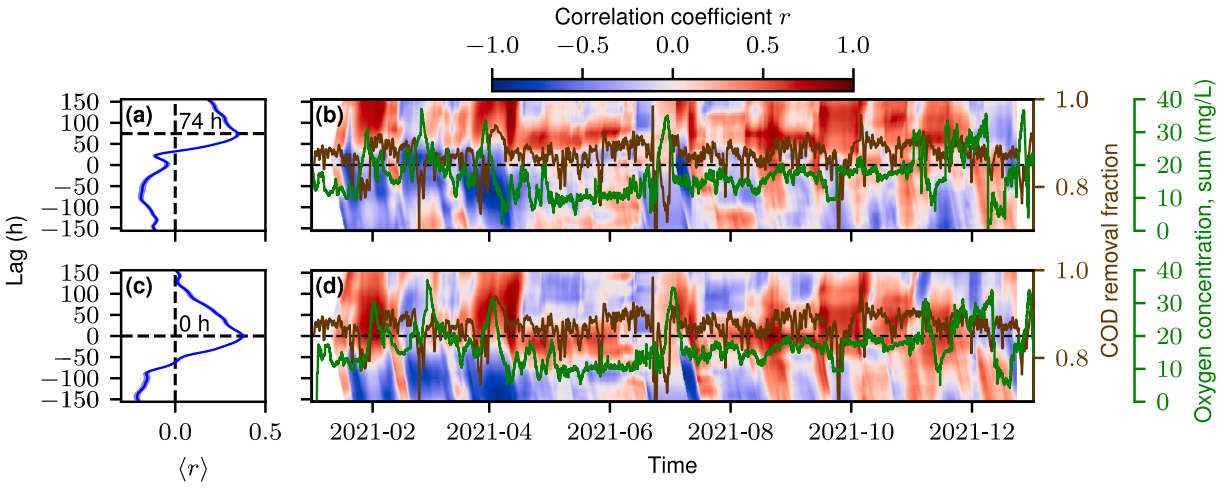

Fig. 7. Cross-correlation between COD removal fraction and effluent DO concentration. (a) Average cross correlation between the effluent COD (green curve) and oxygen concentration (purple curve) computed with TLWCC shown in (b) in colorscale. (c-d) As (a-b) but with the *optimized* lags.

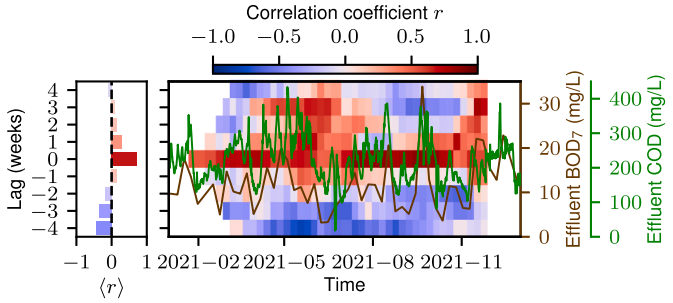

Fig. 8. Time-lagged windowed cross correlation (colorscale) between the effluent BOD_7 and COD shown in blue and green curves, respectively. The automatically measured COD values have been averaged over weekly periods to match with the BOD_7 measurement data. The right panel shows the average correlation coefficient for different lags (in weeks).

changes in the paper quality affecting the water composition. However, even though the available data is limited to 50 weekly BOD_7 measurements, we find a relatively high correlation with an average correlation coefficient of 0.72 with zero lag (simultaneous correlation). This is a promising result in view of the potential to *predict* BOD from automated COD measurements – and thus mitigate the requirements for tedious laboratory measurements of BOD in the long term. In particular, it would be important to assess if excess of critical BOD thresholds (e.g., 10 mg/L) could be predicted with effluent COD monitoring. We utilized a binary classifier to test the predictive value of COD measurements for a threshold of 10 mg/L. In 82% of the cases BOD_7 exceeding the threshold is correctly predicted from COD. This is a reasonable result that supports further data analysis in this direction with more data.

4. Discussion

We have introduced several advanced time-series analysis tools that are widely applicable to analyze multivariate wastewater data, including the preprocessing of the raw data, time-dependent characteristics of individual measured parameters, as well as the mutual correlations between multiple parameters. The presented methodology helps to understand the relevant features and changes in the data, as well as the underlying industrial processes. The tools also serve mechanical or statistical models, or those based on artificial neural networks [37].

