Application of unsupervised learning and process simulation for energy optimization of a WWTP under various weather conditions

Sina Borzooei, Gisele H. B. Miranda, Soroush Abolfathi, Gerardo Scibilia, Lorenza Meucci and Maria Chiara Zanetti

ABSTRACT

This paper outlines a hybrid modeling approach to facilitate weather-based operation and energy optimization for the largest Italian wastewater treatment plant (WWTP). Two clustering methods, *K*-means algorithm and Gaussian mixture model (GMM) based on the expectation-maximization (EM) algorithm, were applied to an extensive dataset of historical and meteorological records. This study addresses the problem of determining the intrinsic structure of clustered data when no information other than the observed values is available. Two quantitative indexes, namely the Bayesian information criterion (BIC) and the Silhouette coefficient using Euclidean distance, as well as two general criteria, were implemented to assess the clustering quality. Furthermore, seven weather-based influent scenarios were introduced to the process simulation model, and sets of aeration strategies are proposed. The results indicate that incorporating weather-based aeration strategies in the operation of the WWTP improves plant energy efficiency.

Key words | cluster analysis, clustering validation, energy optimization, expectation-maximization algorithm, Gaussian mixture models, *K*-means algorithm

HIGHLIGHTS

- A hybrid modeling approach was proposed to improve the energy efficiency of the largest Italian WWTP under various weather conditions.
- Application of *K*-means and Gaussian mixture model was evaluated for weather-based cluster analysis of historical and meteorological data of the WWTP.
- A robust clustering performance evaluation was carried out by the use of Bayesian information criterion, Silhouette coefficient, and two general criteria.
- The best case-specific clustering model was used to study the impact of various weather conditions on the performance of the WWTP using the calibrated process simulation model.
- The incorporation of weather-based aeration strategies can improve the energy efficiency of the WWTP.

Sina Borzooei (corresponding author)
Maria Chiara Zanetti
Department of Environment, Land and Infrastructure Engineering (DIATI),
Politecnico di Torino,
Corso Duca degli Abruzzi, Torino 10129,
Italy
E-mail: *sina.borzooei@polito.it*

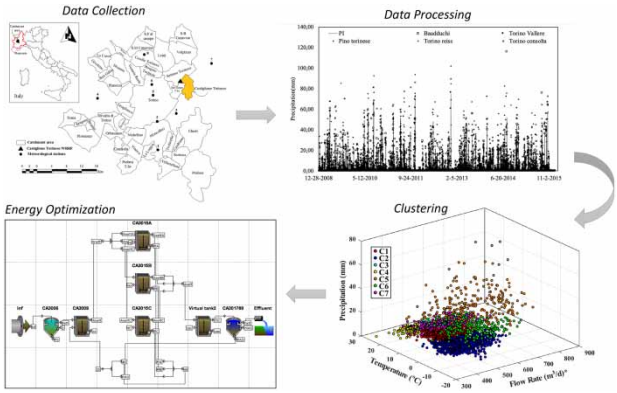
Gisele H. B. Miranda
School of Electrical Engineering and Computer Science,
KTH Royal Institute of Technology,
Lindstedtsvägen 3, Stockholm 10044,
Sweden
and
Science for Life Laboratory,
Tomtebodavägen 23A, Solna 17165,
Sweden

Soroush Abolfathi
Warwick Water Research Group,
School of Engineering, University of Warwick,
Coventry CV4 7AL,
UK

Gerardo Scibilia
Lorenza Meucci
SMAT (Società Metropolitana Acque Torino) Research Center,
Corso Unità d'Italia 235/3, Torino 10127,
Italy

doi: 10.2166/wst.2020.220

GRAPHICAL ABSTRACT



INTRODUCTION

Variation in weather conditions can profoundly influence the performance of wastewater treatment plants (WWTPs) due to fluctuations in the influent hydraulic characteristics and organic loads as well as thermal changes (Vo *et al.* 2014). At the operational level, preparedness and prompt response to weather changes have rarely been observed. Despite the recent advances in real-time monitoring, instrumentation, and control of WWTPs (Campisano *et al.* 2013), a large number of utilities continue to operate in relatively static mode for a wide range of weather and influent conditions (Zhu *et al.* 2015). The efficient and adaptive management of WWTPs under the influence of various weather conditions is a complex and challenging task given that an efficient management strategy should offer a trade-off between energy consumption and stringent regulations on effluent limits while handling the weather-related fluctuations. Coupling the available mechanistic mathematical models of the wastewater treatment process (Monod 1949; Dold *et al.* 1981; Henze *et al.* 2000) and sewer systems (Alley & Smith 1982; Rossman 2010) with weather data can provide valuable information about resultant dominant impacts of weather conditions on the performance of receiving treatment systems. However, continuous monitoring of WWTPs is required to capture the temporal patterns of key parameters included in these mechanistic models. This usually can be performed using high-precision and advanced sensors that are associated with high cost, as well as labour-intensive procedures to conduct necessary sampling campaigns. The problems associated with quality and quantity of monitoring data are still the most prevalent bottlenecks for more wide-spread use of WWTP models (Martin & Vanrolleghem 2014; Borzooei *et al.* 2016, 2019a). Although projected changes in weather conditions are regionally specific and highly uncertain, some general patterns can be obtained by investigating the metrological data collected in the catchment area of WWTPs. Several linear and non-linear data-driven methods have been proposed to characterize patterns in a plant's historical data. Several studies (Karagozoglu & Altin 2003; Mines *et al.* 2007) implemented linear data-driven approaches to identify empirical relationships between environmental data and influent flowrate and wastewater characteristics of WWTPs. However, due to the complex and multi-dimensional nature of the problem as well as spatial and temporal heterogeneity of environmental data, linear data-driven methods are not the best approach to study the impact of weather conditions on WWTPs. To this end, several non-linear data-driven methods have been implemented for knowledge discovery from various variables in water and wastewater treatment sectors (Karunasinghe & Liong 2006; Hu 2011; Abolfathi *et al.* 2016; Zhu *et al.* 2015, 2018).

