A new predictive model for the outlet turbidity in micro-irrigation sand filters fed with effluents using Gaussian process regression

P.J. García Nieto[a,*], E. García-Gonzalo[a], J. Puig-Bargués[b], C. Soler-Torres[b], M. Duran-Ros[b], G. 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 used in microirrigation systems must remove suspended particle load for avoiding emitter physical clogging. Turbidity is a parameter related to suspended particle load that it is easy and quick to measure and it is also included in some guidelines for reusing effluents in irrigation. Currently, there are not sufficiently accurate models available to predict outlet turbidity for sand filters, which would be useful for both irrigators and engineers. The aim of this study was to obtain a predictive model able to perform an early detection of the sand filter outlet value of turbidity. This study presents a powerful and effective Bayesian nonparametric approach, termed Gaussian process regression (GPR) model, for predicting the output turbidity ($Turb_o$) from data corresponding to 637 samples of different sand filters using reclaimed effluent. This optimization technique involves kernel parameter setting in the GPR training procedure, which significantly influences the regression accuracy. To this end, the most important parameters of this process are monitored and analyzed: type of filter,

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

height of the filter bed (H), filtration velocity (v) and filter inlet values of the electrical conductivity (CE_i), dissolved oxygen (DO_i), pH_i, turbidity ($Turb_i$) and water temperature (T_i). The results of the present study are two-fold. In the first place, the significance of each variable on the filtration is presented through the model. Secondly, a model for forecasting the outlet turbidity was obtained with success. Indeed, regression with optimal hyperparameters was performed and a coefficient of determination equal to 0.8921 for outlet turbidity was obtained when this new predictive GPR–based model was applied to the experimental dataset. The agreement between experimental data and the model confirmed the good performance of the latter.

Keywords: Gaussian process regression (GPR); Bayesian statistics; Machine learning techniques; Drip irrigation; Clogging

1. Introduction

Shortage of fresh water resources has stimulated the use of reclaimed effluents with microirrigation systems since these systems offer several agronomic, environmental and health advantages regarding other irrigation methods (Trooien and Hills, 2007; Tal, 2016). However, the use of effluents pose an increased emitter clogging risk due to their higher salt, nutrients, solid and biological concentrations. Thus, the greatest challenge when using effluents is preventing emitter clogging to keep microirrigation systems operating as designed (Trooien and Hills, 2007). Despite a proper selection of emitter reduces emitter clogging (Zhou et al., 2019), operation and maintenance practices such as filtration, water treatment, dripline flushing and monitoring system performance are required when effluents are used (Trooien and Hills, 2007).

Sand media filters are considered the standard for protection of microirrigation systems (Trooien and Hills, 2017) since they usually remove more particles and therefore reduce emitter clogging (Ravina et al., 1997; Capra and Scicolone, 2007; Duran-Ros et al., 2009; Tripathi et al., 2014; Wen-Yong et al., 2015). However, investment and maintenance costs for sand filters are greater (Pujol et al., 2011) and require high technological and professional standards (Capra and Scicolone, 2007), which is aligned with the growth of precision microirrigation (Madramootoo and Morrison, 2013). In this regard, advanced techniques such as neural networks (ANN), gene expression programming (GEP) (Martí et al., 2013), support vector machines (SVM) (García-Nieto et al., 2016) have been used for predicting the filtered volume and the value of dissolved oxygen – an indicator of the water quality – at sand media filter outlets. More recently, García-Nieto et al. (2017, 2018) used hybrid algorithms and gradient boosted regression trees for modeling pressure loss in these filters. However, prediction of turbidity values at microirrigation sand filter outlet has not been completely successful (Puig-Bargués et al., 2012) although better results have been obtained in a pilot multi-media filter (Hawari and Alnahhal, 2016). Turbidity is a parameter related to suspended load (Stevenson and Bravo, 2019) that it is easy and quick to measure using specific sensors. Accurate prediction of turbidity is becoming interesting since several guidelines for using reclaimed effluents in irrigation (e.g. USEPA, 2012; Alcalde-Sanz and Gawlik, 2017) include thresholds values for this parameter.

Thus, the application of the innovative methodology that combines the Gaussian process regression (GPR) approach (Rasmussen, 2003; Kuhn and Johnson, 2018; Ebden, 2015) with the optimization algorithm Limited-memory Broyden-Fletcher-

Goldfarb-Shanno (LBFGSB) (Liu and Nocedal, 1989; Byrd et al., 1994; Zhu et al., 1997) to foretell the outlet turbidity in sand media filters used in microirrigation systems could be an interesting approach since, at the knowledge of the authors, has not been yet addressed in previous investigations. GPR is a machine learning method developed on the basis of statistical theory and Bayesian theory. It is a nonparametric regression method and can be considered a complex model with capability to model nonlinearities and variable interactions (Rasmussen, 2003; Ebden, 2015). When this method is compared with other machine learning techniques (Hastie et al., 2003; Mather and Johnson, 2015), GPR has several advantages (Rasmussen and Williams, 2006): (1) GPR has an important generalization capacity; (2) the hyperparameters in GPR can be self-adaptively calculated; and (3) the GPR outputs have clear probabilistic meaning. In this study, the LBFGSB method was applied successfully to optimize the GPR hyperparameters. Previous researches show that GPR is an effective tool in many fields, such as irrigation mapping (Chen et al., 2018), wind engineering and industrial aerodynamics (Ma et al., 2019), applied geophysics (Noori et al., 2019), applied demography (Wu and Wang, 2018), psychology (Schulz et al., 2018), mechanical engineering (Kong et al., 2018), environmental engineering (Liu et al., 2018), tracking and positioning (Ko et al., 2007a), deformation observation (Rogers and Girolami, 2016), system identification and control (Ko et al., 2007b) and so on. However, it has never been used in microirrigation sand filters.

