Hindawi
Computational Intelligence and Neuroscience
Volume 2019, Article ID 2859429, 12 pages
https://doi.org/10.1155/2019/2859429

Research Article

Gaussian Process Regression Tuned by Bayesian Optimization for Seawater Intrusion Prediction

George Kopsiaftis,[1] Eftychios Protopapadakis,[1] Athanasios Voulodimos,[2] Nikolaos Doulamis,[1,3] and Aristotelis Mantoglou[1]

[1]*National Technical University of Athens, 15773 Athens, Greece*
[2]*Department of Informatics and Computer Engineering, University of West Attica, 12243 Athens, Greece*
[3]*Institute of Communication and Computer Systems (ICCS), Zografou 15773, Athens, Greece*

Correspondence should be addressed to Athanasios Voulodimos; thanosv@mail.ntua.gr

Received 20 April 2018; Revised 25 November 2018; Accepted 11 December 2018; Published 17 January 2019

Academic Editor: José Alfredo Hernández-Pérez

Accurate prediction of the seawater intrusion extent is necessary for many applications, such as groundwater management or protection of coastal aquifers from water quality deterioration. However, most applications require a large number of simulations usually at the expense of prediction accuracy. In this study, the Gaussian process regression method is investigated as a potential surrogate model for the computationally expensive variable density model. Gaussian process regression is a nonparametric kernel-based probabilistic model able to handle complex relations between input and output. In this study, the extent of seawater intrusion is represented by the location of the 0.5 kg/m^3 iso-chlore at the bottom of the aquifer (seawater intrusion toe). The initial position of the toe, expressed as the distance of the specific line from a number of observation points across the coastline, along with the pumping rates are the surrogate model inputs, whereas the final position of the toe constitutes the output variable set. The training sample of the surrogate model consists of 4000 variable density simulations, which differ not only in the pumping rate pattern but also in the initial concentration distribution. The Latin hypercube sampling method is used to obtain the pumping rate patterns. For comparison purposes, a number of widely used regression methods are employed, specifically regression trees and Support Vector Machine regression (linear and nonlinear). A Bayesian optimization method is applied to all the regressors, to maximize their efficiency in the prediction of seawater intrusion. The final results indicate that the Gaussian process regression method, albeit more time consuming, proved to be more efficient in terms of the mean absolute error (MAE), the root mean square error (RMSE), and the coefficient of determination (R^2).

1. Introduction

Seawater intrusion (SI) in coastal aquifers is a complex physical phenomenon, consisting of several physical processes. A number of approaches have been proposed to simulate SI, considering different components. Dispersion mechanisms and water density changes are considered critical components in the accurate representation of SI [1]. Both mechanisms are incorporated in the mathematical description of what is known as variable density (VD) models. Although accurate, VD models are CPU intensive and entail long runtimes because the resulting model equations are solved using complex numerical methods (e.g., finite differences and finite element methods). The time-consuming simulations hinder the exploitation of the high accuracy VD models in applications which require a large number of iterations, such as coastal groundwater management, parameter estimation, sensitivity analysis, and uncertainty analysis. Because of the long runtimes, it is also rather impractical to incorporate VD models in real-time systems, e.g., decision support systems [2]. A common method to tackle the duration problem is the use of very fast approximation models, which could efficiently substitute the original VD models, without compromising the accuracy of

the results. These models are usually called surrogate models, metamodels, model emulators, lower fidelity models, proxy models, and response surfaces [2–5].

Surrogate model practice is based on the notion that original model response(s) could be approximated by a computationally more efficient model, for a range of values of the selected model variables [5]. In the present study, the Gaussian process regression (GPR) as a surrogate model for SI is examined. Rajabi and Ketabchi [6] summarized the advantages of GPRs compared with other surrogate models in the following: (i) GPRs provide both an approximation of the original high-fidelity model results and a probabilistic estimate of the approximation uncertainties [7, 8], (ii) GPRs' structure is relatively simple based on the mean and covariance functions [9], (iii) GPRs are flexible with regard to the probability distributions of the input data, (iv) GPRs can efficiently cope with models of different complexity [10, 11], (v) GPRs provide the ability to calculate the mean and standard deviation, and (vi) GPRs provide the ability to incorporate prior knowledge of the outputs in the metamodel construction process [12].