Data clustering is a non-linear data-driven approach that can be implemented to partition spatial and temporal datasets. Data clustering is a general task that can be performed by several algorithms that can be categorized based on their definition of clusters and how to find them (Jain 2010). Clustering algorithms can be conceptually categorized into connectivity-based (e.g., hierarchical clustering), centroid-based (e.g., *K*-means clustering), distribution-based (e.g., Gaussian mixture models (GMM)) and density-based (e.g., density-based spatial clustering of

applications with noise (DBSCAN)). Since there is no objectively correct general-purpose clustering algorithm, the most appropriate clustering model needs to be identified by trial for a specific problem (Estivill-Castro 2002). The clustering results are highly sensitive to the choice of clustering algorithm and variables, initial assumptions, and data pre-processing (Hsu 2015).

This study adopts the *K*-means and GMM using the expectation-maximization (EM) algorithms for clustering meteorological and historical data of the largest Italian WWTP located at Castiglione Torinese, Italy. Further, a robust evaluation methodology is proposed to compare clustering models. Finally, the best case-specific clustering configuration was used with the calibrated process simulation model to study the impact of various weather conditions on the performance of the WWTP.

METHODS

Data collection and pre-processing

The analyzed database in this study was collected from the largest Italian WWTP located in Castiglione Torinese, Italy. The plant provides primary and secondary treatment using the biological nutrient removal activated sludge processes to treat approximately 590,000 m^3/d combined municipal and industrial wastewater of about 2.1 million of equivalent inhabitants. The dataset consists of volumetric influent flow-rate (Q_{in}), ammonium ($N\text{-}NH_4$), chemical oxygen demand (COD), and total suspended solids (TSS) concentrations collected by daily composite sampling from the influent of the plant from 2009 to 2016. Two environmental variables, namely precipitation (P_I) and temperature (T), were collected from the Piedmont Environmental Protection Agency. Daily spatial average values of P_I and T were obtained following the procedure reported in Borzooei *et al.* (2019c) from eight meteorological stations equipped with tipping-bucket rain gauges, located throughout the catchment area of 38 municipalities in the Piedmont region.

Prior to data clustering, data pre-processing was performed in two main steps: (i) screening of missing elements, outlier detection, and removal followed by data imputation and (ii) data normalization. Screening of the dataset to find the missing elements and outlier detection based on a statistical parametric approach were carried out using the *Stats* package in the R environment (R Core Team 2013). Further, the missing and/or removed observations were filled by the use of a cubic Hermite interpolation method (Catmull & Rom 1974) following the procedure reported in Borzooei *et al.* (2019b). The final step of data pre-processing was the normalization, which was performed by centralization by mean and scaling by the standard deviation method, as proposed in Zhu *et al.* (2015).

Finally, an extensive 20-day composite as well as two-day dynamic sampling campaigns were carried out in one of the wastewater treatment modules of Castiglione Torinese WWTP. The plant consists of four, almost the same, wastewater treatment modules, each resembling a typical modified Ludzack–Ettinger (MLE) activated sludge system with primary clarifiers. Due to restricted resources in the study, the decision was made to focus the modeling project on a wastewater module of the WWTP. Further details about the model development and its data collection process can be found in Borzooei *et al.* (2019a).

Clustering algorithms

Two clustering algorithms, *K*-means and GMM, were implemented to partition the two environmental variables (T, and P_I) as well as one WWTP-related attribute (Q_{in}). The *K*-means clustering algorithm using an iterative refinement technique was implemented in the *Stats* and *ClusterR* packages of the R environment (R Core Team 2013). The initialization or the selection of the initial centroids was performed using the *K*-means++ algorithm (Arthur & Vassilvitskii 2007). The squared (Euclidean) distance function was used as a similarity criterion, as it was found that other distance functions could lead to a diverging algorithm. The stop criterion was defined as the convergence of the mean and the standard deviation of intra- and inter-cluster distance. For further conceptual information about *K*-means clustering algorithms, Jain (2010) should be consulted.

Application of the *K*-means algorithm for clustering has several benefits such as its easy implementation, guaranteed convergence, relatively short computational time etc.; however, it has some critical drawbacks which should be considered before its implementation. Since the centroids are randomly chosen, different runs of *K*-means may yield very distinct clustering results. Besides, *K*-means does not perform well when the clusters are not round-shaped. Another issue with *K*-means is that the assignment of data-points to clusters is deterministic, i.e., a single observation will be assigned to a single cluster, even if this observation is in an overlapping region.

GMM, on the other hand, is a probabilistic clustering method that can overcome some of the *K*-means drawbacks.

This method provides a combination of K Gaussian distributions as a weighted sum of Gaussian density functions as follows (Reynolds 2009):

$$p(x|\lambda) = \sum_{i=1}^{M} w_i g(x|\mu_i, \Sigma_i) \quad (1)$$

where x corresponds to a D-dimensional vector composed of continuous data values, and w_i are the mixture weights, which should satisfy $\sum_{i=1}^{M} w_i = 1$. The Gaussian densities are represented by $g(x|\mu_i, \Sigma_i)$, each component being described by the following D-variate function:

$$g(x|\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma_i|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(x-\mu_i)' \sum_{i}^{-1}(x-\mu_i)\right\} \quad (2)$$

where μ_i is the mean vector and Σ_i accounts for the covariance matrix. A Gaussian mixture model is parametrized by $\lambda = \{w_i, \mu_i, \Sigma_i\}$, such that $i = 1, \ldots, M$.