In our case example of a WWTP we demonstrated with a few examples that the introduced methodology is useful to understand the main characteristics of the data, including the type of COD fluctuations in different scales captured by DDFA, and the frequency modes extracted with EMD. Particular attention was laid on the mutual correlations of the measured parameters. TLWCC proved to give valuable insights not only into the type, magnitude and development of cross-correlations, but also into the relevant time delays between, e.g., influent and effluent COD. In addition, simulated annealing combined with TLWCC provides the characteristic delays in the aeration process and may thus be a useful tool in the design of optimized oxidation protocols at the WWTP. Finally, TLWCC also gives valuable information on how the continuously monitored effluent COD affects effluent BOD_7, which is a critical parameter in the assessment of pollution potential.

The main limitation in the present study is the amount of data. We focused on a single WWTP with measurement data from (mainly) one year. For BOD_7 we are limited by only about 50 weekly measurements. However, even with the amount of data the results are promising and call for further studies with additional data from additional WWTPs. In addition, the methodology is also directly applicable to, e.g., municipal wastewaters and also to other situations producing time-dependent data of environmental parameters.

5. Conclusions

In conclusion, we have shown that advanced time-series analysis provides powerful methods to analyze industrial wastewater data. The application of the Potts model provides a valuable tool to remove spurious step-function-type artifacts in the raw data caused by incompatible frequencies in the measurements and data update. On the other hand, the recent extensions to the detrended fluctuation analysis (DFA) and empirical mode decomposition (EMD), i.e., *dynamical* DFA and *complete ensemble* EMD with adaptive noise (CEEMDAN), respectively, are useful to characterize the fluctuations and frequency components in the data, for example, in COD. Further, time-lagged windowed cross correlation (TLWCC) captures the dynamical cross correlations between influent and effluent COD, as well as between COD removal fraction and DO concentration in aeration. With simulated annealing we were able to optimize the lags for maximum correlation, which gives insight into the relevant delays in the aeration process. Finally, we showed that automatically and frequently measured COD values are relatively strongly correlated with (weekly) BOD_7 values obtained in laboratory measurements. Hence, with more data and validation it may be possible to reliable predict the probabilities for critical BOD values with continuous COD measurements. This would aid us to develop more effective and

sustainable monitoring protocols.

CRediT authorship contribution statement

Esko Toivonen: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Esa Räsänen:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

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

Data availability

The data that has been used is confidential.

Acknowledgements

This work was supported by the *Circular Economy of Water in Industrial Processes* (CEIWA) Co-Innovation Project of Business Finland, grant number 563/31/2021. The authors would like to thank all the company representatives for providing the data and for their helpful comments. The authors also acknowledge Marika Kokko, Jukka Rintala, Silja Mustonen, Matti Vilkko, Hussain Ahmed, Jani Tomperi and Henri Pörhö for fruitful collaboration in the project and for many useful discussions.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jwpe.2024.105231.