The GPR results are compared with other widely used methods, specifically, linear regression (LR), support vector machine regression (SVMR), binary regression decision tree (BRDT), and ensemble tree learners (ETL). It should be noted that the examined methods are all univariate. A Bayesian optimization is employed in all surrogate models, to improve their efficiency.

The remainder of this study is structured as follows: Section 2 presents a brief survey of the related work. In Section 3, the seawater intrusion model is described, whereas section 4 presents the proposed Gaussian process regression scheme and the Bayesian optimization process. In Section 5, the experimental evaluation is provided, and finally, section 6 concludes the paper.

2. Related Work

Approximation models have been widely used during the last decade in water resources (e.g., [13, 14]) and especially in groundwater modelling. Razavi et al. [5] and Asher et al. [2] performed an extended review of surrogate model applications in water resources field. Regarding coastal aquifers, surrogate models have been widely used for the prediction of SI, substituting the complex fluid flow and transport processes. For example, Bhattacharjya et al. [15] employed artificial neural networks (ANN) to approximate density-dependent flow in coastal aquifers. In more recent studies, Roy and Datta [16] used the fuzzy C-mean clustering method to predict SI, while Lal and Datta [17] investigated the ability of Support Vector Machine regression (SVMr) to predict the location of SI toe and concluded that the method surpasses other widely used metamodelling methods, such as genetic programming (GP).

A significant number of relative studies are devoted to the use of surrogate models in coastal aquifer management problems to cope with the computational burden, which arises from simulation-optimization schemes [18]. A well-established metamodel which is very common in coastal aquifer management literature is the artificial neural networks (ANNs) [5, 18]. Bhattacharjya and Datta [19] employed an ANN model to approximate a density-dependent model in a genetic optimization framework. Rao et al. [20] and Kourakos and Mantoglou [21] incorporated ANNs in a simulation-optimization scheme to replace the SEAWAT numerical code. Kourakos and Mantoglou [22] proposed a pumping optimization method based on modular neural networks and an Evolutionary Annealing Simplex optimization algorithm. Ataie-Ashtiani et al. [23] combined a simulation-optimization procedure with ANNs to develop an efficient model for the multi-objective management of groundwater lenses in small islands. Christelis and Mantoglou [24] used cubic radial basis functions (RBFs) in two adaptive metamodeling frameworks: (1) the adaptive-recursive approach and (2) the metamodel-embedded evolutionary strategy. The latter proved to be computationally more efficient, providing solutions near the global optimum. Christelis et al. [25] employed two surrogate-based optimization (SBO) frameworks, under restricted computational budgets to improve the efficiency of optimization algorithms in problems of moderate and large dimensionalities. In a more recent study, Christelis and Mantoglou employed variable-fidelity surrogate models and evolutionary algorithms to calculate the maximum allowed pumping rates in coastal aquifers. Sreekanth and Datta in several studies [26–28] examined genetic programming (GP) as a potential surrogate model in multiobjective management of SI in coastal aquifers and compared the proposed method with modular neural network metamodels. Roy and Datta examined several surrogate models to predict SI in coastal aquifers and employed all these models in coastal aquifer management problems [29–31]. A review of surrogate models, focusing on SI and coastal aquifer management, is presented by Roy and Datta [32].