Therefore, instead of identifying clusters only by their centroids, a set of probability distributions are fitted to the data observations. Consequently, the assumption that the observations are Gaussian-distributed is less restrictive than assuming that the clusters are round-shaped. Also, in addition to the number of centroids, two parameters are used to describe each cluster: mean (centroid) and standard deviation. Cluster methods based on GMM, as well as other statistical models, work iteratively to find the best parameter fit to a dataset. For this purpose, GMM makes use of an optimization method. Although several techniques can be employed for estimating λ, the expectation-maximization (EM) algorithm is often used to obtain the maximum likelihood estimator (Reynolds 2009; Steele & Raftery 2010). The EM starts with an initial model λ, and iteratively estimates a new model parametrized by $\bar{\lambda}$, such that $p(X|\bar{\lambda}) \geq p(X|\lambda)$, p being the GMM likelihood and X the training vectors. As such, the GMM clustering approach can be summarized as the following steps:

- Random selection of Gaussian parameters and fitting them to the data-points.
- Iterative optimization of the Gaussian parameters.
- Assignment of data-points to their closest distributions, after convergence of local minima.

Clustering performance evaluation

In contrast to supervised learning, clustering analyses do not have general and sound evaluation matrices to assess the results obtained from implementing various clustering algorithms (Duda *et al.* 2012; Machicao *et al.* 2018). There are two crucial measures to assess clustering performance: the optimal number of clusters and clustering quality. The two clustering algorithms implemented in this study required setting the number of clusters (K) in advance. Since identifying the conceptually right K value is a non-trivial problem, the performance of these two algorithms for different K values was evaluated and compared. Therefore, the dataset was clustered by both algorithms with K ranging from 2 to 10 with enough iterations to achieve stable and convergent results. An approximation to the Bayes factor by Bayesian information criterion (BIC), as well as the Silhouette coefficient using Euclidean distance, were further implemented to compare the models. Using BIC and the Silhouette coefficient to assess clustering quality can avoid the compared models being nested. Consequently, both issues of determining K and the optimum clustering algorithm were solved simultaneously by choosing the best model.

Bayesian methods have been widely used to select the number of components in the finite mixture of models since frequentist inferences could not sufficiently address this issue (Steele & Raftery 2010). The BIC, proposed by Schwarz (1978), is a criterion for model selection, which can provide an approximation to the integrated likelihood over the model parameters. The model with the lowest BIC is preferred among a finite set of models. The Silhouette coefficient is the average of a series of quantitative measures, ranging from -1 to 1, which evaluates the cohesion and separation characteristics of attributes within the cluster (Rousseeuw 1987). A high value of the Silhouette coefficient of a cluster indicates that the objects are well matched to the cluster and poorly matched to other neighboring clusters.

In addition to the abovementioned quantitative criteria, two further measures as general criteria were considered in the clustering evaluation: first, the resulting cluster from each model was checked for representing the distinctive and meaningful characteristics in accordance with the general objectives of the study, and second, to avoid too low and high population in clusters, the number of attributes in each cluster was monitored to be less than 95% and more than 5% of a total number of observations.

Wastewater treatment process simulation

Wastewater treatment processes in Castiglione Torinese WWTP were modeled using GPS-X ver. 6.5.1 simulator (Snowling 2014). The modeling boundary conditions were set according to the available plant information, including

data collected during sampling campaigns. The modeling project was conducted for half of a single wastewater treatment module, which resembles the MLE configuration with a primary clarifier. The developed model consists of biokinetic activated sludge model no. 1 (Henze *et al.* 2000), aeration, hydraulic and transport approximated by a 'tanks-in-series' approach, primary and secondary clarifier, influent, effluent, and energy consumption sub-models. The implemented sub-models for treatment processes in the Castiglione Torinese WWTP are summarized in Table 1.

Two performance assessment criteria (PAC) were considered for this study to evaluate the model performance under various weather-based scenarios. Firstly the energy-based PAC was introduced as the daily delivered power blower in the aeration units, modeled by the adiabatic compression equation (Mueller *et al.* 2002). The airflow rate in the aeration units was modeled by linear proportional-integral (PI) controllers calibrated based on the measured dissolved oxygen (DO) data. Secondly, a new function of the net moving average effluent quality index ($EQI_{n\text{-}a}$) (kg pollution per unit time) was defined as follows:

$$EQI_{n-a} = \frac{1}{T.1000}\int_{t}^{t+T} Q_e(t)\sum_{i=1}^{n} w_i.\max\left[0,\ (C_i(t) - C_{i,\text{limit}})\right].d(t) \quad (3)$$

where $Q_e(t)$ is the effluent flow rate time function (m^3/d), n is the number of effluent quality parameters, $C_i(t)$ and $C_{i,\text{limit}}$ are the effluent concentration-time function (g/m^3) and limits respectively, w_i is the weight factor of the parameter i, and T is the period considered for the moving average calculation (d) (Borzooei *et al.* 2019b). Five effluent quality parameters, namely BOD_5, COD, TSS, TKN, and NO_3 were considered for the measurement of $EQI_{n\text{-}a}$ in this study, and their corresponding weights were assigned according to Nopens *et al.* (2010). The value of the $C_{i,\text{limit}}$ corresponding to each of the effluent quality parameters was considered based on EU Directive 91/271/EEC (EEC Council 1991).