The main objective of the present study was to predict the outlet turbidity ($Turb_o$) in sand media filters that worked with reclaimed effluents using Gaussian Processes (GPs) in combination with the LBFGSB parameter optimization technique.

The structure of this paper is organized 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

A filtration platform with three sand media filters fed with the reclaimed effluent of the wastewater treatment plant of Celrà (Girona, Spain) was used for carrying out the experiment. Each one of the filters had a different underdrain design: inserted domes (model FA-F2-188, Regaber, Parets del Vallès, Spain), arm collector (model FA1M, Lama, Sevilla, Spain) and porous media (prototype designed by Bové et al. (2017)) (see Fig. 1).

All the filters were filled with silica sand CA-07MS (Sibelco Minerales SA, Bilbao, Spain) with the same characteristics: 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. Two media heights were tested for each filter: 20 and 30 cm, respectively.

Each filter operated on a 8 h daily basis and not simustaneously with the other two. Sligth changes on the operation time were sporadically set for solving different operation and maintenance issues. Two filtration velocities were used for each filter: 30 and 60 m/h, respectively. 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 during 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, turbidity and dissolved oxygen) and after (only turbidity and dissolved oxygen) being filtered were measured and recorded every minute in a supervisory control and data acquisition system (SCADA) fully described by Solé-Torres et al. (2019).

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 turbidity as a function of different experimentally measured parameters that the GPR–based model needs as input. The output variable is the outlet turbidity which is an indicator of the quality of the filtered effluent and it is directly related to physical clogging risk of emitters of microirrigation systems. The operation input variables are as follows:

- Filter: Each one of the three filter designs (porous, dome and arm collector) described in section 2.1. It is a categorical variable;
- Height of the filter bed (cm): this is an operation variable for sand filters. Two different filter bed heights of 20 and 30 cm were tested for each filter;
- Filtration velocity (m/h): it is a variable related to filter operation. Two filtration velocities (30 and 60 m/h) were tested for each filter since these follow within the common range of velocities suggested by the manufacturers;
- Electrical conductivity (μS/cm): it is a general measure of water quality related to salinity, which is a constraint for using microirrigation (Tal, 2016);
- Dissolved oxygen (mg/l): it is a variable related to the ability of water to support aquatic life. This is a common parameter used for controlling biological treatment in wastewater plants;
- pH: it measures water acidity or alkalinity;
- Water temperature (ºC): temperature of the effluent at the filter inlet;
- Input turbidity (FNU): this a key parameter for water quality that measures water clarity, which depends on suspended solid load;
- Filtered volume (m^3): it measures the volume of effluent filtered in each filtration cycle.

2.3. Gaussian process regression (GPR)

GPs 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 latent functions, which are unobserved, and the likelihood (of the latent function) and observations shape this prior to produce posterior probabilistic estimates. Consequently, the joint distribution of training and test data is a multidimensional GP, and the predicted distribution is estimated by conditioning on the training data (Camps-Valls, 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 fully stated exactly through the mean and covariance: $x : N\left(\mu, \sigma^2\right)$. On the other hand, a Gaussian process can be seen as a generalization of the Gaussian probability distribution and applies over functions. From the functional space point of view, a Gaussian procedure 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

Suppose that $D=\left\{\left(\mathbf{x}_i, y_i\right) / i=1,2,\ldots,N\right\}$ depicts the training dataset of the Gaussian approach. Moreover, the feature vectors $\mathbf{x}_i \in \Re^n$ comprise the extracted features or the merged features and the pertinent segregation parameters. The observed target values y_i reproduce the outlet turbidity (Turb$_\text{o}$) measured in a filtration process, respectively. $X=\left\{\mathbf{x}_i\right\}_{i=1}^N$ depicts the input matrix of training dataset, $\mathbf{y}=\left\{y_i\right\}_{i=1}^N$ symbolizes the output vector. A Gaussian process $f\left(\mathbf{x}\right)$ defines a prior over functions, which can be converted into a posterior over functions once we have seen some data. A Gaussian process 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 and Williams, 2006; Marsland, 2014):

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

so that

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

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 it is also reasonable if there is no prior knowledge about the mean variable, as is the case in this study.

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

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

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 Gaussian process. 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 given by:

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

so that ε is the additive white noise and besides suppose that Gaussian noise is independent and identically distributed such that $\varepsilon : N\left(0, \sigma_n^2\right)$, 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 (Vidales, 2019):

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

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 $\overline{f}^*$ 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 Gaussian process, the observed values and the function values at new test points obey a joint Gaussian previous distribution which can be expressed as:

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

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 and ask how likely predictions for the $\mathbf{f}^*$ are. This can be expressed as:

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

where

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

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

Afterwards, 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

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 maximizes the marginal likelihood. That is, the optimal hyperparameters are achieved by maximizing the log marginal likelihood.