Gaussian process metamodels have been widely used in engineering optimization applications, but according to Razavi et al. [5] and Asher et al. [2], they have not yet attracted much attention in groundwater field, particularly SI. Stone [33] presented a Bayesian emulation methodology as an alternative to Monte Carlo in the analysis of stochastic groundwater models. Zhang et al. [34] employed an adaptive Gaussian process-based method to identify contaminant source in groundwater problems, whereas Crevillén-García et al. [35] used Gaussian process method to perform uncertainty analysis in a convectively enhanced dissolution process model. Raghavendra and Deka [36] used GPR and adaptive neuro fuzzy inference system (ANFIS) to forecast groundwater level time series. In a recent study, Rajabi and Ketabchi [6] used Gaussian process emulators in a simulation-optimization framework to address the computational challenges arising from the large number of the required simulations. Roy and Datta [37] incorporated three metamodels, particularly ANFIS, GPR, and multivariate adaptive regression spline (MARS), in a multiobjective optimization framework to quantify the influence of sea-level rise on coastal aquifer management. In this specific study, they concluded that the ANFIS-based metamodel

proved to be more efficient and inexpensive compared with the other two metamodels. The authors also performed a comparative analysis between several surrogate models [38], including GPR, in a coupled simulation-optimization methodology under parameter uncertainty. In this specific study, they concluded that the GPR metamodels and their ensemble (EGPR) proved to be more efficient in terms of prediction compared with other similar methods, such as the MARS metamodel and the regression tree (RT) metamodel.

3. Seawater Intrusion Model

3.1. Variable Density and Salt Transport Model. As mentioned in section 1, VD models are based on the spatial variability of groundwater density, which ranges from saline water density to freshwater density. The driving force of the seawater/freshwater mixing is the dispersion mechanism, which results in the existence of a transition zone across the entire coastline. The width and exact position of the zone depends on the aquifer parameters and the pumping regime. In the current study, thermal and viscosity effects are neglected and the density changes are attributed only to concentration effect. The flow and solute transport equations are used to describe mathematically the VD model. The two equations form a coupled differential equation system, which could be expressed as follows [39]:

$$-\nabla \cdot (\rho \mathbf{q}) + \rho_s q_s = \rho S_f \frac{\partial h_f}{\partial t} + n \frac{\partial \rho}{\partial C} \frac{\partial C}{\partial t}, \quad (1)$$

$$\frac{\partial C}{\partial t} = \nabla \cdot (\mathbf{D} \cdot \nabla C) - \nabla \cdot (\mathbf{v} C) - \frac{q_s}{n} C_s, \quad (2)$$

where ρ is the fluid density, $\mathbf{q}$ is the specific discharge vector, ρ_s is the density of water entering from a source of leaving through a sink, q_s is the volumetric flow rate per unit volume of porous medium representing sources and sinks, S_f is the specific storage, h_f is the freshwater head, n is the porosity, C is the solute concentration, $\mathbf{D}$ is the hydrodynamic dispersion tensor, $\mathbf{v}$ is the fluid velocity vector, and C_s is the solute concentration of water entering or leaving through sources and sinks, respectively. Because solute reaction is not considered, fluid density is only a function of the solute concentration C, according to the following equation:

$$\rho = \rho_o \left(1 + \frac{\varepsilon}{(C_s - C_o)} (C - C_o) \right), \quad (3)$$

in which ρ_o is the freshwater density, ε is the density difference ratio (equation (4)), C_o is the reference concentration, and C_s is the maximum concentration. In this study, the following values are used for the parameters of equation (3): $\rho_o = 1000\,\text{kg/m}^3$, $C_o = 0\,\text{kg/m}^3$, and $C_s = 35\,\text{kg/m}^3$.

The density difference ratio is expressed as

$$\varepsilon = \frac{\rho_s - \rho_o}{\rho_o}, \quad (4)$$

where ρ_s stands for the maximum seawater density. In this study, we consider $\rho_s = 1025\,\text{kg/m}^3$.

The Darcy flux term $\mathbf{q}$ of equation (1) for constant viscosity and freshwater properties could be expressed as

$$q_x = -K_{fx} \left(\frac{\partial h_f}{\partial x} \right), \quad (5)$$

$$q_y = -K_{fy} \left(\frac{\partial h_f}{\partial y} \right), \quad (6)$$

$$q_z = -K_{fz} \left(\frac{\partial h_f}{\partial z} + \frac{\rho - \rho_f}{\rho} \right), \quad (7)$$

where q_x, q_y, and q_z are the components of the specific discharge in the principal directions, K_{fx}, K_{fy}, and K_{fz} are the components of the freshwater hydraulic conductivity in the same directions, and ρ_f is the freshwater density.