Table 1 | Sub-models implemented in the developed model for Castiglione Torinese WWTP

Unit process	Process model
Influent wastewater	COD states influent (Snowling 2014)
Primary settling	Ideal clarifier (Snowling 2014)
Pre-denitrification	ASM1 (Henze *et al.* 2000)
Aeration system	ASM1 (Henze *et al.* 2000)
Secondary clarification	Simple 1-D (Takács *et al.* 1991)
Denitrification in secondary clarifiers	ASM1 (Henze *et al.* 2000)
Energy consumption	Operating cost (Snowling 2014)

Eventually, the developed model was calibrated following an iterative step-wise calibration procedure reported in Borzooei *et al.* (2019a), where the calibration parameters were identified based on sensitivity analysis, available calibration protocols and full-scale observations. For the aeration units, the α factors (ratio of process water to clean water mass transfer coefficients) were numerically tuned based on the recorded DO and airflow data collected in the sampling period. For calibration of the aeration energy model, the pressure drop in piping and diffuser downstream of the blower (ΔP_a) and the combined blower and motor efficiency (e) were numerically adjusted based on the energy audit data reported in Panepinto *et al.* (2016).

RESULTS AND DISCUSSION

Weather-based clustering approach

Following the data pre-processing steps, amongst the total 17,520 records collected in 2,920 days, 57 days with missing values and 118 days with outliers were identified; thus, data of 2,745 days were further analyzed for clustering. Two weather-related attributes, namely P_I and T as well as Q_{in} were considered for the clustering approach. For a given number of clusters from two to ten, K-means and GMM algorithms were employed to generate the clusters. Each cluster was further evaluated initially by BIC index as a model selection criterion followed by the average Silhouette coefficient. Figure 1 depicts the results of the clustering evaluation for both methods adopted in this study.

Figure 1(a) shows the variation of BIC value as the number of clusters increases; for K-means the BIC value steadily declines with increasing K as the likelihood saturates. However, for the GMM, the BIC value declines in the beginning, followed by a jump from $K = 7$ to $K = 8$, indicating a higher degree of complexity associated with GMM. Additionally, the lower value of the BIC in GMM compared with K-means for $K = 7$, 9 and 10 is a measure of the posterior probability and trueness of the GMM in these conditions (Schwarz 1978). By looking at the BIC, which is a model selection criterion, and following the two general clustering evaluation criteria defined previously, the ideal value of K is 7, since it is the lowest BIC given $K \leq 7$.

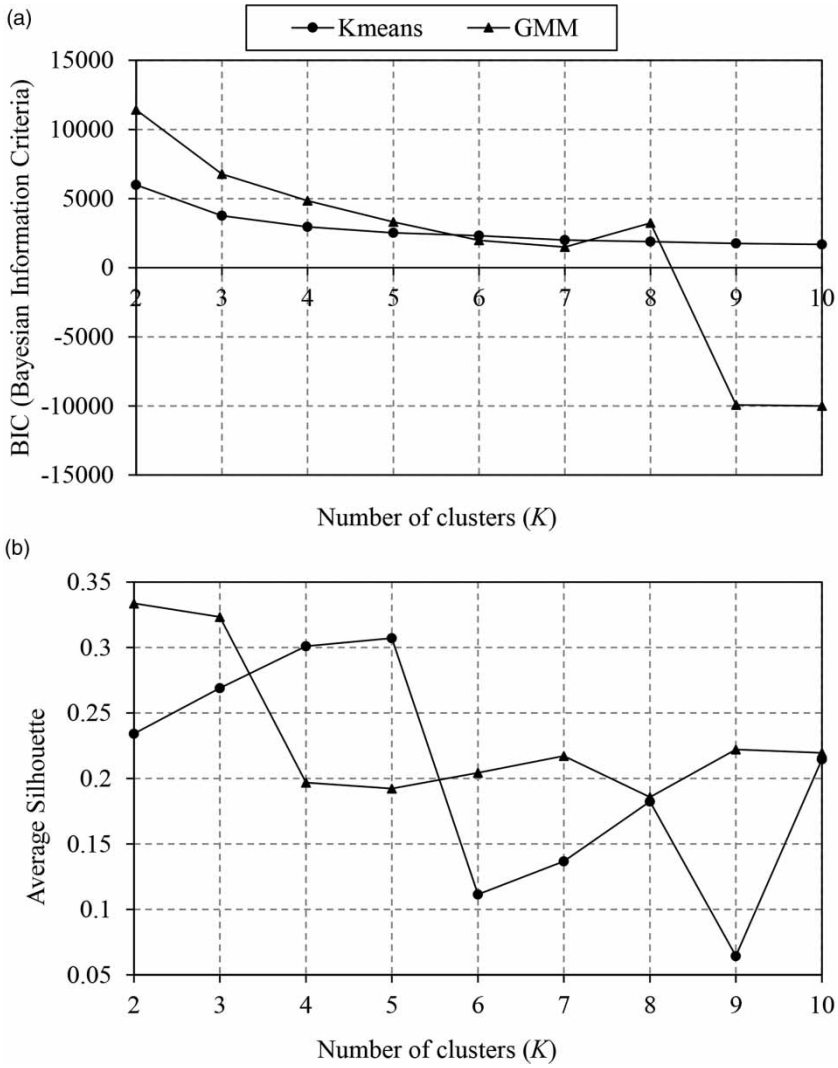


Figure 1 | Clustering evaluation: (a) BIC and (b) average Silhouette.