The marginal likelihood $P\left(\mathbf{y}|X\right)$ is obtained, using Bayes' rule, as:

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

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

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

where $K_y = K + \sigma_n^2 I, K = K(X, X)$ and $| \, |$ is the determinant. In this expression, the first term is a data-fit term, the second term (always positive), substracted from it, is a model complexity penalty, and the last term is just a normalization constant. Then, this expression shows that the criterion of maximum marginal likelihood avoids the problem of over-fitting because if two models are explaining the observed data, then the simplest one will be chosen (Murphy, 2012).

The optimal hyperparameters $\Theta' = \arg\max_{\Theta} \log p\left(\mathbf{y} \middle| X, \Theta\right)$ can be calculated using any standard gradient-based optimizer after parameter initialization. In this study, the variant of the limited-memory Broyden-Fletcher-Goldfarb-Shanno algorithm, denomined LBFGSB algorithm (Liu and Nocedal, 1989; Byrd et al., 1994; Zhu et al., 1997) is used.

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 is the outlet turbidity. To predict the outlet turbidity from other input operation parameters, it is necessary to choose the model that best fits the experimental data. In this sense, to determine the goodness–of–fit, the criterion considered here was the coefficient of determination R^2 (Picard and Cook, 1984; Freedman et al., 2007). A dataset takes values t_i, each of which has an associated modelled value y_i. The former are termed the observed values and the latter are often referred to as the predicted values. The dataset variability is measured through different sums of squares as follows (Freedman et al., 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;

- $SS_{err} = \sum_{i=1}^{n} (t_i - y_i)^2$: 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:

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

so that a coefficient of determination value of 1.0 points out that the regression curve fits the data perfectly.

Two additional criteria considered in this study were the root mean square error (RMSE) and mean absolute error (MAE) (Hastie et al., 2003; Wasserman, 2003). These statistics are also used frequently to evaluate the forecasting capability of a mathematical model. Indeed, the root mean square error (RMSE) and mean absolute error (MAE) are given by the expressions (Freedman et al., 2007; Wasserman, 2003):

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (t_i - y_i)^2}{n}} \tag{14}$$

$$\text{MAE} = \frac{\sum_{i=1}^{n} |t_i - y_i|}{n} \tag{15}$$

If the root mean square error (RMSE) has a value of zero, it means that there is no difference between the predicted and observed data. Mean Absolute Error (MAE) is the average vertical distance between each point and the identity line. MAE is also the average horizontal distance between each point and the identity line. MAE has a clear interpretation as the average absolute difference between t_i and y_i.

Besides, it is well known that the GPR technique depends strongly on the following hyperparameters (Friedman and Roosen, 1995; Xu et al., 2004; Vidoli, 2011):

- Variance (σ_f^2): is the signal variance and controls the vertical scale of the kernel function;
- Lengthscale (ℓ): the characteristic length-scale and 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 $\varepsilon : N\left(0,\sigma_n^2\right)$, then σ_n^2 is the variance of this noise.

At this point, we have constructed a model (specifically in this study, the novel GPR–based model) taking as dependent variable the outlet turbidity (output variable) from the other eight remaining variables (input variables) in granular filters (Tien, 2012; Bové et al., 2015), studying their effect in order to optimize its calculation through the analysis of the coefficient of determination R^2 with success.

Additionally, as previously mentioned, this GPR technique is greatly dependent on their hyperparameters: variance (σ^2); lengthscale (ℓ) and the Gaussian noise variance (σ_n^2). The traditional way of performing hyperparameter optimization 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 variant of the limited-memory Broyden-Fletcher-Goldfarb-Shanno algorithm, denomined LBFGSB algorithm (Liu and Nocedal, 1989; Byrd et al., 1994; Zhu et al., 1997) is used due to its features of rapid convergence and moderate memory requirement for large-scale problems. Moreover, LBFGSB is an iterative algorithm. After initialization with a starting point and boundary constraints, it iterates through five phases (Fei et al., 2014): (1) gradient projection; (2) generalized Cauchy point calculation; (3) subspace minimization; (4) line searching; and (5) limited-memory Hessian approximation. It is important to observe LBFGSB is an iterative algorithms that requires initialization and is sensitive to the initial value of the hyperparameters.

3. Results and discussion

The new predictive model created, employed as input variables eight different operation variables. All of them are presented in Table 1. The total number of samples measured experimentally was 637, but after removing samples with missing data, we have worked with data from 547 filtration cycles.

Table 1

Set of operation physical input variables used in this study along with their mean, median, standard deviation (STD) and mean absolute deviation (MAD).

In order to tackle this study, we divided the dataset in a training set with 80% of the data, and testing set with the remainder 20% of the data. A model is constructed and optimized with the training data and then, it is tested with the test data set.

The outlet turbidity is used as output dependent variable of the proposed GPR–based model. The prediction performed from the independent variables (Tien, 2012) was satisfactory as it was already stated before, the GPR technique is influenced by the selection of the GPR hyperparameters much as the variance σ^2 and lengthscale ℓ for the RBF kernel, the Gaussian noise variance σ_n^2 and objective function value.

Table 2 points out the optimal hyperparameters of the best fitted GPR–based model found with the LBFGSB optimization technique. Usually, the traditional way of performing hyperparameter optimization in most computational codes 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. Indeed, the grid search is a brute force method and, as such, almost any optimization method improves its efficiency. The LBFGSB method used here belongs to quasi-Newton methods, a class of hill-climbing optimization techniques that seek a stationary point of a function. It is an iterative method for solving nonlinear optimization problems.

