Prediction of outlet dissolved oxygen in micro-irrigation sand media filters using a Gaussian process regression

Paulino J. García–Nieto[a,*], Esperanza García–Gonzalo[a], Jaume Puig–Bargués[b], Miquel Duran–Ros[b], Francisco Ramírez de Cartagena[b], Gerard Arbat[b]

[a]Department of Mathematics, Faculty of Sciences, University of Oviedo, 33007 Oviedo, Spain

[b]Department of Chemical and Agricultural Engineering and Technology, University of Girona, 17003 Girona, Catalonia, Spain

Abstract

Sand media filters are a key component of micro-irrigation systems since they help preventing emitter clogging, which greatly affects the system performance. Dissolved oxygen is an irrigation water quality parameter related to organic matter loading. Low values of dissolved oxygen can cause crop root hypoxia and, therefore, agronomic problems. Thus, an accurate prediction of dissolved oxygen values could be of great interest, especially if effluents are used in micro-irrigation systems. The aim of this study was to obtain a predictive model able to forecast the dissolved oxygen values at the outlets of sand media filters. In this study, a Gaussian process regression (GPR) model was used for predicting the output dissolved oxygen (DO_o) from data corresponding to 547 filtration cycles of different sand filters using reclaimed effluent. This optimisation technique involves kernel parameter setting in the GPR training procedure, which significantly influences the regression accuracy. To this end, the height of the filter bed, filtration velocity and filter inlet values of the electrical conductivity, dissolved oxygen, pH, turbidity and water temperature were monitored

[*]Corresponding author. Tel.: +34-985103417; fax: +34-985103354.
E-mail address: lato@orion.ciencias.uniovi.es (P.J. García–Nieto).

and analysed. The significance of each variable on filtration performance is presented and a model for forecasting the outlet dissolved oxygen obtained. Regression with optimal hyperparameters was performed and a coefficient of determination of 0.90 for DO_o was obtained when this new predictive GPR–based model was applied to the experimental dataset. Agreement between experimental data and the model confirmed the good performance of the latter.

Keywords: Gaussian process regression; Bayesian statistics; Machine learning techniques; Drip irrigation; Clogging; Effluents

Nomenclature

Abbreviations

ANN	Artificial neural network
DE	Differential evolution
DO	Dissolved oxygen
GPR	Gaussian process regression
GEP	Gene expression programming
R^2	Coefficient of determination
RBF	Radial basis function
SCADA	Supervisory control and data acquisition
SE	Squared-exponential
SVM	Support vector machine
v	Filtration velocity, m h^{-1}

Symbols

DO_i	Dissolved oxygen at filter inlet, mg l^{-1}

DO_o	Dissolved oxygen at filter outlet, mg l^{-1}
δ_{ij}	Kronecker delta function
ε	Additive white noise
ℓ	Length-scale for the RBF kernel
σ_f^2	Variance for the RBF kernel
σ_n^2	Gaussian noise variance

1. Introduction

The substitution of conventional irrigation water by reclaimed effluents in areas of low water availability is a common management strategy despite of its potential pollution and health hazards (Ait-Mouheb et al., 2018). Among the different irrigation techniques used, micro-irrigation shows several environmental and health advantages related mainly to the reduced effluent exposure to humans and plants. However, one of the most important disadvantages of applying effluents with micro-irrigation is emitter clogging which can cause irrigation nonuniformity and system failure (Trooien & Hills, 2007). In order to avoid emitter clogging, micro-irrigation systems require effective filtration (Nakayama, Boman, & Pitts, 2007) and sand media filters are the standard for the protection of micro-irrigation systems using effluents (Trooien & Hills, 2017).

The level of dissolved oxygen (DO) decreases with the increased organic matter, commonly present in wastewaters. So, DO, which can be determined easier and quicker using sensors, is an indicator of irrigation water quality. Low DO values in the irrigation water cause root oxygen deficiency, leading to low yields (Bhattarai, Midmore, & Pendergast, 2008) and low quality (Zhou, Zhou, Xu, Muhammad, & Li, 2019). Usually, DO increases through micro-irrigation systems, especially when water is released by the emitters (Maestre–Valero & Martínez-Álvarez, 2010). The DO increase is slight in sand media filters but it is considerably affected by the filter performance (Elbana, Ramírez de Cartagena, & Puig-Bargués, 2012; Solé–Torres, Puig–Bargués, Duran–Ros, Arbat, Pujol, & Ramírez de Cartagena, 2019b). Thus, the development of accurate models for forecasting DO at filter outlets can be very useful for the appropriate management of both sand filter performance and irrigation water quality. Optimal efficiency of drip irrigation systems is required for implementing smart irrigation techniques which aim to provide optimum use of the water resources (Canales-Ide, Zubelzu & Rodríguez-Sinobas, 2019).

In this regard, advanced techniques such as artificial neural networks (ANN) (Puig–Bargués, Duran-Ros, Arbat, Barragán, & Ramírez de Cartagena, 2012), gene expression programming (GEP) (Martí et al., 2013) and support vector machines (SVM) (García–Nieto, García–Gonzalo, Arbat, Duran–Ros, Ramírez de Cartagena, & Puig–Bargués, 2016) have been used for predicting the filtered volume and the value of dissolved oxygen at sand media filter outlets. Recently, other machine learning techniques such as gradient boosted regression have been applied to different aspects of the filter operation (García–Nieto et al. 2017, 2018).

Thus, the application of an innovative methodology that combines a Gaussian process regression (GPR) approach (Rasmussen, 2003; Kuhn & Johnson, 2018; Ebden, 2015) with a metaheuristic optimisation algorithm Differential Evolution (DE) (Storn & Price, 1997; Price, Storn, & Lampinen, 2005; Feoktistov, 2006; Chakraborty, 2008; Simon,

2013) to foretell the outlet dissolved oxygen in sand media filters used in microirrigation systems could be an interesting approach since this issue has not yet been yet addressed in previous investigations. GPR is a machine learning method developed on the basis of statistical and Bayesian theory. As a nonparametric regression method it can be considered a complex model with capability to model nonlinearities and variable interactions (Rasmussen, 2003; Ebden, 2015). When GPR is compared with other machine learning techniques, it has several advantages (Rasmussen & Williams, 2006): (1) it has an important generalisation capacity; (2) the hyperparameters in GPR can be self-adaptively calculated; and (3) the GPR outputs have clear probabilistic meaning. In this study, the DE method is applied to optimise the GPR hyperparameters. Previous researches show that GPR is an effective tool in many fields, such as irrigation mapping (Chen, Lu, Luo, Pokhrel, Deb, Huang, & Ran, 2018), wind engineering and industrial aerodynamics (Ma, Xu, & Chen, 2019), applied geophysics (Noori, Hassani, Javaherian, Amindavar, & Torabi, 2019), applied demography (Wu & Wang, 2018), psychology (Schulz, Speekenbrink, & Krause, 2018), mechanical engineering (Kong, Chen, & Li, 2018), environmental engineering (Liu, Yang, Huang, Wang, & Yoo, 2018), tracking and positioning (Ko, Klein, Fox, & Haehnelt, 2007a), deformation observation (Rogers & Girolami, 2016), system identification and control (Ko, Klein, Fox, & Haehnelt, 2007b) and so on. However, it has not been used for predicting micro-irrigation sand filter performance.

The main objective of the this study was to predict the outlet dissolved oxygen (DO_o) in sand media filters operating with reclaimed effluents by using Gaussian processes (GPs) in combination with the DE parameter optimisation technique.