When $K > 7$, some clusters only account for a small fraction of the data resulting in too low a population in some of the clusters, which does not meet the general clustering evaluation criteria explained before.

Figure 1(b) shows the variation of average Silhouette as K increases in both algorithms. The Silhouette coefficient indicates a quantitative interpretation of cohesion and separation of each object in its cluster. The higher the value of the Silhouette, the better the clustering configuration. Figure 1(b) shows that the maximum value of the Silhouette index in the K-means algorithm was observed for $K = 5$, which confirms the viability of the results presented in the previous study (Borzooei *et al.* 2019b). However, for the GMM algorithm, the maximum value of the Silhouette index was observed for $K = 2$ and 3, followed by 9 and 7. The GMM with two clusters ($K = 2$) could only differentiate between days in wet-weather conditions (average $P_I \approx 18.1$ mm) and dry-weather (average $P_I \approx 0$ mm). Although this model had the highest average Silhouette value, the observations in the dry-weather cluster showed a very high standard deviation, confirming the necessity for further partitioning. Additionally, the $K = 2$ model could not cluster observations based on the temperature clustering attributes. The GMM $K = 3$ included two clusters representing dry-weather conditions, which were distinguished by temperature attribute and one cluster related to extreme wet weather. This model was rejected since it could not capture the difference between extreme and normal wet-weather conditions. GMM with $K = 9$ was rejected because it provides clusters including data with

similar characteristics or accounted for less than 5% of the total dataset (general criteria for clustering evaluation). It was eventually decided to use $K = 7$, because this configuration shows a lower BIC for both GMM and K-means. At the same time, the GMM Silhouette index, given $K = 7$, is higher than the same index for K-means clustering. Consequently, the entire dataset was classified into seven weather-based clusters using the GMM algorithm (Figure 2).

Given the importance of the P_I attribute in clustering, the algorithm detects cluster 5 (C5) containing 175 days with extreme wet-weather conditions, identified by average $P_I = 20.9$ mm and consequently very high flowrates, with average temperature. Clusters 6 and 7 (C6 and C7) correspond to 352 wet-weather days with average $P_I = 5.3$ and 3.7 mm, respectively; however, they are distinguished by the temperature. Whilst the average temperature of C6 is the second-lowest among all the clusters (average $T = 9.3\,^{\circ}C$), the temperature of C7 is relatively high (average $T = 21.1\,^{\circ}C$). Cluster 4 (C4) includes 175 days with absolute dry-weather conditions ($P_I \approx 0$ mm) and very high temperature (the highest among the seven clusters), while the flowrate is low. Cluster 2 (C2) and cluster 3 (C3) with 885 and 356 observations are the dry-weather clusters with low and high temperatures, which correspond to low and high flowrate, respectively. Finally, cluster 1 (C1) includes days in dry-weather conditions with medium temperature and flowrate. Table 2 summarizes the average and standard deviation of variables in each cluster.

Comparing the wet-weather clusters (C5, C6, and C7), one can observe that there are positive relations between P_I and Q_{in} in all the clusters. The dilution effect due to wet-weather events prevails only in C5, whereas in C6, all the influent compositions are more likely to be higher than most of the dry-weather clusters. High influent concentrations in C6 and C2 at low temperatures could be due to wastewater production patterns in the catchment area, such as inhabitants traveling in summer. For instance, on days with dry weather and high temperature, the expected influent compositions include relatively medium to low TSS, COD, and NH_4 concentrations. However, it should be noted that among the three influent compositions, COD and NH_4 showed relatively more sensitivity towards the P_I variations in comparison with TSS, which could be the impact of the first flush event (Gupta & Saul 1996).

Weather-based process optimization

The developed and calibrated process model was implemented to assess how the knowledge about the weather-based influent scenarios could facilitate process control and improve process efficiency in the Castiglione WWTP. Several steady-state simulations were performed under the initial conditions set according to weather-based influent scenarios identified by clustering. These influent scenarios contain variations of average Q_{in} and three influent compositions (COD, TSS, and NH_4), as well as air temperature. Two important parameters for simulating the biological activities and aeration energy consumptions, namely wastewater temperature and air temperature at the inlet of the blower in the aeration units, were altered for each scenario based on experience-based empirical relations with air temperature. Scenario-based process optimization was designed to find an optimal or non-dominated solution offering a trade-off between energy consumption in aeration units and effluent quality. Initially, DO set-points of the implemented and calibrated PI controllers in the aeration units were identified as the most sensitive parameters in aeration energy by use of systematic sensitivity analysis (see Borzooei *et al.* (2019a)). For each influent scenario, sets of DO set-points were identified by the use of the Nelder–Mead simplex (polyhedron) algorithm (Nelder & Mead 1965) to minimize the aeration energy consumption while managing the $EQI_{n-a} = 0$ accounting for the operational condition in which all the considered effluent quality parameters are above the limits of EU Directive 91/271/EEC. Further, the weather-based dynamic operation was compared with the aeration strategies of the plant during the sampling period (static DO set-points with manual changes to manage fluctuations) in terms of energy consumption of aeration units in one wastewater treatment module as can be seen in Figure 3.

The simulation results presented in Figure 3 suggest that possible aeration energy savings range from about 4.1% to 6.8% in seven influent scenarios based on clusters. It should be noted that the dynamic operation results account for the conditions with no violation of the effluent constraints; hence, all the proposed strategies can enhance the energy efficiency of the aeration units while meeting all the effluent limits. In general, the highest aeration savings were associated with cold weather as the highest aeration energy consumption and saving was obtained for influent scenarios related to C2 followed by C6. This could be because of the wastewater temperature impact on DO solubility; however, nutrient loading is also a critical point to be considered. On the other hand, the lowest aeration energy savings were associated with warm-weather conditions such as C7 and C3, mainly due to their relatively lower influent compositions.