References

[1] A. Esmaeeli, M.-H. Sarrafzadeh, S. Zeighami, M. Kalantar, S.G. Bariki, A. Fallahi, H. Asgharnejad, S.-B. Ghaffari, A comprehensive review on pulp and paper industries wastewater treatment advances, Ind. Eng. Chem. Res. 62 (21) (2023) 8119–8145, https://doi.org/10.1021/acs.iecr.2c04393.
[2] Water Environment Federation (Ed.), Wastewater Treatment Process Modeling, MOP31, 2nd edition, McGraw-Hill Education, 2014.
[3] L.A. Rasmussen, L. Iordachescu, S. Tumlin, J. Vollertsen, A complete mass balance for plastics in a wastewater treatment plant - macroplastics contributes more than microplastics, Water Res. 201 (2021) 117307, https://doi.org/10.1016/j.watres.2021.117307.
[4] J. Nemcik, F. Krupa, S. Ozana, Z. Slanina, Wastewater treatment modeling methods review, IFAC-PapersOnLine 55 (4) (2022) 195–200, https://doi.org/10.1016/j.ifacol.2022.06.032.
[5] D. Wang, S. Thunéll, U. Lindberg, L. Jiang, J. Trygg, M. Tysklind, N. Souihi, A machine learning framework to improve effluent quality control in wastewater treatment plants, Sci. Total Environ. 784 (2021) 147138, https://doi.org/10.1016/j.scitotenv.2021.147138.
[6] Y. Alali, F. Harrou, Y. Sun, Unlocking the potential of wastewater treatment: machine learning based energy consumption prediction, Water 15 (13) (2023), https://doi.org/10.3390/w15132349.
[7] Y. Zhang, C. Li, H. Duan, K. Yan, J. Wang, W. Wang, Deep learning based data-driven model for detecting time-delay water quality indicators of wastewater treatment plant influent, Chem. Eng. J. 467 (2023) 143483, https://doi.org/10.1016/j.cej.2023.143483.
[8] W. Wang, C. Yang, J. Han, W. Li, Y. Li, A soft sensor modeling method with dynamic time-delay estimation and its application in wastewater treatment plant, Biochem. Eng. J. 172 (2021) 108048, https://doi.org/10.1016/j.bej.2021.108048.
[9] M. Maktabifard, E. Zaborowska, J. Makinia, Achieving energy neutrality in wastewater treatment plants through energy savings and enhancing renewable energy production, Rev. Environ. Sci. Biotechnol. 17 (4) (2018) 655–689, https://doi.org/10.1007/s11157-018-9478-x.
[10] E. Ardern, W.T. Lockett, Experiments on the oxidation of sewage without the aid of filters, J. Chem. Technol. Biotechnol. 33 (10) (1914) 523–539, https://doi.org/10.1002/jctb.5000331005.
[11] E. Ardern, The oxidation of sewage without the aid of filters. Part II, J. Chem. Technol. Biotechnol. 33 (23) (1914) 1122–1124, https://doi.org/10.1002/jctb.5000332304.
[12] D. Jenkins, J. Wanner, Activated Sludge - 100 Years and Counting, IWA Publishing, 2014.
[13] A.H. Slade, R.J. Ellis, M. van den Heuvel, T.R. Stuthridge, Nutrient minimisation in the pulp and paper industry: an overview, Water Sci. Technol. 50 (3) (2004) 111–122, https://doi.org/10.2166/wst.2004.0175.
[14] M. von Sperling, Activated Sludge and Aerobic Biofilm Reactors, IWA Publishing, 2007, https://doi.org/10.2166/9781780402123.
[15] G. Winkler, Image analysis, random fields and Markov chain Monte Carlo methods: a mathematical introduction, in: Stochastic Modelling and Applied Probability, 2nd edition vol. 27, Springer Science & Business Media, 2003, https://doi.org/10.1007/978-3-642-55760-6.
[16] R.B. Potts, Some generalized order-disorder transformations, Math. Proc. Camb. Philos. Soc. 48 (1) (1952) 106–109, https://doi.org/10.1017/S0305004100027419.
[17] G. Winkler, O. Wittich, V. Liebscher, A. Kempe, Don't shed tears over breaks, Jahresbericht der Deutschen Mathematiker-Vereinigung 107 (2) (2005) 57–87.
[18] C.-K. Peng, S.V. Buldyrev, S. Havlin, M. Simons, H.E. Stanley, A.L. Goldberger, Mosaic organization of DNA nucleotides, Phys. Rev. E 49 (1994) 1685–1689, https://doi.org/10.1103/PhysRevE.49.1685.
[19] M. Molkkari, G. Angelotti, T. Emig, E. Räsänen, Dynamical heart beat correlations during running, Sci. Rep. 10 (1) (2020) 13627, https://doi.org/10.1038/s41598-020-70358-7.