The structure of this paper is organised as follows: section 2 introduces the experimental setup and variables involved in this study as well as the GPR method; section 3 describes the results obtained with this model by comparing the GPR results with the experimental measurements, including the importance of the input variables and validating the efficacy of the proposed approach; and finally, section 4 concludes this study with a list of main findings.

2. Materials and methods

2.1. Experimental setup

The experimental setup was composed of 3 media filters fed with the reclaimed effluent from the wastewater treatment plant of Celrà (Girona, Spain). Each filter had a different underdrain design: inserted domes (model FA-F2-188, Regaber, Parets del Vallès, Spain), arm collector (model FA1M, Lama, Gelves, Spain) and porous media (prototype designed by Bové et al. (2017) (see Fig. 1).

Silica sand CA-07MS (Sibelco Minerales SA, Bilbao, Spain) with an effective diameter (*De*, size opening which will pass 10% of the sand) of 0.48 mm and a coefficient of uniformity (ratio of the sizes opening which will pass 60% and 10% of the sand through, respectively) of 1.73 was used as filtration media in the three filters. Media heights of 200 and 300 mm, were tested for each filter.

Each of the three filters operated alone for 8 h per day each. Nominal filtration velocities 30 and 60 m h^{-1} were tested in each filter. Each combination of media height and filtration velocity was tested during 250 h. The filters were automatically backwashed when the pressure loss across them reached 50 kPa for more than 1 min. The backwashing was carried out for 3 min with previously filtered effluent that was chlorinated for achieving 4 ppm target chlorine concentration.

Filtered and backwashed effluent volumes, pressures across the filter and some effluent quality parameters before (pH, temperature, electrical conductivity, DO and turbidity) and after (only DO and turbidity) being filtered were measured and recorded every minute in a supervisory control and data acquisition system (SCADA) fully described by Solé-Torres et al. (2019a). Once the experiment started, the performance of the effluent quality sensors was assessed periodically by comparing its measurements with results obtained by manual sampling and, if necessary, they were calibrated following manufacturer recommendations.

Fig. 1 - Picture of the experimental set-up with the three filter designs: (a) red: arm collector; (b) blue: inserted domes; and (c) green: porous media prototype.

2.2. Variables involved in the model and materials tested

The main objective of this study was to compute the outlet dissolved oxygen as a function of different experimentally measured parameters that the GPR–based model needs as input. The output variable was the outlet dissolved oxygen (DO_o), which is an indicator of the quality of the filtered effluent and it is directly related to the organic load and the hypoxic risk of irrigation water.

The new predictive model used eight different operating variables commonly used for characterising sand media filter performance as input variables (see Table 1) (Puig-Bargués et al., 2012). After removing samples with missing data from the initial 637 samples, 547 satisfactory samples were obtained.

Table 1 - Set of operation physical input variables used in this study and their names along with their mean and standard deviation.

The operating input variables are as follows:

- Filter: three filter designs (porous, dome and arm collector underdrains) as described in section 2.1. This is a categorical variable.
- Height of the filter bed (mm): an operating variable for sand filters. Two different filter bed heights of 200 and 300 mm were tested for each filter.
- Filtration velocity (m h^{-1}): a operating variable related to filter operation. Two filtration velocities (30 and 60 m h^{-1}) were tested for each filter since these follow within the common range of velocities suggested by the manufacturers.
- Electrical conductivity (µS cm^{-1}): a general measure of water quality related to salinity, which is a constraint in microirrigation (Tal, 2016).
- Dissolved oxygen (mg l^{-1}): a variable related to the ability of water to support aerobic processes. This is a common parameter used for both controlling the biological treatment in wastewater plants and measuring irrigation water quality.
- pH: a measure of water acidity or alkalinity.

- Water temperature (ºC): temperature of the effluent at the filter inlet.
- Input turbidity (FNU): a key parameter for water quality that measures water clarity, which depends on suspended solid load.
- Filtered volume (m^3): a measure of the volume of effluent filtered in each filtration cycle.

2.3. Gaussian process regression (GPR)

GPRs are Bayesian state-of-the-art tools for discriminative machine learning (i.e., regression, classification, and dimensionality reduction). GPs assume that a GP prior governs the possible unobserved latent functions and the marginal likelihood of the latent function. Thus, a priori observations shape this to produce posteriori probabilistic estimates. Consequently, the joint distribution of training and test data is a multidimensional GP, and the predicted distribution is estimated by conditioning based on training data (Camps–Valls, 2016; Witten, Frank, Hall, & Pal, 2016).

To fix ideas, a Gaussian distribution is a probability distribution that explains the random variables including vectors and scalars. On the one hand, this kind of distribution is stated exactly through its mean and covariance: $x \sim N\left(\mu,\sigma^2\right)$. On the other hand, a GP can be seen as a generalisation of the Gaussian probability distribution and it applies over functions. From the functional space point of view, a GP is an ensemble of random variables, that is to say, any finite number having a joint Gaussian distribution.

2.3.1. The fundamentals of GPR

Let us assume that $D=\{(\mathbf{x}_i, y_i)/i=1,2,...,N\}$ depicts the training dataset of the Gaussian approach and the feature vectors $\mathbf{x}_i \in \Re^n$ comprises the extracted features or the merged features and the pertinent segregation parameters. The observed target values y_i reproduce the outlet dissolved oxygen measured in a filtration process, respectively. $X=\{\mathbf{x}_i\}_{i=1}^N$ depicts the input matrix of training dataset, $\mathbf{y}=\{y_i\}_{i=1}^N$ symbolises the output vector. A GP $f(\mathbf{x})$ defines a priori over functions, which can be converted into a posteriori over functions once some data is obtained. A GP can be fully stated exactly by using its mean function $m(\mathbf{x})$ and covariance function $k(\mathbf{x},\mathbf{x}')$. In this way, the Gaussian process is indicated as (Rasmussen & Williams, 2006; Marsland, 2014; Witten, Frank, Hall, & Pal, 2016):

$$f(\mathbf{x}) \sim GP(m(\mathbf{x}), k(\mathbf{x},\mathbf{x}')) \tag{1}$$

so that

$$\begin{aligned} m(\mathbf{x}) &= E[f(\mathbf{x})] \\ k(\mathbf{x},\mathbf{x}') &= E\left[(f(\mathbf{x})-m(\mathbf{x}))(f(\mathbf{x}')-m(\mathbf{x}'))^T\right] \end{aligned} \tag{2}$$

The mean function $m(\mathbf{x})$ depicts the anticipated value of the function $f(\mathbf{x})$ at the input point $\mathbf{x}$. The covariance function $k(\mathbf{x},\mathbf{x}')$ can be taken into account as a measurement of the confidence level for $m(\mathbf{x})$, and it is required that $k(\cdot,\cdot)$ be a positive definite kernel. In general, the mean function is set to be zero for notation simplicity, but this is also reasonable if there is no a priori knowledge about the mean variable, as is the case in this study.

The choice of the covariance function is critical for the GP. It describes the assumptions about the latent regression model and, therefore, is also referred to as the prior (Schneider & Ertel, 2010). In this research, the affine mean function and squared-exponential (SE) covariance function are expressed as follows (Shi & Choi, 2011; Witten, Frank, Hall, & Pal, 2016; Kuhn & Johnson, 2018):

$$k_{\mathrm{SE}}\left(\mathbf{x},\mathbf{x}'\right)=\sigma_f^2\exp\left(-\frac{\left\|\mathbf{x}-\mathbf{x}'\right\|^2}{2l^2}\right) \tag{3}$$

being l the characteristic length-scale and σ_f^2 the signal variance. The parameter selection of the SE covariance function has a direct effect on the performance of the GP. Here, l controls the horizontal scale over which the function changes, and σ_f^2 controls the vertical scale of the function.