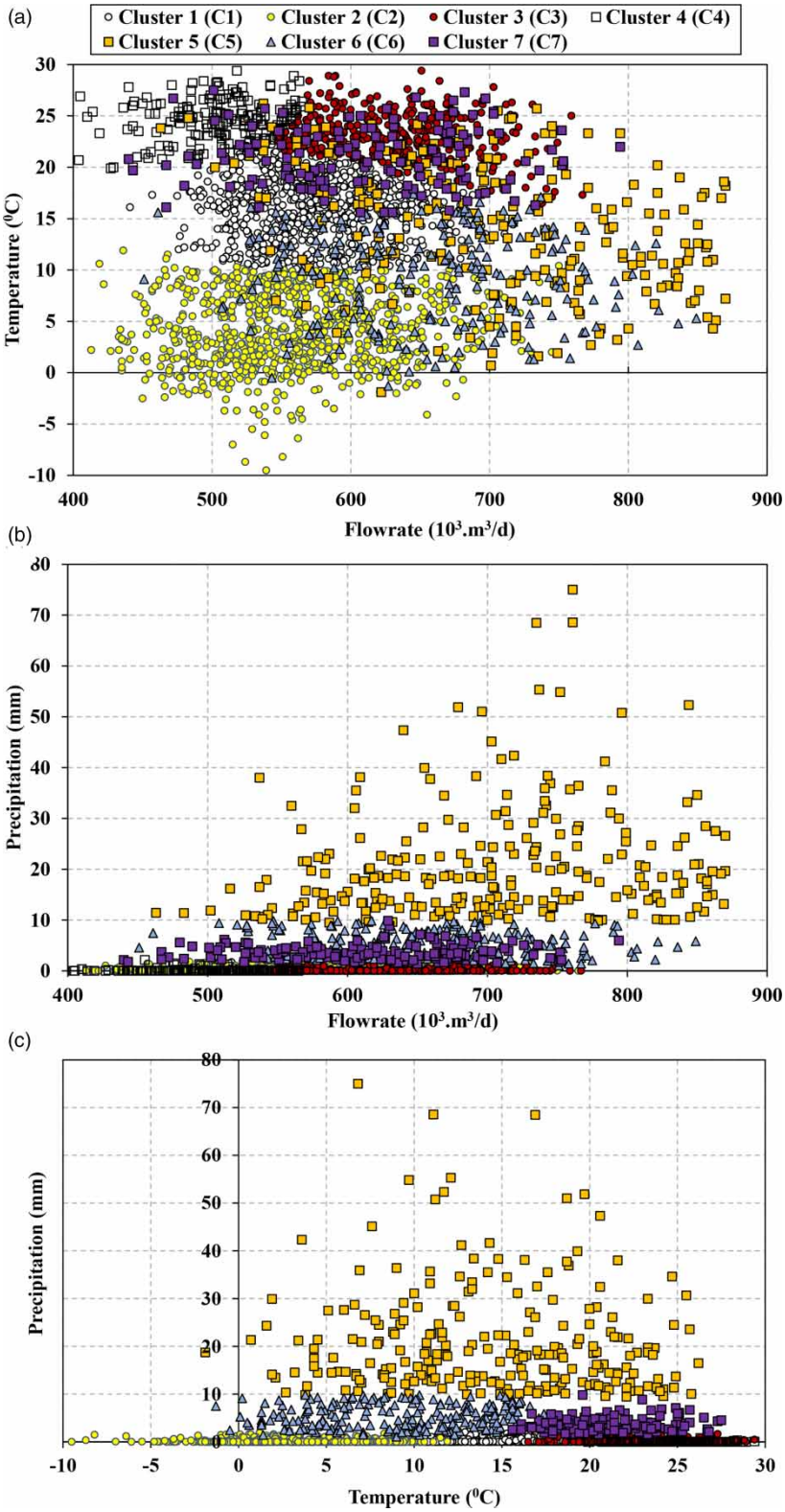


Figure 2 | Visualization of $K = 7$ clustering for GMM: (a) Q_{in} vs T, (b) Q_{in} vs P_I and (c) T vs P_I.

Table 2 | Average (±standard deviation) of attributes in seven clusters obtained by the GMM algorithm

Cluster \ Unit	Precipitation (P_I) (mm)	Temperature (T) (°C)	Flowrate (Q_{in}) ($10^3 \cdot m^3/d$)	Total suspended solids (TSS) (mg/l)	Chemical oxygen demand (COD) (mg/l)	Ammonia (NH_4) (mg/l)
C1	0.1 ± 0.4	15.7 ± 2.9	574 ± 46	186.7 ± 56.1	403.7 ± 93.9	22.7 ± 4.7
C2	0.1 ± 0.1	4.2 ± 3.7	561.7 ± 64.2	200.9 ± 55.9	457.9 ± 97.5	27.4 ± 4.5
C3	0.1 ± 0.3	23.1 ± 2.3	633.4 ± 48.6	182.3 ± 58.6	364.2 ± 95.4	19.5 ± 3.3
C4	≈0.0	24.6 ± 2.0	502.8 ± 39.1	151.6 ± 53.1	314.1 ± 92.1	17.1 ± 4.0
C5	20.9 ± 11.2	14.3 ± 6.1	702.9 ± 93.6	153.9 ± 56.3	291.8 ± 91.6	16.2 ± 5.1
C6	5.3 ± 2.4	9.3 ± 4.5	655.9 ± 76.5	187.8 ± 54.1	406.4 ± 108.2	23.1 ± 6.1
C7	3.7 ± 1.5	21.1 ± 2.8	613.5 ± 73.1	175.3 ± 65.6	361.4 ± 113.9	18.9 ± 4.1