Equations (1) to (7) are the mathematical representation of the VD approach of seawater intrusion. The well-established SEAWAT code is used to solve numerically the aforementioned equation set. SEAWAT is a modular finite difference computer code created by USGS, which couples MODFLOW and MT3DMS, to solve iteratively the fluid flow and solute transport equations [39].

3.2. Coastal Aquifer Case Study. The VD model is applied on a rectangular-shaped unconfined aquifer. The dimensions of the aquifer model are $L = 7000\,\text{m}$, $W = 3000\,\text{m}$, and $d = 25\,\text{m}$. The examined aquifer geometrically resembles a real coastal aquifer located at the central eastern part of the Greek island Kalymnos, specifically the elongate aquifer underlying the Vathi valley [40, 41]. It should be noted that the examined model is an abstraction of the real aquifer, which could be considered as a typical aquifer example for the Aegean Greek islands, in terms of size and shape.

Figure 1 outlines the conceptual model of the aquifer model. A hydrostatic boundary condition (BC) is assigned on the seaside boundary. The aquifer is bounded by impermeable geological formations, with the exception of the inland boundary, where a specified flux BC is applied to simulate the lateral inflow from the adjacent aquifer. The groundwater is replenished by a constant recharge, which is uniformly distributed along the entire surface of the aquifer. For simplification purposes, the aquifer is considered homogeneous and an anisotropic factor is assumed, which represents the differential permeability along the vertical direction. Table 1 presents the values of the basic fluid flow and solute transport parameters. An initial simulation for approximately 200 yr without pumping was performed, until steady flow/steady transport conditions are achieved. The final hydraulic head and concentration values of this simulation are used as the initial conditions for the SI simulations, which are related to the training of the surrogate models. All simulations in the current paper are considered steady state, regarding the fluid flow conditions. This assumption resulted in relatively brief VD simulations, which allowed for the creation of an adequate sample for the calibration of the surrogate models. The duration of each simulation was approximately 1–2 min. The simulations were performed in an i7-4770 quad-core processor 3.4 GHz, with 8 GB RAM.

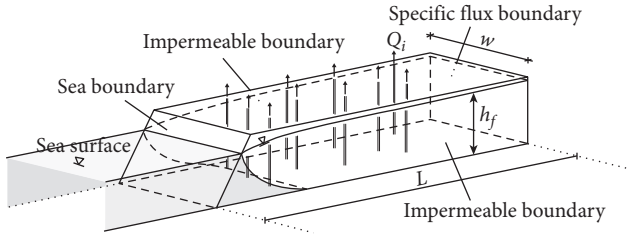

Figure 1: 3D representation of the examined coastal aquifer.

The 0.5 kg/m^3 iso-chlore is considered as an indicative surface, representing the seawater intrusion wedge. Specifically, the location of the intersection of the iso-chlore and the aquifer bottom, known as the toe of the wedge, is used as a measure of the seawater intrusion extend. Further details concerning the calculation of the toe are discussed in the following sections.

4. Gaussian Process Regression and Other Models in Seawater Intrusion

4.1. Gaussian Process Regression. Gaussian process regression (GPR) is a nonparametric kernel-based probabilistic model [8]. Just like other Bayesian methods, GPR do not aim at finding "best-fit" models of the data by relating the underling function $f(\mathbf{x})$ to a specific form (e.g., linear or quadratic). Instead, they calculate posterior predictive distributions for new test inputs. Such an approach enables the quantification of uncertainty as regards model estimates, as well as leveraging the understanding of the uncertainty to improve the robustness of predictions on future test points [43].