The function values $f\left(\mathbf{x}\right)$ are not achievable in most applications. In practice, only the noisy observations are available and they are given by:

$$\mathbf{y}=f\left(\mathbf{x}\right)+\varepsilon \tag{4}$$

so that ε is the additive white noise. Besides, suppose that Gaussian noise is independent and identically distributed such that $\varepsilon\sim N\left(0,\sigma_n^2\right)$, where σ_n is the standard deviation of this noise. Any finite number of the observed values can also constitute an individual Gaussian process as given by (Witten, Frank, Hall, & Pal, 2016; Vidales, 2019):

$$\mathbf{y}\sim GP\left(m\left(\mathbf{x}\right),k\left(\mathbf{x},\mathbf{x}'\right)+\sigma_n^2\delta_{ij}\right)=GP\left(0,k\left(\mathbf{x},\mathbf{x}'\right)+\sigma_n^2\delta_{ij}\right) \tag{5}$$

where δ_{ij} is the Kronecker delta function described as:

$$\delta_{ij} = \begin{Bmatrix} 1 & \text{if} & i = j \\ 0 & & \text{otherwise} \end{Bmatrix}$$

The purpose of the GPR model is to foretell the function value $\bar{f}^*$ and its variance $\operatorname{cov}\left(f^*\right)$ given the new test point $\mathbf{x}^*$. In this sense, X^* depicts the input matrix of test dataset and N^* the size of test dataset. Taking into account the definition of GP, the observed values and the function values at new test points obey a joint Gaussian previous distribution which can be expressed as:

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}^* \end{bmatrix} \sim N\left(0, \begin{bmatrix} K(X,X)+\sigma_n^2 I & K\left(X,X^*\right) \\ K\left(X^*,X\right) & K\left(X^*,X^*\right) \end{bmatrix}\right) \tag{6}$$

where:

- $K(X,X)$: is the covariance matrix of training dataset;
- $K\left(X^*,X^*\right)$: is the covariance matrix of test dataset;
- $K\left(X,X^*\right)$: depicts the covariance matrix obtained from the training and test dataset. Furthermore $K\left(X^*,X\right) = K\left(X,X^*\right)^T$.

Since $\mathbf{y}$ and $\mathbf{f}^*$ are jointly distributed, it is possible to condition the prior on the observations (6) and determine how likely are predictions for $\mathbf{f}^*$. This can be expressed as:

$$\mathbf{f}^* \middle| X^*, X, \mathbf{y} \sim N\left(\bar{\mathbf{f}}^*, \operatorname{cov}\left(\mathbf{f}^*\right)\right) \tag{7}$$

where

$$\bar{\mathbf{f}}^* = E\left[\mathbf{f}^* \middle| X^*, X, \mathbf{y}\right] = K\left(X^*,X\right)\left[K(X,X)+\sigma_n^2 I\right]^{-1}\mathbf{y} \tag{8}$$

$$\operatorname{cov}\left(\mathbf{f}^*\right)=K\left(X^*,X^*\right)-K\left(X^*,X\right)\left[K\left(X,X\right)+\sigma_n^2 I\right]^{-1}K\left(X,X^*\right) \tag{9}$$

The subsequent distribution can be used for the forecast of new test input points. Indeed, $\overline{\mathbf{f}}^*$ is the predicted output value of the GPR model for test point. Additionally, confidence interval (CI) of the predicted output value can be calculated through the variance $\operatorname{cov}\left(\mathbf{f}^*\right)$. For instance, the 95% CI can be determined by $\left[\overline{\mathbf{f}}^*-2\times\sqrt{\operatorname{cov}\left(\mathbf{f}^*\right)},\overline{\mathbf{f}}^*+2\times\sqrt{\operatorname{cov}\left(\mathbf{f}^*\right)}\right]$. As a consequence, the GPR model not only supplies the predicted values but also furnishes the confidence level of the predicted results.

Finally, the GPR model is a nonparametric model since the predicted outputs rely only on the inputs and the observed values $\mathbf{y}$. In this way, parameters $\Theta=\left\{l,\sigma_f,\sigma_n\right\}$ are termed the hyperparameters of the GPR model.

2.3.2. Hyperparameter estimation

In order to carry out this study, the dataset was divided into a training set with 80% of the data, and a testing set with the remainder 20% of the data. A model was constructed and optimised with the training data. It was then tested with the test dataset and the optimisation of the parameters was performed with the help of the differential evolution (DE) technique.

The predictive performance of GPR model depends exclusively on the suitability of the chosen kernel. To estimate the kernel hyperparameters, an exhaustive search over a discrete grid of values can be used, but this can be quite slow. The most usual method considers an empirical Bayes approach that maximises the marginal likelihood. That is, the optimal hyperparameters are achieved by maximising the log marginal likelihood. The marginal likelihood $P(\mathbf{y}|X)$ is obtained, using Bayes' rule, as:

$$P(\mathbf{y}|X) = \int P(\mathbf{y}|f, X)P(f|X)df \tag{10}$$

The term marginal likelihood refers to the marginalisation over the function values $\mathbf{f}$. Since $\mathbf{y} \sim \mathcal{N}[0, K(X,X)]$, the log marginal likelihood can be written as:

$$\log p(\mathbf{y}\hat{\mathbf{u}}X) = -\frac{1}{2}\mathbf{y}K_y^{-1}\mathbf{y} - \frac{1}{2}\log \hat{\mathbf{u}}K_y\hat{\mathbf{u}} - \frac{N}{2}\log(2\pi) \tag{11}$$

where $K_y = K + \sigma_n^2 I, K = K(X,X)$ and $\hat{\mathbf{u}}$ is the determinant. In this expression, the first term is a data-fit term, the second term (always positive), and subtracted from it, is a model complexity penalty, and the last term is simply a normalisation constant. This expression therefore shows that the criterion of maximum marginal likelihood avoids the problem of over-fitting because if two models are explaining the observed data with the simplest one being chosen (Murphy, 2012; Witten, Frank, Hall, & Pal, 2016).

Following parameter initialisation, the optimal hyperparameters $\Theta' = \arg\max_{\Theta} \log p(\mathbf{y}|X, \Theta)$ can be calculated using any standard evolutionary optimiser. In this study, the metaheuristic optimisation algorithm, denominated the DE algorithm (Storn & Price, 1997; Price, Storn, & Lampinen, 2005; Feoktistov, 2006; Simon, 2013), was used. The process is shown in Fig. 2.

Fig. 2 – GPR Model selection using the DE optimisation technique.

2.4. The goodness–of–fit of this approach

Eight predicting variables were used (see section 2.2) to construct the new GPR–based model. The output predicted variable was the outlet dissolved oxygen. To predict the outlet dissolved oxygen from other input operating parameters, it is necessary to choose the model that best fits the experimental data. To determine the goodness–of–fit, the criterion considered here was the coefficient of determination R^2 (Picard & Cook, 1984; Freedman, Pisani, & Purves, 2007). A dataset takes values t_i, each of which has an associated modelled value y_i. The former are usually termed the observed values and the latter often referred to as the predicted values. The dataset variability is measured through different sums of squares as follows (Freedman, Pisani, & Purves, 2007):

- $SS_{tot} = \sum_{i=1}^{n} \left(t_i - \bar{t}\right)^2$: the total sum of squares, proportional to the sample variance.
- $SS_{reg} = \sum_{i=1}^{n} \left(y_i - \bar{t}\right)^2$: the regression sum of squares, also termed the explained sum of squares.
- $SS_{err} = \sum_{i=1}^{n} \left(t_i - y_i\right)^2$: the residual sum of squares.