Table 2

Optimal hyperparameters of the best fitted GPR–based model found with the LBFGSB technique: variance σ_f^2 and lengthscale 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.

Therefore, we have constructed a new predictive model that is the GPR–based model that employs as dependent variable the outlet turbidity in micro-irrigation sand filters fed with effluents.

The value of R^2 was calculated using the optimized model with the testing set. The module Gpy from the Gaussian process framework in python (Gpy, 2014; Martin, 2018), along with the LBFGSB technique (Liu and Nocedal, 1989; Byrd et al., 1994; Zhu et al., 1997), were used to construct the final regression model.

Taking into account the results achieved, the GPR technique in combination with the LBFGSB optimization method is able to build models with a high performance for the estimation of the outlet turbidity 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 0.8921 with a correlation coefficient of 0.9445, and the root mean square error (RMSE) and mean absolute error (MAE) were 0.4335 and 0.2974 for the outlet turbidity, respectively. A computer with a CPU Intel Core i7-4770 @ 3.40 GHz with eight cores and 15.5 GB RAM memory was used, taking 0.2676 seconds to obtain the final outlet turbidity (Turb$_o$) model.

A graphical representation of the terms that form the best fitted GPR–based model for the outlet turbidity (Turb_o) is shown below in Figs. 2 and 3.

Fig. 2. First-order terms for some of the independent variables for the dependent variable output turbidity (Turb_o).

Fig. 3. Second-order terms of some of the independent variables for the dependent variable output turbidity (Turb_o).

3.1. Importance of the variables

The importance of the variables for Gaussian Process models is often done using automatic relevance determination (ARD) (Seeger, 2000). However, this procedure does not provide an adequate technique because it systematically underestimates the relevance of linear input variables in relation with nonlinear ones that have the same relevance in the generation of the squared error (Piironen and Vehtari, 2016). This is consistent with our experience. For instance, it is to be expected that an important variable for Turb_o is Turb_i. This result is not obtained with ARD, where the importance of this variable is relegated to the last positions of the relevance ranking. As an alternative, Paananen and co-workers (Paananen et al., 2019) propose the use the variance of the posterior latent mean. When the value of a single independent variable is modified a small amount, a large variation of the value of the latent mean implies that this variable is relevant. However, this method is not suitable for categorical variables, as it is the case with the *filter* variable, as they do not admit small modifications: either we have one filter or other. Thus, in our study, a different method that accounts for the presence of categorical variables has been used: the importance of the variables has been studied 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 independent variable.

Therefore, as an additional result of these calculations, the significance rankings for the input variables predicting the outlet turbidity (output variable) in this complex nonlinear study are shown in Table 3 and Fig. 4. Thus, for the GPR model the most significant variable in output turbidity prediction is the input turbidity, followed by the filter, electrical conductivity, height of the filter bed, velocity, dissolved oxygen, water temperature and pH.

Table 3

Log marginal likelihood variation value between the full model and the model without the variable for the Turb_o model.

Fig. 4. Relative relevance of the variables in the GPR model for the outlet turbidity (Turb_o).

As it could be anticipated, outlet turbidity is highly dependent on inlet turbidity since suspended particles are retained across filter media, and therefore turbidity is reduced. Less turbidity at filter outlet is to be expected. However, turbidity removal depends also on media particle size (Triphati et al., 2014) and on the interaction between filter type, media height and filtration velocity, considering input turbidity as a co-variable (Solé-Torres et al., 2019b). The results confirm these previous results, but electrical conductivity has also an effect that was not considered before since only one water quality parameter could be included in the analysis carried out by Solé-Torres et al. (2019b). Electrical conductivity measures total dissolved solids (Trooien and Hills, 2007) and is not directly related with turbidity but with the effluent that was used in the experiment it showed a slight effect on outlet turbidity. Further research considering more filtration velocities and media heights could shed more light on their effect on turbidity values.

In conclusion, this research work was able to estimate the outlet turbidity (output variable) in agreement with the actual experimental values observed using the GPR–based model with great accurateness as well as success. Indeed, Fig. 5 shows the comparison among the outlet turbidity values observed and predicted by using the GPR model with the testing set. Therefore, it is mandatory the use of a GPR model with an LBFGSB optimization technique in order to achieve the best effective approach in this regression problem.

Fig. 5. Observed and predicted Turb_o values, taking into account the confidence interval, by using the GPR–based model with the testing set ($R^2 = 0.8921$).

4. Conclusions

Taking into account the experimental and numerical results, the main findings of this study can be summarized as follows:

- Firstly, there are no analytical equations to predict the outlet turbidity from the experimental values; accordingly, the development of alternative diagnostic techniques is very important. In this sense, the new GPR–based method used in this work is a good decision to evaluate the outlet turbidity in sand media filters used in microirrigation systems;
- Secondly, the assumption that the outlet turbidity diagnosis can be accurately modelled by using a hybrid GPR–based model in granular filters was confirmed;
- Thirdly, a reasonable coefficient of determination equal to 0.8921 was obtained when this GPR–based model was applied to the experimental dataset corresponding to the outlet turbidity ($Turb_o$);
- Fourthly, the significance order of the input variables involved in the prediction of the outlet turbidity in sand media filters was set. This is one of the main findings in this work. Specifically, input variable Turbidity ($Turb_i$) could be considered the most influential parameter in the prediction of the outlet turbidity. In this regard, it is also important to highlight the influential role of the type of filter in the dependent variable outlet turbidity;
- Finally, the influence of the hyperparameters setting of the GPR approach on the outlet turbidity 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 determination of the outlet turbidity in sand media filters broadly 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

Alcalde–Sanz, L., Gawlik, B.M., 2017. Minimum quality requirements for water reuse in agricultural irrigation and aquifer recharge - Towards a water reuse regulatory instrument at EU level. EUR 28962 EN, Publications Office of the European Union, Luxembourg.