Gaussian processes can be considered as the extension of multivariate Gaussians to infinite-sized collections of variables of real value. More specifically, a Gaussian process is a collection of random variables $\{f(\mathbf{x}) : \mathbf{x} \in \mathrm{X}\}$ defined by its mean function $\mu(\mathbf{x})$ and a covariance function $k(\mathbf{x}, \mathbf{x}')$ so that

$$\begin{bmatrix} f(\mathbf{x}_1) \\ \vdots \\ f(\mathbf{x}_n) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu(\mathbf{x}_1) \\ \vdots \\ \mu(\mathbf{x}_n) \end{bmatrix} \begin{bmatrix} k(\mathbf{x}_1, \mathbf{x}_1) & \dots & k(\mathbf{x}_1, \mathbf{x}_n) \\ \vdots & \ddots & \vdots \\ k(\mathbf{x}_n, \mathbf{x}_1) & \dots & k(\mathbf{x}_n, \mathbf{x}_n) \end{bmatrix} \right). \tag{8}$$

The above statement can be rewritten as follows:

$$f(\cdot) \sim \mathcal{GP}(\mu(\cdot), k(\cdot, \cdot)), \tag{9}$$

where each dimension of the Gaussian corresponds to an element x from the index set X. Furthermore, the respective component of the random vector represents the $f(\mathbf{x})$ value. Typically, the prior distribution over functions $f(\cdot)$ is expected to be a zero-mean GP prior.

Consider a training set $\mathcal{L} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ of i.i.d. examples from some unknown distribution, where $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. A GPR model assumes that a response y_i satisfies the following equation:

$$y_i = f(\mathbf{x}_i) + \epsilon_i, \tag{10}$$

Table 1: Basic parameters of the flow and transport model (source [42]).

Parameters	Values
K_x	100 m/d
K_y	100 m/d
K_z	1 m/d
Longitudinal dispersivity	50 m
Transverse dispersivity	5 m
Vertical dispersivity	0.5 m
Density ratio	0.025
Recharge	8.22×10^{-5} m/d
Lateral inflow	3696 m^3/d

where ϵ_i are i.i.d. noise variables, so that $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Let $\mathcal{U} = \left\{(\mathbf{x}_i^{(u)}, y_i^{(u)})\right\}_{i=1}^n$ be a set of i.i.d. testing points drawn from the same unknown distribution as $\mathcal{L}$. Recall that both training and test points must have a joint multivariate Gaussian distribution.

Then, it can be proved that [1]

$$\overline{y}^{(u)} \left| \overline{y}, X, X^u \sim \mathcal{N}\left(\mu^{(u)}, \Sigma^{(u)}\right), \right. \tag{11}$$

with mean value and covariance defined as

$$\mu^{(u)} = K\left(\mathbf{X}^{(u)}, \mathbf{X}\right) \cdot \left(K(\mathbf{X}, \mathbf{X}) + \sigma^2 \cdot I\right)^{-1} \cdot \overline{y}, \tag{12}$$

$$\begin{aligned} \Sigma^{(u)} = {} & K\left(\mathbf{X}^{(u)}, \mathbf{X}^{(u)}\right) + \sigma^2 \cdot I - K\left(\mathbf{X}^{(u)}, \mathbf{X}\right) \\ & \cdot \left(K(\mathbf{X}, \mathbf{X}) + \sigma^2 \cdot I\right)^{-1} \cdot K\left(\mathbf{X}, \mathbf{X}^{(u)}\right), \end{aligned} \tag{13}$$

respectively. Note that $K(\mathbf{X}^{(u)}, \mathbf{X}) \in \mathbb{R}^{n \times n}$ is defined as $K(\mathbf{X}^{(u)}, \mathbf{X})_{ij} = k(\mathbf{x}_i^{(u)}, \mathbf{x}_j)$, $i, j = 1, \dots, n$. The same hold for the $K(\mathbf{X}, \mathbf{X})$, $K(\mathbf{X}^{(u)}, \mathbf{X}^{(u)})$, and $K(\mathbf{X}, \mathbf{X}^{(u)})$ cases.