Note that in the previous sums, $\bar{t}$ is the mean of the *n* observed data:

$$\bar{t} = \frac{1}{n} \sum_{i=1}^{n} t_i \tag{12}$$

Taking into account the above sums, the coefficient of determination is defined via:

$$R^2 \equiv 1 - \frac{SS_{err}}{SS_{tot}} \tag{13}$$

Thus, a coefficient of determination value of 1.0 indicates that the regression curve fits the data perfectly.

The value of R^2 was calculated using the optimised model with the testing dataset. The module Gpy from the Gaussian process framework found in Python (Gpy, 2014; Martin, 2018), along with the DE technique (Storn & Price, 1997; Price, Storn, & Lampinen, 2005; Simon, 2013) were used to construct the final regression model.

It is well known that the GPR technique depends strongly on the following hyperparameters (Friedman & Roosen, 1995; Aggarwal, 2015; Larose, 2015; Witten, Frank, Hall, & Pal, 2016; Tan, Steinbach, Karpatne, & Kumar, 2018):

- Variance (σ_f^2):the signal variance that controls the vertical scale of the kernel function.
- Lengthscale (ℓ):the characteristic length-scale that controls the horizontal scale over which the kernel function changes.
- Gaussian noise variance (σ_n^2): if ε is the additive white noise and the Gaussian noise is independent and identically distributed such that $\varepsilon \sim N\left(0,\sigma_n^2\right)$, then σ_n^2 is the variance of this noise.

1. A novel GPR–based model was constructed selecting as the dependent variable the outlet dissolved oxygen from the other eight remaining variables which were designated as input variables in the granular filters (Tien, 2012; Bové, Arbat,

Duran–Ros, Pujol, Velayos, Ramírez de Cartagena, & Puig–Bargués, 2015) and studying their effect in order to optimise calculation by analysing R^2.

As previously mentioned, this GPR technique is greatly dependent on the hyperparameters: variance (σ^2); lengthscale (ℓ) and the Gaussian noise variance (σ_n^2). The traditional way of performing hyperparameter optimisation has been *grid search*, or a *parameter sweep*, which is simply an exhaustive searching through a manually specified subset of the hyperparameter space of a learning algorithm. In this study, the metaheuristic optimisation algorithm, the DE algorithm (Storn & Price, 1997; Price, Storn, & Lampinen, 2005; Feoktistov, 2006; Simon, 2013) was used for multidimensional real-valued functions but it did not use the gradient of the problem being optimised, thus the DE did not require the optimisation problem to be differentiable, as is required by classic optimisation methods such as the gradient descent and quasi-Newton methods. Like other algorithms in this evolutionary category, the DE maintains a population of candidate solutions, which are recombined and mutated to produce new individuals which are chosen according to the value of their performance function (Storn & Price, 1997). What characterises DE is the use of test vectors, which compete with individuals in the current population in order to survive.

Additionally, the importance of the variables was studied. As categorical variables are present, the chosen method depends on removing a variable, evaluating the new model performance and comparing it with the performance of the full model. The greater the decrease in the goodness-of-fit parameter, the greater the importance of the removed independent variable.

3. Results and discussion

As stated earlier, the outlet dissolved oxygen was used as output dependent variable of the proposed GPR–based model. The prediction performed from the independent variables (Tien, 2012) was satisfactory.

Table 2 shows the optimal hyperparameters of the best fitted GPR–based model found with the DE technique. The objective function value, in this case the marginal likelihood was optimised to a value of 239 using the DE technique using the training set.

Table 2 - Optimal hyperparameters of the best fitted GPR–based model found with the DE technique: variance σ_f^2 and lengthscale ℓ for the RBF kernel, the Gaussian noise variance σ_n^2 for the optimised models for the training set.

Taking into account the results achieved, the GPR technique in combination with the DE meta-heuristic optimisation method was able to build models with a high performance for estimating the outlet dissolved oxygen in micro-irrigation sand filters fed with effluents using the test set. Indeed, the coefficient of determination (R^2) of the fitted GPR model was of 0.9023 with a correlation coefficient of 0.9499 for the outlet dissolved oxygen.

A graphical representation of the terms that formed the best fitted GPR–based model for the outlet dissolved oxygen (*DO_o)* is shown below in Figs. 3 and 4. The first order terms, that is, the variations of the dependent variable when all the variables but one are constant (its median value) is shown in Fig 3. The graphs suggest that the variable DO_i is the main influence for the variations in DO_i, while other variables as pH and temperature do not significantly affect this variable as these curves are almost constant. The same effect can be shown in the surfaces that represent the second order relationships, that is, leaving all the independent variables constant but two. Again, it can be seen that the main influence in rapid change of output variable was due to the DO_i.

Fig. 3 - First-order terms for some of the independent variables for the dependent variable output dissolved oxygen (DO_o).

Fig. 4 - Second-order terms of some of the independent variables for the dependent variable output dissolved oxygen (DO_o).

The significance rankings for the input variables predicting the outlet dissolved oxygen (output variable) in this complex nonlinear study are shown in Table 3 and Fig. 5. As there are some categorical variables such as the filter type involved, the method where discarding one independent variable from the model at a time and taking into account the decrease in goodness-of-fit, in this case, the marginal likelihoods, is shown in Table 3. The result is, that for the GPR model, the most significant variable in DO_o prediction is the DO_i, followed by (in order) the type of filter, water temperature, height of the filter bed, pH, velocity, turbidity, and electrical conductivity.

Table 3 - Log marginal likelihood variation value between the full model and the model without the variable for the outlet dissolved oxygen (DO_o) model.

Fig. 5 - Relative relevance of the variables in the GPR model for the outlet dissolved oxygen (DO_o).

As it could be anticipated, DO_o was highly dependent on DO_i since organic pollutants are retained across filter media and chlorination of filter backwashing water reduced microorganisms level, and therefore less oxygen is consumed and dissolved oxygen could increase. However, DO removal depended also on media particle size (Elbana, Ramírez de Cartagena, & Puig-Bargués, 2012) and on the interaction between filter type and filtration velocity, considering input inlet DO as a co-variable (Solé–Torres, Puig–Bargués, Duran–Ros, Arbat, Pujol, & Ramírez de Cartagena, 2019b). The filter type had also a contribution on the results since different underdrain designs affect backwashing performance and frequency (Burt, 2010), which is directly related to DO removal (Enciso-Medina, Multer, & Lamm, 2011; Elbana, Ramírez de Cartagena, & Puig-Bargués, 2012). The third parameter is temperature, but this is also logical since DO values are temperature dependent.

The importance of DO_i for estimating DO_o has been previously observed by Martí et al. (2013) and García–Nieto et al. (2016), working with different types of models. Martí et al. (2013) observed that pH, EC and pressure loss, but not temperature, García–Nieto et al. (2016) found that inlet turbidity and pressure loss were also considered as influential parameters for predicting DO_o. Thus, the results highlight the importance of correctly assessing the performance of each prediction model.

In conclusion, this research was able to estimate the outlet dissolved oxygen (output variable) in agreement with the actual experimental values observed using the GPR–based model with accuracy as well as success. Indeed, Fig. 6 shows the comparison among the DO_o values observed and those predicted by using the GPR model with the testing set. The values predicted by the model using the samples of the testing dataset show a very good agreement with the observed values. As it can be seen, predicted values are very close to the observed values or within the 95% confidence interval. This is to be expected since the coefficient of determination was equal to 0.90. Therefore, in order to achieve the best effective approach in this regression problem it is mandatory the use of a GPR model with a DE optimisation technique.

Fig. 6 - Observed and predicted DO_o values, taking into account the confidence interval, by using the GPR–based model with the testing set ($R^2 = 0.9023$).