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. Biosyst. Eng. 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. Agric. Water Manage. 179, 296–305.

Byrd, R.H., Lu, P., Nocedal, J., Zhu, C., 1994. A limited-memory algorithm for bound constrained optimization. SIAM J. Sci. Comp. 16, 1190–1208.

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 Geosci. Remote S. Mag. 4(2), 58–78.

Capra, A., Scicolone, B., 2007. Recycling of poor quality urban wastewater by drip irrigation systems. J. Clean. Prod. 15(16), 1529–1534.

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 Sens. Environ. 204, 197–211.

Duran–Ros, M., Puig–Bargués, J., Arbat, G., Barragán, J., Ramírez de Cartagena, F., 2009. Effect of filter, emitter and location on clogging when using effluents. Agr. Water Manage. 96(1), 67–79.

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

Fei, Y., Rong, G., Wang, B., Wang, W., 2014. Technical section: parallel L-BFGS-B algorithm on GPU. Comput. Graph. 40, 1–9.

Freedman, D., Pisani, R., Purves, R., 2007. Statistics. 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 . Comput. Electron. Agric. 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. Biosyst. Eng. 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. Comput. Electron. Agric. 139, 65–74.

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

Hastie, T., Tibshirani, R., Friedman, J.H., 2003. The Elements of Statistical Learning. Springer–Verlag, New York.

Hawari, A. H., Alnahhal, W., 2016. Predicting the performance of multi-media filters using artificial neural networks. Water Sci. Tech. 74 (9), 2225–2233.

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. IEEE, San Diego, CA, USA, pp. 1901–1907.

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. IEEE, Roma, Italy, pp. 742–747.

Kong, D., Chen, Y., Li, N., 2018. Gaussian process regression for tool wear prediction. Mech. Syst. Signal Pr. 104, 556–574.

Kuhn, M., Johnson, K., 2018. Applied Predictive Modeling. Springer, New York.

Liu, D.C., Nocedal, J., 1989. On the limited memory BFGS method for large scale optimization. Math. Program. 45, 503–528.

Liu, H., Yang, C., Huang, M., Wang, D., Yoo, C., 2018. Modeling of subway indoor air quality using Gaussian process regression. J. Hazard. Mater. 359, 266–273.

Ma, X., Xu, F., Chen, B., 2019. Interpolation of wind pressures using Gaussian process regression. J. Wind Eng. Ind. Aerod. 188, 30–42.

Madramootoo, C.A., Morrison, J., 2013. Advances and challenges with micro-irrigation. Irrig. Drain. 62 (3), 255–261.

Marsland, S., 2014. Machine Learning: An Algorithmic Perspective. Chapman and Hall/CRC Press, Boca Raton, FL, USA.

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. Comput. Electron. Agric. 99, 176–185.

Martin, O., 2018. Bayesian Analysis with Python. Packt Publishing, Birmingham, UK.

Mather, A.L., Johnson, R.L., 2015. Event-based prediction of stream turbidity using a combined cluster analysis and classification tree approach. J. Hydrol. 530, 751–761.

Murphy, K.P., 2012. Machine Learning: A Probabilistic Perspective. The MIT Press, Cambridge, MA, USA.

Noori, M., Hassani, H., Javaherian, A., Amindavar, H., Torabi, S., 2019. Automatic fault detection in seismic data using Gaussian process regression. J. Appl. Geophys. 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). Naha, Okinawa, Japan, pp. 1743–1752.

Picard, R., Cook, D., 1984. Cross-validation of regression models. J. Am. Stat. Assoc. 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). IEEE, Vietri sul Mare, Italy, pp. 1–6.

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. Biosyst. Eng. 111(1), 126–132.

Pujol, J., Duran–Ros, M., Arbat, G., Ramírez de Cartagena, F., Puig–Bargués, J., 2011. Private micro-irrigation costs using reclaimed water. Span. J. Agric. Res. 9(4), 1120–1129.

Rasmussen, C.E., 2003. Gaussian Processes in Machine Learning: Summer School on Machine Learning. Springer, Berlin, Heidelberg.

Rasmussen, C.E., Williams, C.K.I., 2006. Gaussian Processes for Machine Learning. The MIT Press, Cambridge, MA, USA.

Ravina, I., Paz, E., Sofer, Z., Marm, A., Schischa, A., Sagi, G., Yechialy, Z., Lev, Y. 1997. Control of clogging in drip irrigation with stored treated municipal sewage effluent. Agric.Water Manage. 33(2–3), 127–137.

Rogers, S., Girolami, M., 2016. A First Course in Machine Learning. Chapman and Hall/CRC, Boca Raton, FL, USA.

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. IEEE, Taipei, Taiwan, pp. 255–260.

Schulz, E., Speekenbrink, M., Krause, A., 2018. A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions. J. Math. Psychol. 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. MIT Press Cambridge, MA, USA, Vol. 12, pp. 603–609.