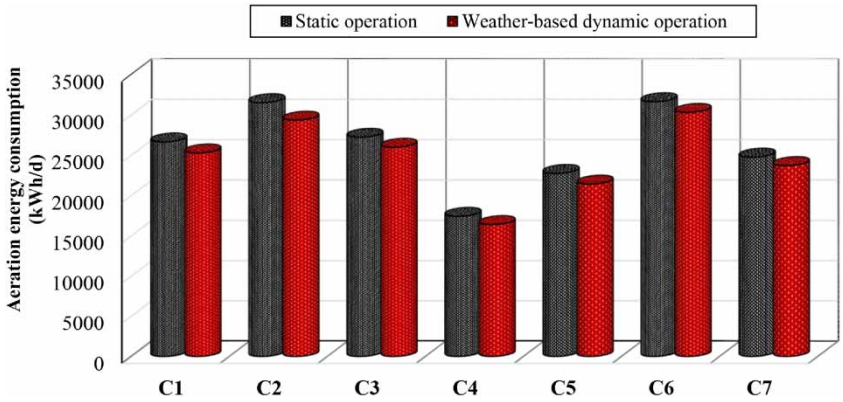


Figure 3 | Potential aeration energy savings in steady-state conditions of the seven influent scenarios for one wastewater treatment module of Castiglione Torinese WWTP.

CONCLUSIONS

This study presented the application of *K*-means and Gaussian mixture clustering algorithms to assess the impact of weather variations on WWTP influent characteristics. An average Silhouette coefficient and Bayesian Information Criterion were used as quantitative clustering evaluation measures to select the clustering algorithm and the optimum number of clusters. The results of the clustering evaluation confirmed that $K = 7$ for the GMM algorithm provides the best clustering configuration. Further analysis of the selected configuration highlighted the importance of precipitation rate (P_I), by categorizing the observation days into dry (clusters C1, C2, C3, and C4), wet (clusters C6 and C7), and extreme wet (cluster C5) weather conditions. The results also showed that the temperature (T) attribute is categorized into low, medium, and high conditions. Clusters with high P_I have high influent flowrate (Q_{in}) except for the case of C3, which represents dry weather, high T, and Q_{in}. Dry-weather and low-temperature days (C2), accounting for almost 30% of observations, deserve a special note since they include relatively higher influent TSS, COD, and NH_4 concentrations. Extreme wet-weather days in cold and warm conditions (C5) identified by $P_I \geq 10$ mm are of importance for operational preparedness since the Q_{in} is likely to be double on these days with a dilution effect in all influent compositions. Finally, seven influent scenarios based on clustering results were developed and introduced to the calibrated process simulation model. A series of steady-state model-based optimization scenarios were proposed to reduce the energy consumption of the aeration units in Castiglione Torinese WWTP. The results from the process simulation model

confirmed that aeration energy consumption among the seven influent scenarios could be reduced up to 6.8%.

REFERENCES

Abolfathi, S., Yeganeh-Bakhtiary, A., Hamze-Ziabari, S. M. & Borzooei, S. 2016 Wave runup prediction using M5′ model tree algorithm. *Ocean Eng.* **112**, 76–81.

Alley, W. M. & Smith, P. E. 1982 *Distributed Routing Rainfall-Runoff Model; Version II*. Open-File Report 82-344, US Geological Survey, Reston, VA, USA.

Arthur, D. & Vassilvitskii, S. 2007 k-means^{++}: the advantages of careful seeding. In: *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, Society for Industrial and Applied Mathematics, PA, USA, pp. 1027–1035.

Borzooei, S., Zanetti, M., Genon, G., Ruffino, B., Godio, A., Campo, G., Panepinto, D., Lorenzi, E., De Ceglia, M. & Binetti, R. 2016 Modelling and calibration of the full scale WWTP with data scarcity. In: *Proceedings of International Symposium on Sanitary and Environmental Engineering*, Rome, Italy.

Borzooei, S., Amerlinck, Y., Abolfathi, S., Panepinto, D., Nopens, I., Lorenzi, E., Meucci, L. & Zanetti, M. C. 2019a Data scarcity in modelling and simulation of a large-scale WWTP: stop sign or a challenge. *J. Water Process Eng.* **28**, 10–20.

Borzooei, S., Miranda, G. H. B., Teegavarapu, R., Scibilia, G., Meucci, L. & Zanetti, M. C. 2019b Assessment of weather-based influent scenarios for a WWTP: application of a pattern recognition technique. *J. Environ. Manage.* **242**, 450–456.

Borzooei, S., Teegavarapu, R., Abolfathi, S., Amerlinck, Y., Nopens, I. & Zanetti, M. C. 2019c Data mining application in assessment of weather-based influent scenarios for a WWTP: getting the most out of plant historical data. *Water Air Soil Pollut.* **230**, 5.

Campisano, A., Cabot Ple, J., Muschalla, D., Pleau, M. & Vanrolleghem, P. A. 2013 Potential and limitations of modern equipment for real time control of urban wastewater systems. *Urban Water J.* **10**, 300–311.