Additionally,

$$\begin{aligned} \overline{y} &= \left[y_1, \dots, y_n\right]^T, \\ \overline{y}^{(u)} &= \left[y_1^{(u)}, \dots, y_n^{(u)}\right]^T, \\ \mathbf{X} &= \begin{bmatrix} x_{11} & \dots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{nm} \end{bmatrix}, \quad \mathbf{X} \in \mathbb{R}^{n \times m}, \end{aligned} \tag{14}$$

$\mathbf{X}^{(u)}$ is defined in a similar way.

As such, we can estimate any new value as the mean of a posterior predictive distribution. We should also note that with the rise of training samples number, the confidence region size reduces, so as to reflect the decreasing uncertainty in the model estimates.

4.2. Alternative Regressors for Comparison. Regression trees is an alternative approach to nonlinear regression. The core idea lies in sub-dividing the space into smaller regions and then fit simple models to them [42, 44]. Provided a training set $\mathcal{L}$, a set of branches is created. Each binary split is performed according to a specific feature (from the m available). Then, a new value prediction is defined as

$$y(\mathbf{x}) = \frac{1}{c}\sum_{i=1}^{c} y_i, \tag{15}$$

where c is the number of observations available at the specific cell.

Support Vector Machine regression is another approach. The function used to predict new values (for linear support vector regression) is defined as [45]

$$y(\mathbf{x}) = \sum_{i=1}^{n} \left(\boldsymbol{\alpha}_i - \boldsymbol{\alpha}_i^*\right) \cdot \mathbf{x}_i, \mathbf{x} + b, \tag{16}$$

where $\langle \cdot, \cdot \rangle$ is the dot product, $\boldsymbol{\alpha}_i, \boldsymbol{\alpha}_i^*$ are Langrage multipliers, so that $\boldsymbol{\alpha}_i \cdot \boldsymbol{\alpha}_i^* = 0, i = 1, \ldots, n$, and b is a bias term. SV algorithm can be made nonlinear by simply preprocessing the training patterns $\mathbf{x}_i$ using a kernel function $k(\cdot, \cdot)$. The regression is performed as

$$y(\mathbf{x}) = \sum_{i=1}^{n} \left(\boldsymbol{\alpha}_i - \boldsymbol{\alpha}_i^*\right) \cdot k\left(\mathbf{x}_i, \mathbf{x}\right) + b. \tag{17}$$

4.3. Bayesian Optimization of Model Parameters. A variety of widely used machine learning techniques contain a significant number of parameters to be decided (e.g., SVM kernel type and parameters and ANN layers and type of activation functions). The performance of any algorithm depends on the selection of these hyperparameters [46–48]. Typically, hyperparameter tuning involves grid search, random search, and genetic algorithms, among many other techniques [49]. Such techniques require many (nonconvex) function evaluations.

Bayesian optimization (BayesOpt) is a surrogate modelling technique that can optimize an objective function that is expensive to evaluate, reducing the number of actual function evaluations required [50, 51]. It is built on Bayesian inference and Gaussian processes and is applicable in cases where closed-form expression for the objective function is not known but can obtain observations (possibly noisy) of this function at sampled values.

BayesOpt builds a probabilistic proxy model for the objective, using outcomes of past experiments as training data. The proxy model (e.g., Gaussian process) is much cheaper to calculate but it can provide adequate information on where we should evaluate the true objective function to get a good result. Assume a vector $\mathbf{P} = \{\mathbf{p}_1, \ldots, \mathbf{p}_m\}$ for a set of m hyperparameters to be tuned. Given a set of training paradigms $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, we need to find $\mathbf{P}^* = \underset{\mathbf{P}}{\operatorname{argmin}}\, g(\mathbf{P}|\{(\mathbf{x}_i, y_i)\}_{i=1}^n)$, where g is a cost function (e.g., cross-entropy cost and quadratic cost).

The entire optimization approach is guided by an appropriate acquisition function (AF), which defines the next point (i.e., set of hyperparameters) to be evaluated. As such, any AF needs to balance between exploration and exploitation.

Exploration refers to region search where the uncertainty is high, expecting to find a new set of parameters that improve model's performance. Exploitation, on the other hand, is a region search close to already calculated high estimated values (i.e., regression performance scores).