Shi, J.Q., Choi, T., 2011. Gaussian Process Regression Analysis for Functional Data. Chapman and Hall/CRC Press, Boca Raton, FL, USA.

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. Biosyst. Eng. 187, 292–304.

Stevenson, M., Bravo, C., 2019. Advanced turbidity prediction for operational water supply planning. Decis. Support Syst. 119, 72–84.

Tal, A., 2016. Rethinking the sustainability of Israel's irrigation practices in the drylands. Water Res. 90, 387–394.

Tien, C., 2012. Principles of Filtration. Elsevier, Kidlington, Oxford, UK.

Tripathi, V.K., Rajput, T.B.S., Patel, N., 2014. Performance of different filter combinations with surface and subsurface drip irrigation systems for utilizing municipal wastewater. Irrigation Sci. 32(5), 379–391.

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

USEPA 2012 Guidelines for Water Reuse. EPA/600/R-12/618. US Environmental Protection Agency, Washington D.C. and Cincinnati, Ohio.

Vidales, A., 2019. Machine Learning with Matlab: Gaussian Process Regression, Analysis of Variance and Bayesian Optimization. Independently published.

Wasserman, L., 2003. All of Statistics: A Concise Course in Statistical Inference. Springer, New York.

Wen–Yong, W., Yan, H., Hong–Lu, L., Shi–Yang, Y., Yong, N., 2015. Reclaimed water filtration efficiency and drip irrigation emitter performance with different combinations of sand and disc filters. Irrig. Drain. 64 (3), 362–369.

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

Zhou, B., Zhou, H., Puig–Bargués, J., Li, Y., 2019. Using an anti-clogging relative index (CRI) to assess emitters rapidly for drip irrigation systems with multiple low-quality water sources. Agric. Water Manage. 221, 270–278.

Zhu, C., Byrd, R.H., Lu, P., Nocedal, J., 1997. Algorithm 778: L–BFGS–B: Fortran subroutines for large-scale bound-constrained optimization. ACM T. Math. Software 23(4), 550–560.

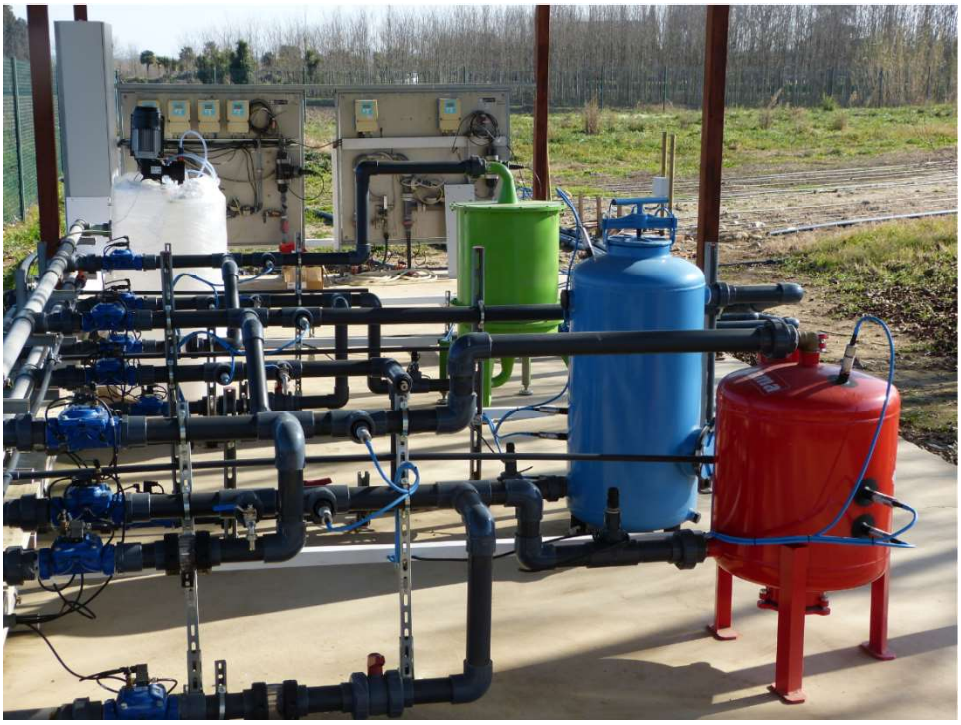

Fig. 1. Picture of the experimental set-up with the three filter designs: (a) red: arm collector; (b) blue: inserted domes; and (c) green: porous media prototype.