Catmull, E. & Rom, R. 1974 A class of local interpolating splines. In: *Computer Aided Geometric Design* (R. E. Barnhill & R. F. Riesenfeld, eds), Academic Press, New York, USA, pp. 317–326.

Dold, P. L., Ekama, G. A. & Marais, G. R. 1981 A general model for the activated sludge process. In: *Water Pollution Research and Development* (S. H. Jenkins, ed.), Pergamon Press, Oxford, UK, pp. 47–77.

Duda, R. O., Hart, P. E. & Stork, D. G. 2012 *Pattern Classification*. John Wiley & Sons, New York, USA.

EEC Council 1991 *Council Directive 91/271/EEC of 21 May 1991 Concerning Urban Waste-Water Treatment*. EEC Council Directive.

Estivill-Castro, V. 2002 Why so many clustering algorithms: a position paper. *ACM SIGKDD Explor. Newsl.* **4** (1), 65–75.

Gupta, K. & Saul, A. J. 1996 Specific relationships for the first flush load in combined sewer flows. *Water Res.* **30**, 1244–1252.

Henze, M., Gujer, W., Mino, T. & van Loosdrecht, M. C. 2000 *Activated Sludge Models ASM1, ASM2, ASM2d and ASM3*. IWA Publishing, London, UK.

Hsu, D. 2015 Comparison of integrated clustering methods for accurate and stable prediction of building energy consumption data. *Appl. Energy* **160**, 153–163.

Hu, W. X. 2011 The application of artificial neural network in wastewater treatment. In: *2011 IEEE 3rd International Conference on Communication Software and Networks*, IEEE, Piscataway, NJ, USA, pp. 338–341.

Jain, A. K. 2010 Data clustering: 50 years beyond K-means. *Pattern Recognit. Lett.* **31**, 651–666.

Karagozoglu, B. & Altin, A. 2003 Flow-rate and pollution characteristics of domestic wastewater. *Int. J. Environ. Pollut.* **19**, 259–270.

Karunasinghe, D. S. K. & Liong, S.-Y. 2006 Chaotic time series prediction with a global model: artificial neural network. *J. Hydrol.* **323**, 92–105.

Machicao, J., Corrêa Jr, E. A., Miranda, G. H. B., Amancio, D. R. & Bruno, O. M. 2018 Authorship attribution based on life-like network automata. *PLoS One* **13** (3), e0193703.

Martin, C. & Vanrolleghem, P. A. 2014 Analysing, completing, and generating influent data for WWTP modelling: a critical review. *Environ. Model. Softw.* **60**, 188–201.

Mines, R. O., Lackey, L. W. & Behrend, G. H. 2007 The impact of rainfall on flows and loadings at Georgia's wastewater treatment plants. *Water Air Soil Pollut.* **179**, 135–157.

Monod, J. 1949 The growth of bacterial cultures. *Annu. Rev. Microbiol.* **3**, 371–394.

Mueller, J. A., Boyle, W. C. & Pöpel, H. J. 2002 *Aeration: Principles and Practice*. Water Quality Management Library Volume 11, CRC Press, Boca Raton, FL, USA.

Nelder, J. A. & Mead, R. 1965 A simplex method for function minimization. *Comput. J.* **7**, 308–313.

Nopens, I., Benedetti, L., Jeppsson, U., Pons, M.-N., Alex, J., Copp, J. B., Gernaey, K. V., Rosen, C., Steyer, J.-P. & Vanrolleghem, P. A. 2010 Benchmark Simulation Model No. 2: finalisation of plant layout and default control strategy. *Water Sci. Technol.* **62**, 1967–1974.

Panepinto, D., Fiore, S., Zappone, M., Genon, G. & Meucci, L. 2016 Evaluation of the energy efficiency of a large wastewater treatment plant in Italy. *Appl. Energy* **161**, 404–411.

Reynolds, D. A. 2009 Gaussian mixture models. In: *Encyclopedia of Biometrics* (S. Z. Li & S. Jain, eds), Springer, Boston, MA, USA, pp. 659–663.

Rossman, L. A. 2010 *Storm Water Management Model: User's Manual, Version 5.0*. EPA/600/R-05/040, US Environmental Protection Agency, Cincinnati, OH, USA.

Rousseeuw, P. J. 1987 Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65.

Schwarz, G. 1978 Estimating the dimension of a model. *Ann. Stat.* **6**, 461–464.

Snowling, S. 2014 *GPS-X Technical Reference*. Hydromantis Environmental Software, Hamilton, ON, Canada.

Steele, R. J. & Raftery, A. E. 2010 Performance of Bayesian model selection criteria for Gaussian mixture models. *Front. Stat. Decis. Mak. Bayesian Analys.* **2**, 113–130.

Takács, I., Patry, G. G. & Nolasco, D. 1991 A dynamic model of the clarification-thickening process. *Water Res.* **25**, 1263–1271.

R Core Team 2013 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Vo, P. T., Ngo, H. H., Guo, W., Zhou, J. L., Nguyen, P. D., Listowski, A. & Wang, X. C. 2014 A mini-review on the impacts of climate change on wastewater reclamation and reuse. *Sci. Total Environ.* **494–495**, 9–17.

Zhu, J.-J., Segovia, J. & Anderson, P. R. 2015 Defining influent scenarios: application of cluster analysis to a water reclamation plant. *J. Environ. Eng.* **141**, 04015005.

Zhu, J.-J., Kang, L. & Anderson, P. R. 2018 Predicting influent biochemical oxygen demand: balancing energy demand and risk management. *Water Res.* **128**, 304–313.

First received 3 October 2019; accepted in revised form 27 April 2020. Available online 7 May 2020