[20] C. Peng, S. Havlin, H.E. Stanley, A.L. Goldberger, Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series, Chaos 5 (1) (1995) 82–87, https://doi.org/10.1063/1.166141.
[21] J.R. Thompson, J.R. Wilson, Multifractal detrended fluctuation analysis: Practical applications to financial time series, Math. Comput. Simul. 126 (2016) 63–88, https://doi.org/10.1016/j.matcom.2016.03.003.
[22] R. Hardstone, S.-S. Poil, G. Schiavone, R. Jansen, V. Nikulin, H. Mansvelder, K. Linkenkaer-Hansen, Detrended fluctuation analysis: a scale-free view on neuronal oscillations, Front. Physiol. 3 (2012), https://doi.org/10.3389/fphys.2012.00450.
[23] N.E. Huang, Z. Shen, S.R. Long, M.C. Wu, H.H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, H.H. Liu, The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis, Proceedings: Mathematical, Physical and Engineering Sciences 454 (1971) (1998) 903–995, https://doi.org/10.1098/rspa.1998.0193.
[24] N.E. Huang, Z. Shen, S.R. Long, A new view of nonlinear water waves: the Hilbert spectrum, Annu. Rev. Fluid Mech. 31 (1) (1999) 417–457, https://doi.org/10.1146/annurev.fluid.31.1.417.
[25] W.U. Zhaohua, N.E. Huang, Ensemble empirical mode decomposition: a noise-assisted data analysis method, Adv. Adapt. Data Anal. 01 (01) (2009) 1–41, https://doi.org/10.1142/S1793536909000047.
[26] M.E. Torres, M.A. Colominas, G. Schlotthauer, P. Flandrin, A complete ensemble empirical mode decomposition with adaptive noise, in: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2011, pp. 4144–4147, https://doi.org/10.1109/ICASSP.2011.5947265.
[27] P.J.J. Luukko, J. Helske, E. Räsänen, Introducing libeemd: a program package for performing the ensemble empirical mode decomposition, Comput. Stat. 31 (2) (2016) 545–557, https://doi.org/10.1007/s00180-015-0603-9.
[28] A.J. Quinn, V. Lopes-dos Santos, D. Dupret, A.C. Nobre, M.W. Woolrich, EMD: empirical mode decomposition and Hilbert-Huang spectral analyses in Python, Journal of Open Source Software 6 (59) (2021) 2977, https://doi.org/10.21105/joss.02977.
[29] P. Schober, C. Boer, L.A. Schwarte, Correlation coefficients: appropriate use and interpretation, Anesth. Analg. 126 (5) (2018) 1763–1768, https://doi.org/10.1213/ANE.0000000000002864.
[30] S.M. Boker, J.L. Rotondo, M. Xu, K. King, Windowed cross-correlation and peak picking for the analysis of variability in the association between behavioral time series, Psychol. Methods 7 (3) (2002) 338–355, https://doi.org/10.1037/1082-989X.7.3.338.
[31] N.C. Silver, W.P. Dunlap, Averaging correlation coefficients: should Fisher's z transformation be used? J. Appl. Psychol. 72 (1987) 146–148, https://doi.org/10.1037/0021-9010.72.1.146.
[32] J. Dréo, A. Pétrowski, P. Siarry, E. Taillard, Metaheuristics for Hard Optimization Methods and Case Studies, 1st edition, Springer, Berlin Heidelberg, 2006, https://doi.org/10.1007/3-540-30966-7.
[33] P.M. Pardalos, H.E. Romeijn, H. Tuy, Recent developments and trends in global optimization, J. Comput. Appl. Math. 124 (1) (2000) 209–228, https://doi.org/10.1016/S0377-0427(00)00425-8.
[34] M.D.S.G. Tsuzuki, T.D.C. Martins, Simulated Annealing: Strategies, Potential Uses and Advantages, Mathematics Research Developments, Nova Science Publishers, Inc, 2014.
[35] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, E. Teller, Equation of state calculations by fast computing machines, J. Chem. Phys. 21 (6) (1953) 1087–1092, https://doi.org/10.1063/1.1699114.
[36] O. Ashrafi, L. Yerushalmi, F. Haghighat, Wastewater treatment in the pulp-and-paper industry: a review of treatment processes and the associated greenhouse gas emission, J. Environ. Manag. 158 (2015) 146–157, https://doi.org/10.1016/j.jenvman.2015.05.010.
[37] G. Wang, Q.-S. Jia, M. Zhou, J. Bi, J. Qiao, A. Abusorrah, Artificial neural networks for water quality soft-sensing in wastewater treatment: a review, Artif. Intell. Rev. 55 (1) (2022) 565–587, https://doi.org/10.1007/s10462-021-10038-8.