4. Conclusions

Taking into account the experimental observations and numerical predictions, the main findings of this study can be summarised as follows:

- Firstly, the development of novel data-driven diagnostic techniques is very useful to predict the DO_o from the experimental measurements. In this sense, the new GPR–based method used here is useful to evaluate the outlet dissolved oxygen in sand media filters used in microirrigation systems.

- Secondly, the assumption that the outlet dissolved oxygen diagnosis can be accurately modelled by using a hybrid GPR–based model in granular filters was confirmed.
- Thirdly, a reasonable coefficient of determination (0.9023) was obtained when this GPR–based model was applied to the experimental dataset corresponding to the DO_o.
- Fourthly, the significance order of the input variables involved in the prediction of the outlet dissolved oxygen in sand media filters was set. This is one of the main findings in this work. Specifically, input variable dissolved oxygen (DO_i) could be considered the most influential parameter in the prediction of the DO_o. In this regard, it is also important to highlight the influential role of the type of filter in the dependent variable outlet dissolved oxygen.
- Finally, the influence of the hyperparameters setting of the GPR approach on the DO_o regression performance was set up.

In summary, this methodology could be applied to other filtration processes with similar or distinct filter media types with success, but it is always necessary to take into account the characteristics of each filter and experiment. Consequently, an effective GPR–based model is a good practical solution to the problem of the determining DO_o in the sand media filters usually used in microirrigation systems.

Acknowledgements

Authors wish to acknowledge the computational support provided by the Department of Mathematics at University of Oviedo as well as financial support of the Spanish Research Agency through grants AGL2015-63750-R and RTI2018-094798-B-100. Additionally, we would like to thank Anthony Ashworth for his revision of English grammar and spelling of the manuscript.

References

Aggarwal, C.C. (2015). *Data mining: the textbook*. New York, USA: Springer.

Ait-Mouheb, N., Bahri, A., Ben Thayer, B., Benyahia, B., Bourrié, G., Cherki, B. et al. (2018). The reuse of reclaimed water for irrigation around the Mediterranean Rim: a step towards a more virtuous cycle? *Regional Environmental Change*, 18, 693–705.

Bhattarai, S.P., Midmore, D.J. & Pendergast, L. (2008). Yield, water-use efficiencies and root distribution of soybean, chickpea and pumpkin under different subsurface drip irrigation depths and oxygenation treatments in vertisols. *Irrigation Science*, 26(5), 439–450.

Bové, J., Arbat, G., Duran–Ros, M., Pujol, T., Velayos, J., Ramírez de Cartagena, F., & Puig–Bargués, J. (2015). Pressure drop across sand and recycled glass media used in micro irrigation filters. *Biosystems Engineering*, 137, 55–63.

Bové, J., Puig–Bargués, J., Arbat, G., Duran–Ros, M., Pujol, T., Pujol, J., & Ramírez de Cartagena, F. (2017). Development of a new underdrain for improving the efficiency of microirrigation sand media filters. *Agricultural Water Management*, 179, 296–305.

Camps–Valls, G., Verrelst, J., Munoz–Mari, J., Laparra, V., Mateo–Jimenez, F., & Gomez–Dans, J. (2016). A survey on Gaussian processes for earth-observation data analysis: a comprehensive investigation. *IEEE Geoscience and Remote Sensing Magazine*, 4(2), 58–78.

Canales-Ide, F., Zubelzu, S., & Rodríguez-Sinobas, L. (2019). Irrigation systems in smart cities coping with water scarcity: The case of Valdebebas, Madrid (Spain). Journal of Environmental Management, 247, 187-195.

Chakraborty, U.K. (2008). *Advances in differential evolution*. Berlin: Springer.

Chen, Y., Lu, D., Luo, L., Pokhrel, Y., Deb, K., Huang, J., & Ran, Y. (2018). Detecting irrigation extent, frequency, and timing in a heterogeneous arid agricultural region using MODIS time series, Landsat imagery, and ancillary data. *Remote Sensing of Environment*, 204, 197–211.

Ebden, M. (2015). Gaussian processes: a quick introduction. https://arxiv.org/pdf/1505.02965.pdf.

Elbana, M., Ramírez de Cartagena, F., & Puig-Bargués, J. (2012). Effectiveness of sand media filters for removing turbidity and recovering dissolved oxygen from a reclaimed effluent used for micro-irrigation. *Agricultural Water Management*, 111, 27–33.

Enciso-Medina, J., Multer, W.L. & Lamm, F.R. (2011). Management, maintenance, and water quality effects on the long-term performance of subsurface drip irrigation systems. *Applied Engineering in Agriculture*, 27 (6), 969–978.

Feoktistov, V. (2006). *Differential evolution: in search of solutions*. New York: Springer.

Freedman, D., Pisani, R., & Purves, R. (2007). *Statistics*. New York: W.W. Norton & Company, New York.

García–Nieto, P.J., García–Gonzalo, E., Arbat, G., Duran–Ros, M., Ramírez de Cartagena, F., & Puig–Bargués, J. (2016). A new predictive model for the filtered volume and outlet parameters in micro-irrigation sand filters fed with effluents using the hybrid PSO–SVM–based approach. *Computers and Electronics in Agriculture*, 125, 74–80.

García–Nieto, P.J., García–Gonzalo, E., Arbat, G., Duran–Ros, M., Ramírez de Cartagena, F., & Puig–Bargués, J. (2018). Pressure drop modelling in sand filters in micro-irrigation using gradient boosted regression trees. *Biosystems Engineering*, 171, 41–51.

García–Nieto, P.J., García–Gonzalo, E., Bové, J., Arbat, G., Duran–Ros, M., & Puig–Bargués, J. (2017). Modeling pressure drop produced by different filtering media in microirrigation sand filters using the hybrid ABC–MARS–based approach, MLP neural network and M5 model tree. *Computers and Electronics in Agriculture,* 139, 65–74.

GPy, 2014. A Gaussian process framework in python. http://github.com/SheffieldML/GPy.

Ko, J., Klein, D.J., Fox, D., & Haehnelt, D. (2007a). GP-UKF: Unscented Kalman filters with Gaussian process prediction and observation models. In 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems (pp. 1901–1907). San Diego, CA, USA: IEEE.

Ko, J., Klein, D.J., Fox, D., & Haehnelt, D. (2007b). Gaussian processes and reinforcement learning for identification and control of an autonomous blimp. In

Proceedings 2007 IEEE International Conference on Robotics and Automation (pp. 742–747). Roma, Italy: IEEE.

Kong, D., Chen, Y., & Li, N. (2018). Gaussian process regression for tool wear prediction. *Mechanical Systems and Signal Processing*, 104, 556–574.

Kuhn, M., & Johnson, K. (2018). *Applied predictive modeling*. New York, USA: Springer.

Larose, D.T. (2015). Data mining and predictive analytics. New York, USA: Wiley.

Liu, H., Yang, C., Huang, M., Wang, D., & Yoo, C. (2018). Modeling of subway indoor air quality using Gaussian process regression. *Journal of Hazardous Materials*, 359, 266–273.

Ma, X., Xu, F., & Chen, B. (2019). Interpolation of wind pressures using Gaussian process regression. *Journal of Wind Engineering & Industrial Aerodynamics*, 188, 30–42.

Maestre-Valero, J.F., & Martínez-Álvarez, V. (2010). Effects of drip irrigation systems on the recovery of dissolved oxygen from hypoxic water. *Agricultural Water Management*, 97, 1806–1812.

Marsland, S. (2014). *Machine learning: an algorithmic perspective*. Boca Raton, FL, USA: Chapman and Hall/CRC Press.