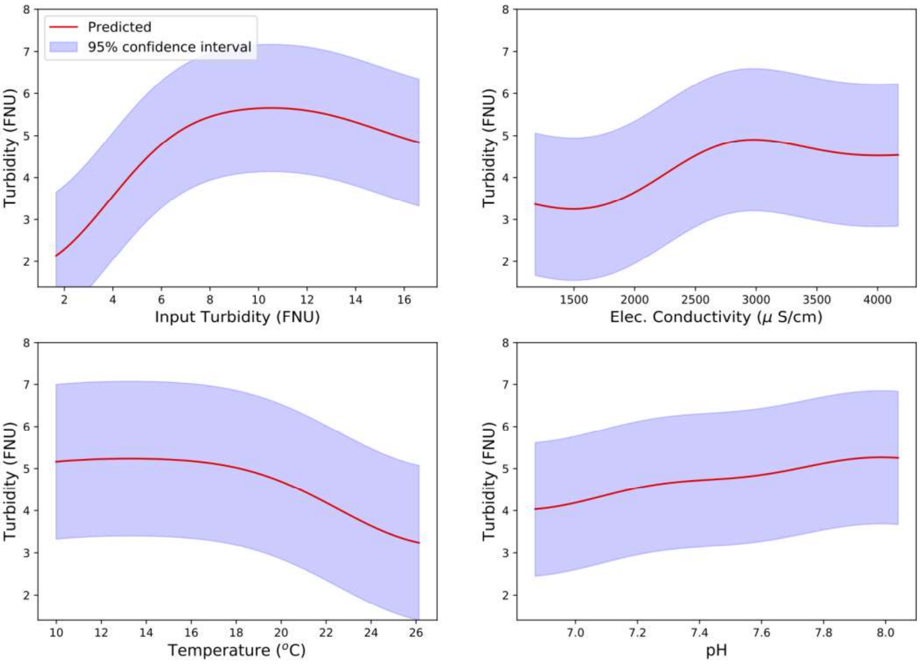

Fig. 2. First-order terms for some of the independent variables for the dependent variable output turbidity $\left(\mathrm{Turb}_o\right)$.

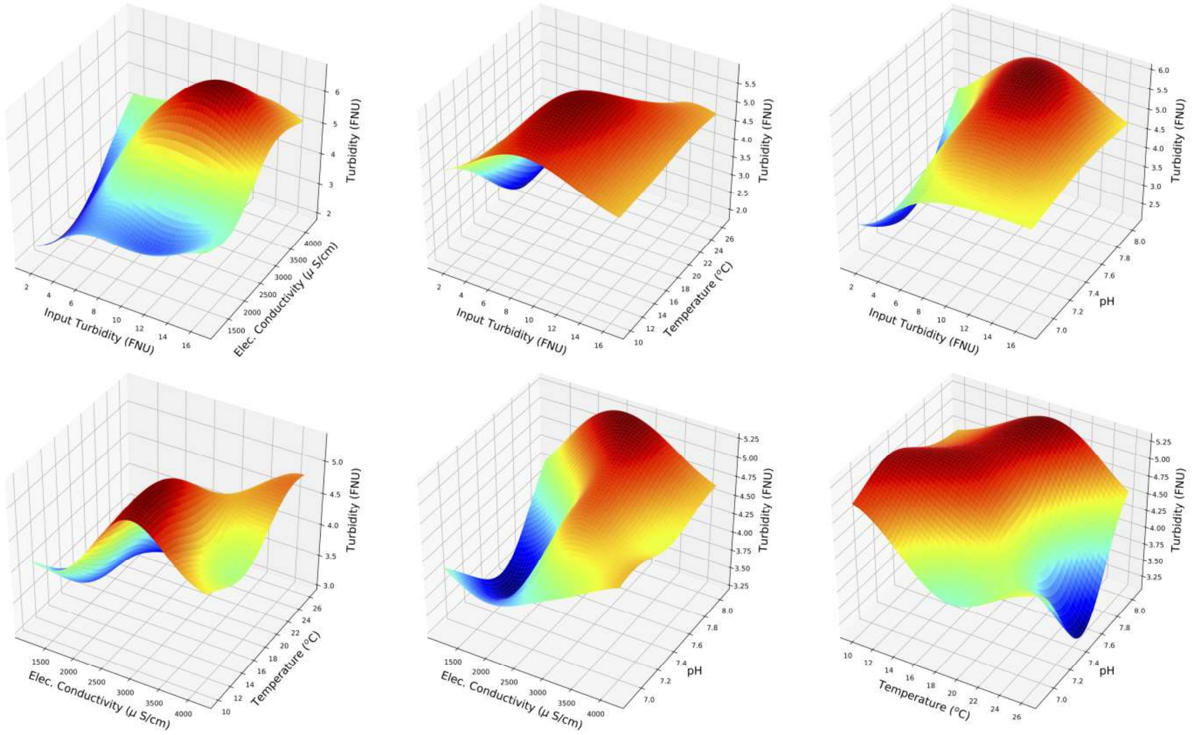

Fig. 3. Second-order terms of some of the independent variables for the dependent variable output turbidity (Turb_o).