5. Experimental Evaluation

5.1. Data Preprocessing and Experiment Setup. The training sample consists of 4000 variable sets. Each set has 40 input variables: (1) the pumping rates of the 10 wells and (2) the distance of the SI toe from 30 observation points, uniformly distributed across the sea boundary. The Latin hypercube sample (LHS) statistical method was used to generate the 4000 pumping rate patterns. As mentioned in Section 2.2, the 0.5 kg/m^3 is selected as an indicative concentration value for the SI extend. The distance between the toe and the observation wells represents the initial position of the SI wedge. Regarding the initial position of the SI toe, the variable sets are divided into four categories of 1,000 samples. In the first category, the concentration results from the zero-pumping rate simulation define the initial location of the SI toe. In the remaining categories, the solute transport and hydraulic head results of the previous category simulations are used as the initial conditions for the following simulations.

The output set consists of 30 variables, which represent the final location of the SI toe, calculated as the distance from the same observation points. Figure 2 presents the initial and final position of the SI toe for a specific set of pumping rates, representing, along with the pumping rates, the input/output variables used to train the surrogate model.

5.2. Experimental Results

5.2.1. Hyperparameter Optimization. Each regressor's hyperparameters were optimized using Bayesian optimization over 5k-fold cross-validation sets. The final parameter values are summarized in Table 2. An interesting remark is that, for the specific setup, simpler models (i.e., least square regression vs linear kernel SVMs, and linear kernel SVM vs Gaussian RBF or polynomial kernels) perform slightly better, during the optimization process.

Figure 3 illustrates the normalized performance scores of the investigated regressors for the training set, and Figure 4 provides a further insight into the actual differences (in meters), on average, for the trained models. Errors in estimation do not surpass 10 meters for the GPR and 20 meters for the TreeEns. The other regressors achieve an average error greater than 40 meters.

Figure 5 illustrates the optimization time (in minutes) required for the identification of the best possible hyperparameters, using Bayesian optimization. GPR and SVR had significantly higher training times. It is also intriguing that, for different observation points, SVR and TreeEns regressors' optimization times had increased variance. Figure 6 illustrates the case.

5.2.2. Statistical Evaluation. Statistical errors calculate the sum of differences between actual (simulated) and forecasted (regressor estimated) values. The statistical measurements used were the mean absolute error (MAE) and the root mean square error (RMSE). Low error scores suggest a good regression model.

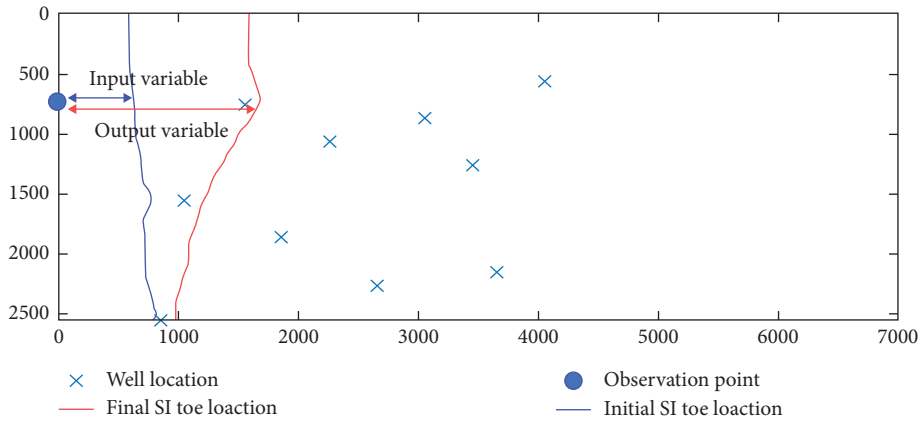

Figure 2: Initial and final position of the SI toe.

Table 2: Value ranges of optimized hyperparameters of regressors.