Martí, P., Shiri, J., Duran–Ros, M., Arbat, G., Ramírez de Cartagena, F., & Puig–Bargués, J. (2013). Artificial neural networks vs. Gene Expression Programming for estimating outlet dissolved oxygen in micro-irrigation sand filters fed with effluents. *Computers and Electronics in Agriculture*, 99, 176–185.

Martin, O. (2018). *Bayesian analysis with python*. Birmingham, UK: Packt Publishing.

Murphy, K.P. (2012). *Machine learning: a probabilistic perspective*. Cambridge, MA, USA: The MIT Press.

Nakayama, F.S., Boman, B.J., & Pitts, D.J. (2007). Maintenance. In: Lamm, F.R., Ayars, J.E. & Nakayama, F.S. (Eds.), Microirrigation for Crop Production. Design, Operation, and Management (pp. 389–430). Amsterdam, Netherlands: Elsevier.

Noori, M., Hassani, H., Javaherian, A., Amindavar, H., & Torabi, S. (2019). Automatic fault detection in seismic data using Gaussian process regression. *Journal of Applied Geophysics*, 163, 117–131.

Paananen, T., Piironen, J., Andersen, M.R., & Vehtari, A. (2019). Variable selection for Gaussian processes via sensitivity analysis of the posterior predictive distribution. In Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS), Proceedings of Machine Learning Research (PMLR) (pp. 1743–1752). Naha, Okinawa, Japan: arXiv:1712.08048 [stat.ME], Cornell University, USA.

Picard, R., & Cook, D. (1984). Cross-validation of regression models. *Journal of the American Statistical Association*, 79(387), 575–583.

Piironen, J., & Vehtari, A. (2016). Projection predictive model selection for Gaussian processes. In 2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP) (pp. 1–6). Vietri sul Mare, Italy: IEEE.

Price, K., Storn, R.M., & Lampinen, J.A. (2005). *Differential evolution: A practical approach to global optimization*. Berlin: Springer.

Puig–Bargués, J., Duran–Ros, M., Arbat, G., Barragán, J., & Ramírez de Cartagena, F. (2012). Prediction by neural networks of filtered volume and outlet parameters in micro-irrigation sand filters using effluents. *Biosystems Engineering*, 111(1), 126–132.

Rasmussen, C.E. (2003). *Gaussian processes in machine learning: summer school on machine learning*. Berlin, Germany: Springer.

Rasmussen, C.E., & Williams, C.K.I. (2006). *Gaussian processes for machine learning*. Cambridge, MA, USA: The MIT Press.

Rogers, S., & Girolami, M. (2016). *A first course in machine learning*. Boca Raton, FL, USA: Chapman and Hall/CRC.

Schneider, M., & Ertel, W. (2010). Robot learning by demonstration with local Gaussian process regression. In: The 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems (pp. 255–260). Taipei, Taiwan: IEEE.

Schulz, E., Speekenbrink, M., & Krause, A. (2018). A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions. *Journal of Mathematical Psychology*, 85, 1–16.

Seeger, M. (2000). Bayesian model selection for support vector machines, Gaussian processes and other kernel classifiers. In NIPS'99 Proceedings of the 12th International Conference on Neural Information Processing Systems (vol. 12, pp. 603–609). Cambridge, MA, USA: The MIT Press.

Shi, J.Q., & Choi, T. (2011). *Gaussian process regression analysis for functional data*. Boca Raton, FL, USA: Chapman and Hall/CRC Press.

Simon, D. (2013). *Evolutionary optimization algorithms*. New York: Wiley.

Solé–Torres, C., Duran–Ros, M., Arbat, G., Pujol, J., Ramírez de Cartagena F., & Puig–Bargués, J. (2019a). Assessment of field water uniformity distribution in a microirrigation system using a SCADA system. *Water*, 11(7), 1346–1359.

Solé–Torres, C., Puig–Bargués, J., Duran–Ros, M., Arbat, G., Pujol, J., & Ramírez de Cartagena, F. (2019b). Effect of underdrain design, media height and filtration velocity on the performance of microirrigation sand filters using reclaimed effluents. *Biosystems Engineering*, 187, 292–304.

Storn, R., & Price, K. (1997). Differential evolution - a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, 11, 341–359.

Tan, P.–N., Steinbach, M., Karpatne, A., Kumar, V. (2018). *Introduction to data mining*. Oxford, UK: Pearson.

Tien, C. (2012). *Principles of filtration*. Kidlington, Oxford, UK: Elsevier.

Trooien, T.P., & Hills, D.J. (2007). Application of biological effluent. In Lamm, F.R., Ayars, J.E., & Nakayama, F.S. (Eds.), Microirrigation for Crop Production. Design, Operation and Management (pp. 329–356). Amsterdam: Elsevier.

Vidales, A. (2019). *Machine learning with MATLAB: Gaussian process regression, analysis of variance and Bayesian optimization*. Independently published.

Witten, I.H., Frank, E., Hall, M.A., Pal, C.J. (2016). *Data mining: practical machine learning tools and techniques*. Cambridge, MA, USA: Morgan Kaufmann.

Wu, R., & Wang, B. (2018). Gaussian process regression method for forecasting of mortality rates. *Neurocomputing*, 316, 232–239.

Zhou, Y., Zhou, B., Xu, F., Muhammad, T., Li, Y. (2019). Appropriate dissolved oxygen concentration and application stage of micro-nano bubble water oxygenation in greenhouse crop plantation. *Agricultural Water Management*, 223, 105713.

Figure

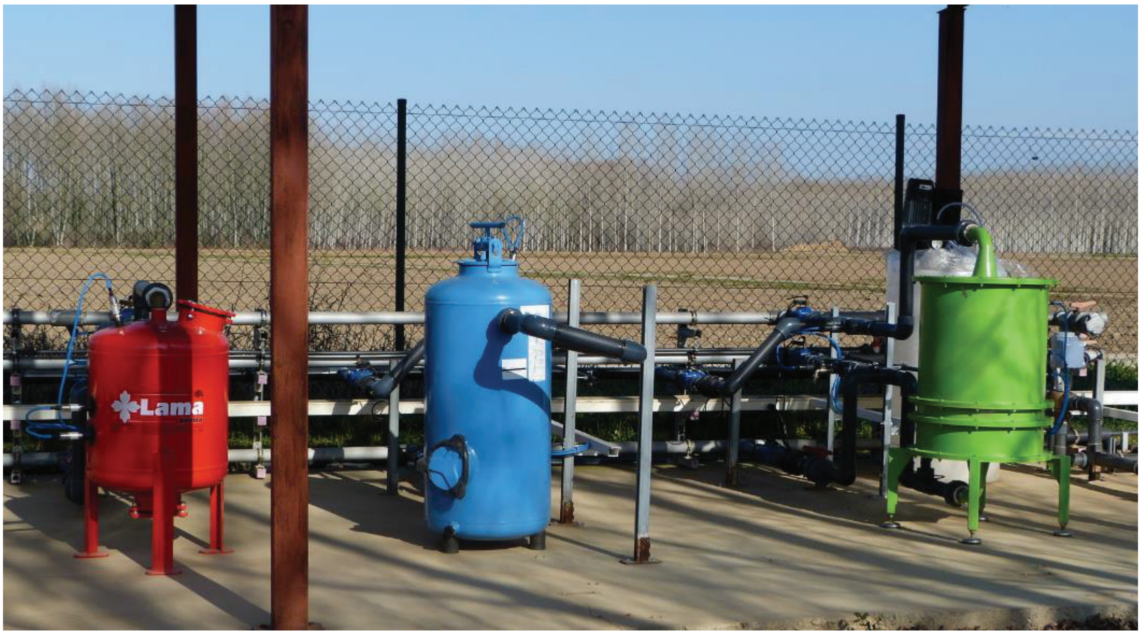

Fig. 1 - Picture of the experimental set-up with the three filter designs: (a) red: arm collector; (b) blue: inserted domes; and (c) green: a porous media prototype.