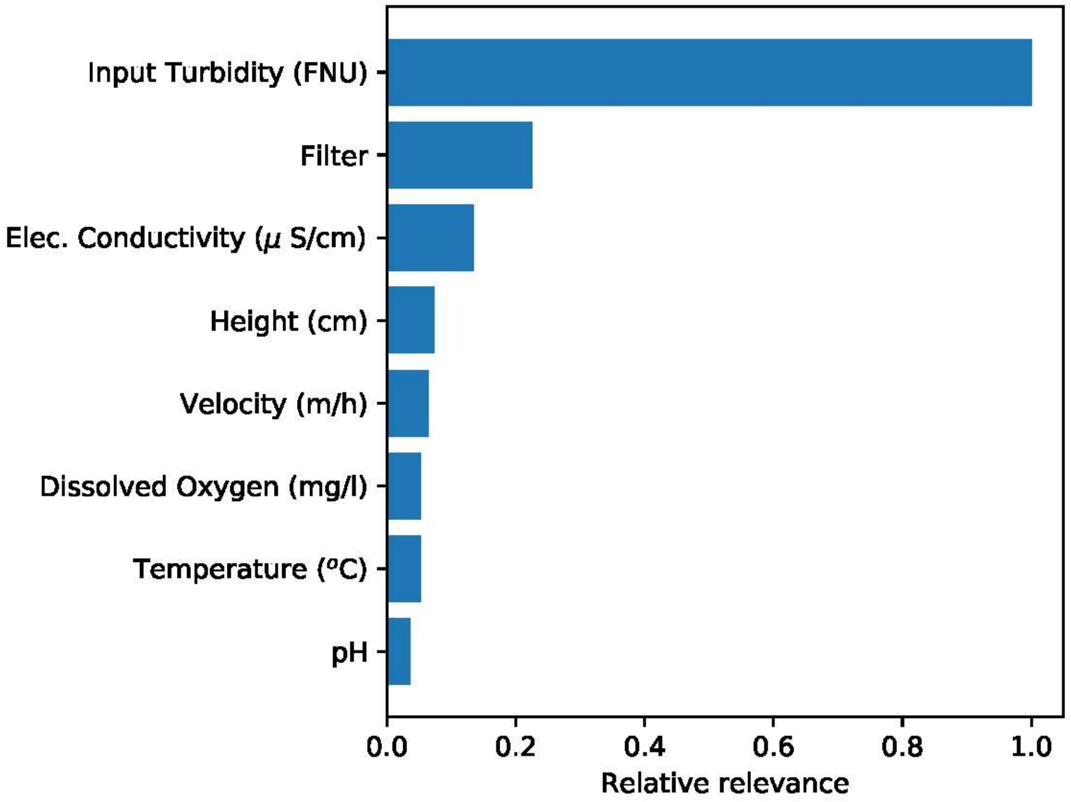

Fig. 4. Relative relevance of the variables in the GPR model for the outlet turbidity ($Turb_o$).

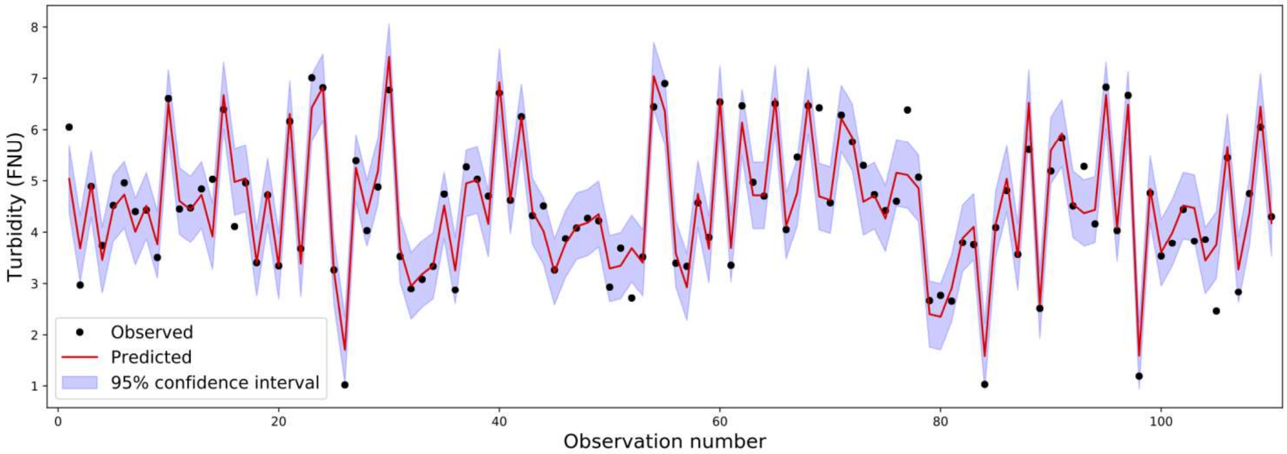

Fig. 5. Observed and predicted $\mathrm{Turb_o}$ values, taking into account the confidence interval, by using the GPR–based model with the testing set ($R^2 = 0.8921$).

Table 1

Set of operation physical input variables used in this study along with their mean, median, standard deviation (STD) and mean absolute deviation (MAD).

Input variables	**Name of the variable**	**Mean**	**Median**	**STD**	**MAD**
Filter media type	Filter	--	--	--	--
Height of the filter bed (cm)	H	25.631	30.000	4.9601	0.0000
Filtration velocity (m/h)	v	49.909	60.000	14.174	0.0000
Electrical conductivity (μ S/cm)	CE_i	2575.6	2639.0	497.68	285.00
Dissolved oxygen (mg/l)	DO_i	3.3529	3.3300	0.9860	0.6700
pH	pH_i	7.3526	7.3800	0.2229	0.1400
Input turbidity (FNU)	$Turb_i$	6.1029	5.8000	2.5898	1.5800
Water temperature (ºC)	T_i	20.002	19.960	3.3486	2.6200

Table 2

Optimal hyperparameters of the best fitted GPR–based model found with the LBFGSB technique: variance σ_f^2 and lengthscale 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 fun. value
$Turb_o$	1.05	1.56	0.0298	174

Table 3

Log marginal likelihood variation value between the full model and the model without the variable for the Turb$_o$ model.

Variable	**Likelihood variation**
Input Turbidity (FNU)	566
Filter	128
Electrical Conductivity (μ S/cm)	76
Height (cm)	42
Velocity (m/h)	36
Dissolved Oxygen (mg/l)	30
Temperature (ºC)	29
pH	20