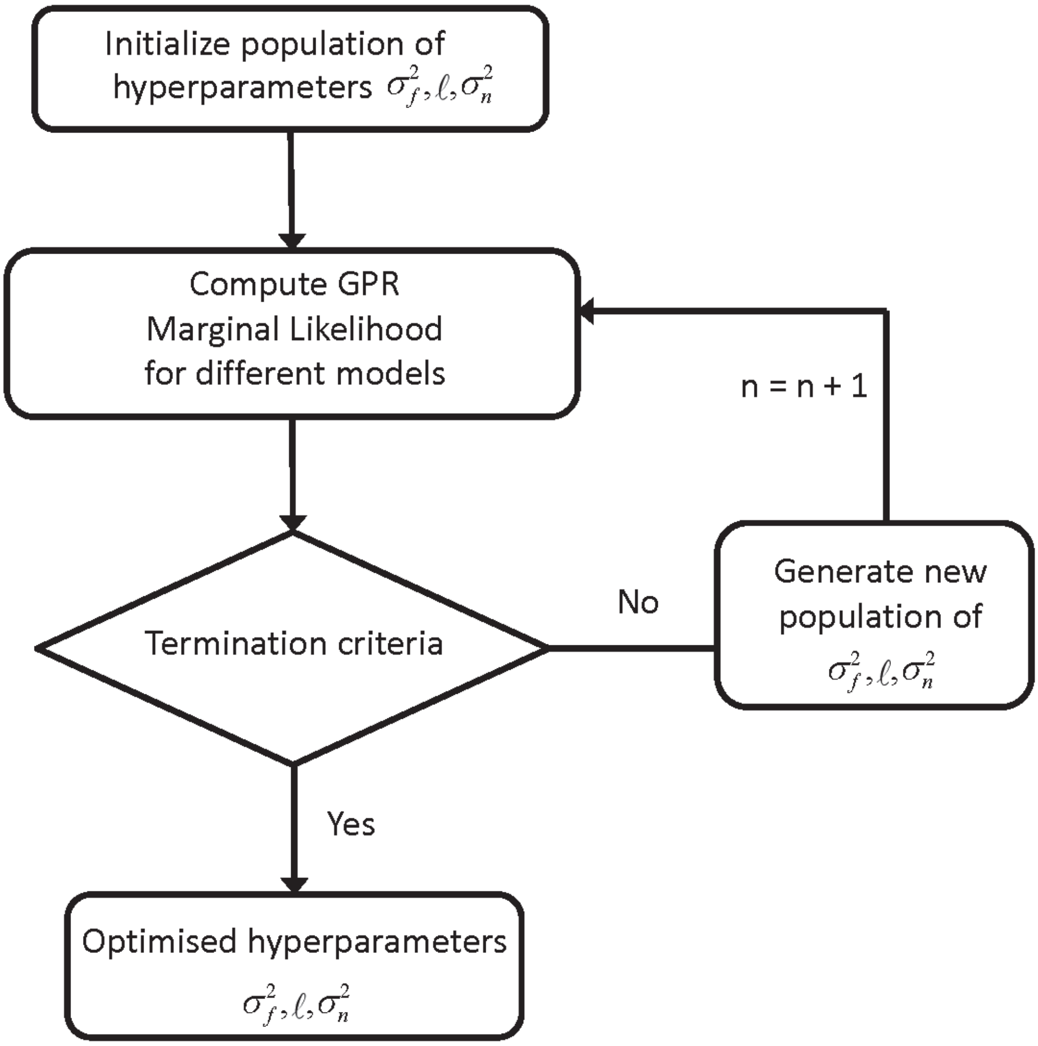
Initialize population of hyperparameters $\sigma_f^2, \ell, \sigma_n^2$
Compute GPR Marginal Likelihood for different models
n = n + 1
Termination criteria
No
Generate new population of $\sigma_f^2, \ell, \sigma_n^2$
Yes
Optimised hyperparameters $\sigma_f^2, \ell, \sigma_n^2$

Fig. 2 – GPR Model selection using the DE optimisation technique.

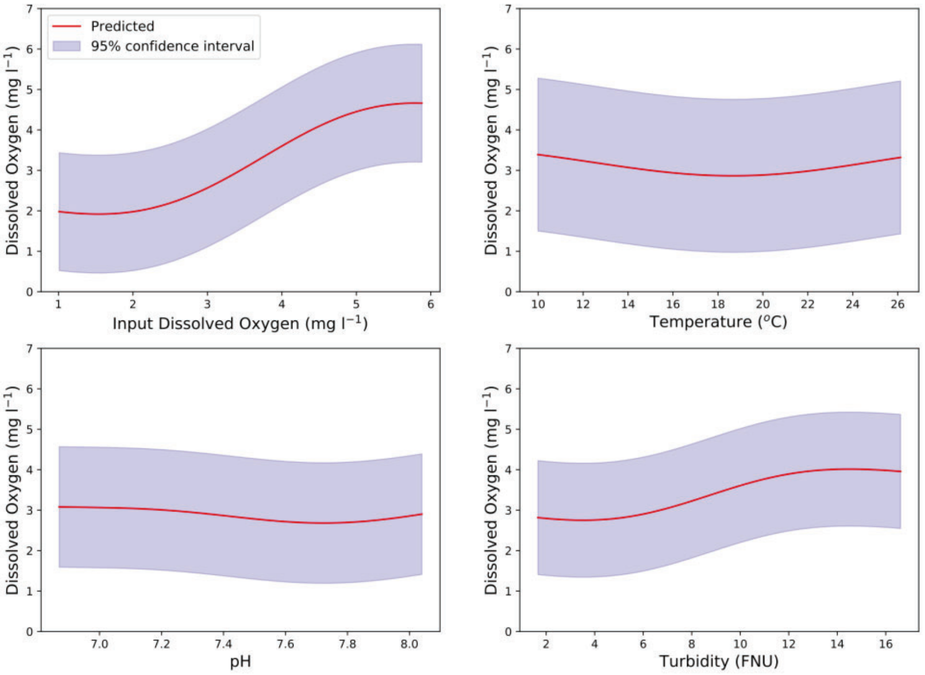

Fig. 3 - First-order terms for some of the independent variables for the dependent variable output dissolved oxygen (*DO_o)*.

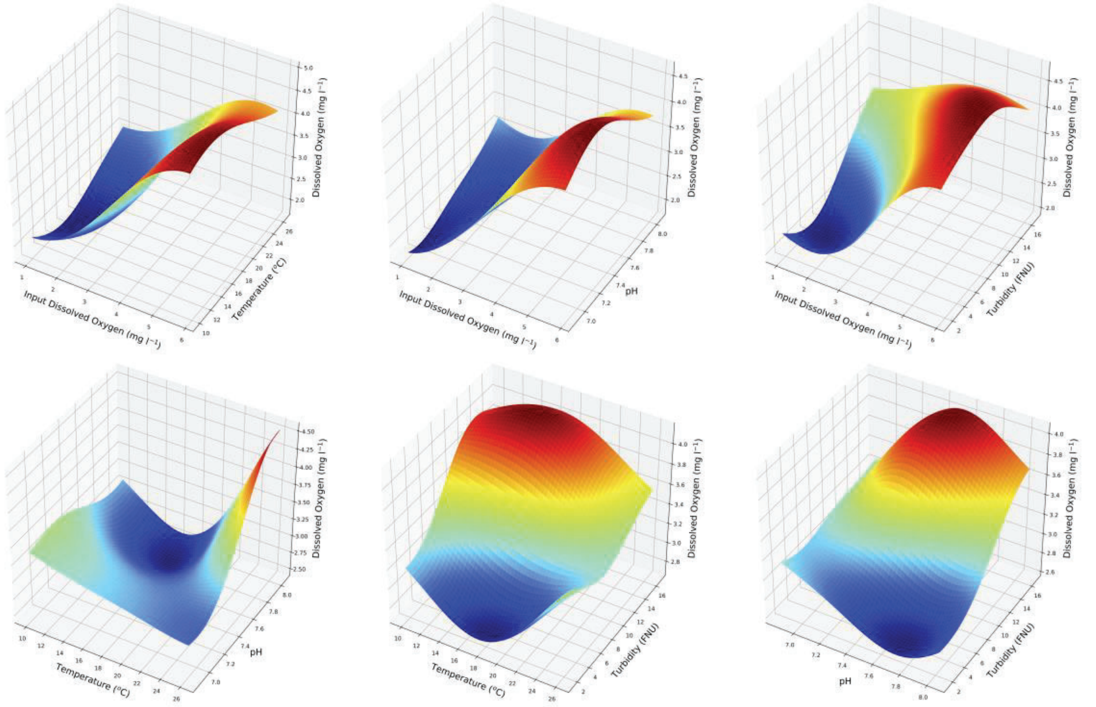

Fig. 4 - Second-order terms of some of the independent variables for the dependent variable output dissolved oxygen (*DO$_o$)*.

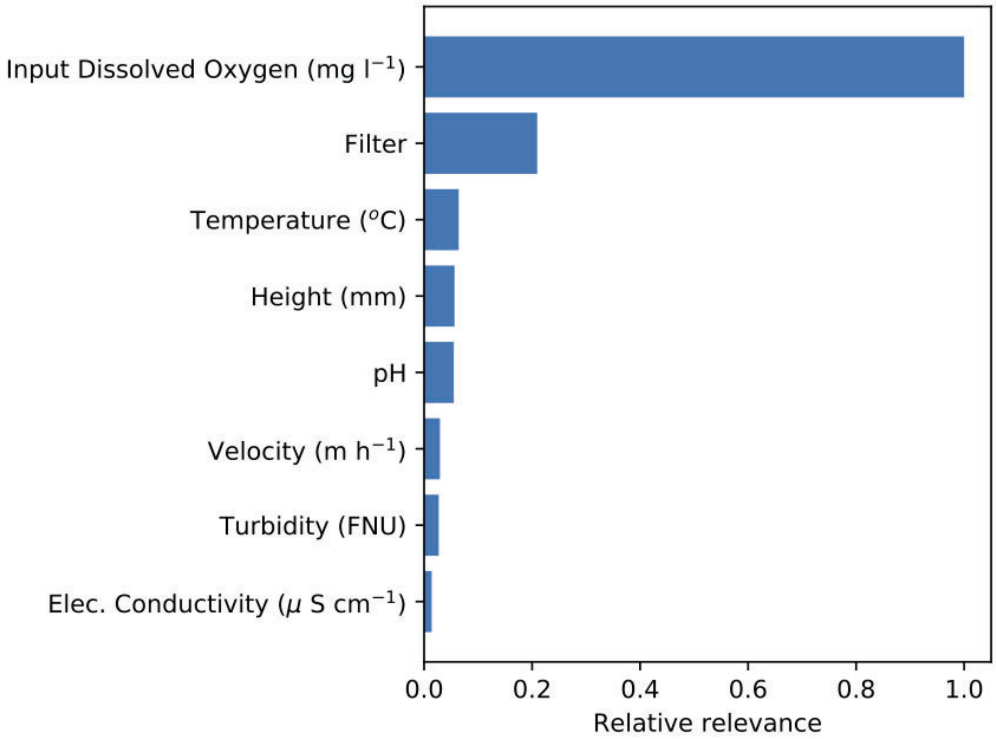

Fig. 5 - Relative relevance of the variables in the GPR model for the outlet dissolved oxygen (DO_o).

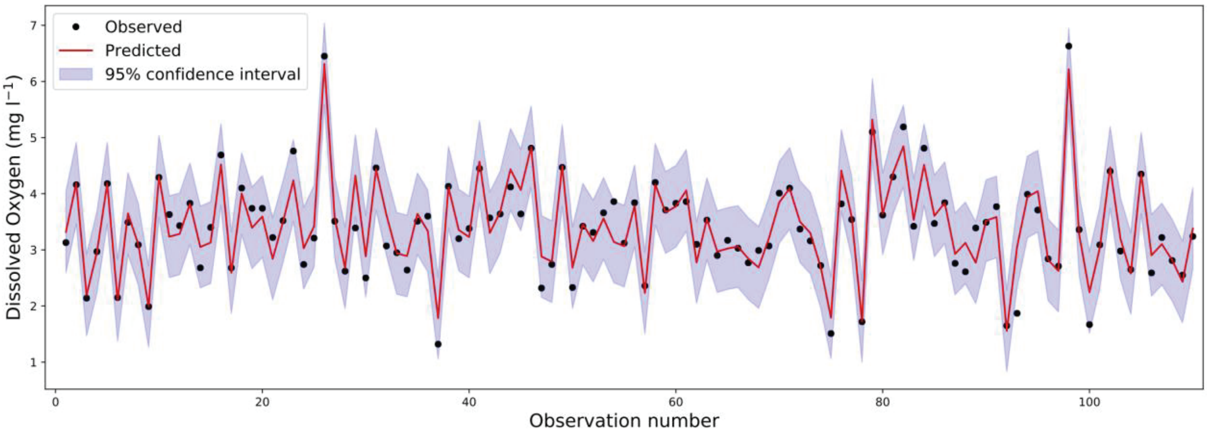

Fig. 6 - Observed and predicted DO_o values, taking into account the confidence interval, by using the GPR–based model with the testing set ($R^2 = 0.9023$).

Table 1 - Set of operation physical input variables used in this study and their names along with their means and standard deviations.

Input variables	Name of the variable	Mean	Standard deviation
Filter media type	Filter	--	--
Height of the filter bed (mm)	H	256.31	49.601
Filtration velocity (m h^{-1})	v	49.909	14.174
Electrical conductivity (μ S cm^{-1})	CE_i	2575.6	497.68
Input dissolved oxygen (mg l^{-1})	DO_i	3.3529	0.9860
pH	pH_i	7.3526	0.2229
Input turbidity (FNU)	$Turb_i$	6.1029	2.5898
Water temperature (ºC)	T_i	20.002	3.3486

Table 2 - Optimal hyperparameters of the best fitted GPR–based model found with the DE technique: variance σ_f^2 and length-scale ℓ for the RBF kernel, the Gaussian noise variance σ_n^2, and the corresponding objective function value for the optimized models for the training set.

Output variable	σ_f^2	ℓ	σ_n^2	Objective function value
DO_o	1.57	1.97	0.0636	239

Table 3 - Log marginal likelihood variation value between the full model and the model without the variable for the DO_o model.

Variable	Likelihood variation
Input dissolved Oxygen (mg l^{-1})	589.62
Filter	123.51
Water temperature (ºC)	37.77
Height (mm)	332.1
pH	32.45
Velocity (m h^{-1})	17.31
Input turbidity (FNU)	15.96
Electrical Conductivity (µS cm^{-1})	8.25