Regressor name	Parameter(s) name(s)	Observed value range (number of points appeared)
GPR	Kernel function	Squared exponential (30/30)
	Sigma	0.013 ± 0.005 (26/30)
LRM	Learner	Least squares
	Initial bias	-0.85 ± 0.04 (30/30)
SVM	Kernel function	Polynomial (8/20)
		Gaussian RBF (11/30)
		Linear (11/30)
BRDT	Max splits	500 ± 200 (8/30)
		900 ± 200 (8/30)
		1300 ± 200 (3/30)
		≥ 1501 (11/30)
	Number of variables to sample	All
TreeEns	Number of learners	200 ± 200 (17/30)
		≥ 400 (13/30)

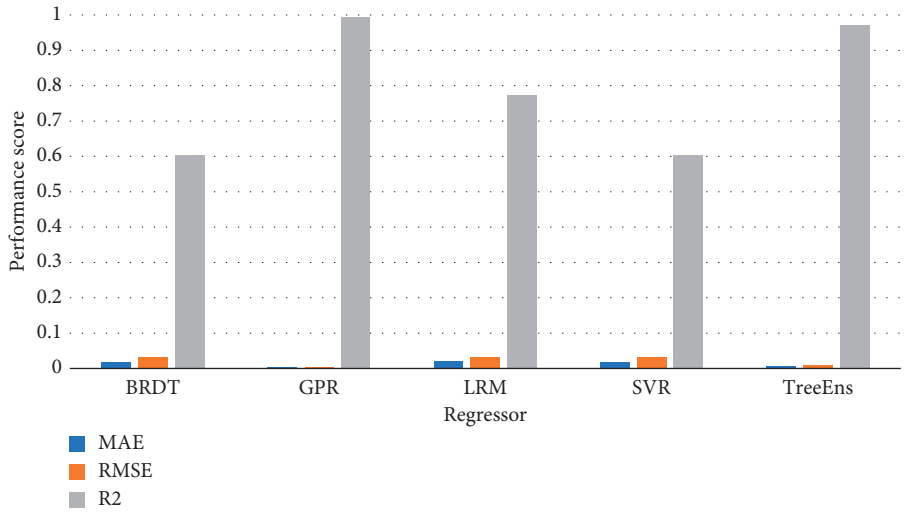

Figure 3: Normalized performance scores for the training set.

An additional performance score, i.e., coefficient of determination, R^2, is used. R^2 provides a measure of how well-observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model. Values close to 1, i.e., $R^2 \approx 1$, indicate that the regression predictions approximate extremely well the actual data outputs.

Figure 7 illustrates the average performance scores, for the proposed statistical errors. GPR surpasses all other repressors in all performance fields.

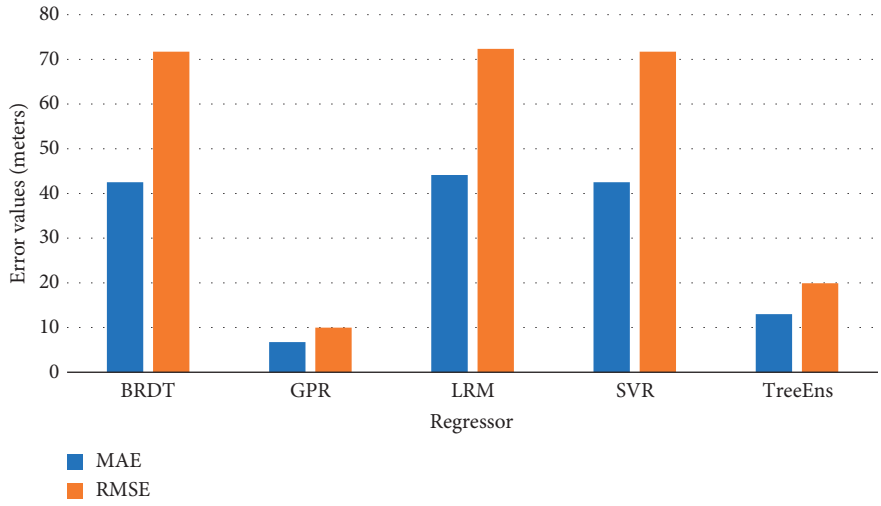

Figure 4: Error values (meters) for the training set.

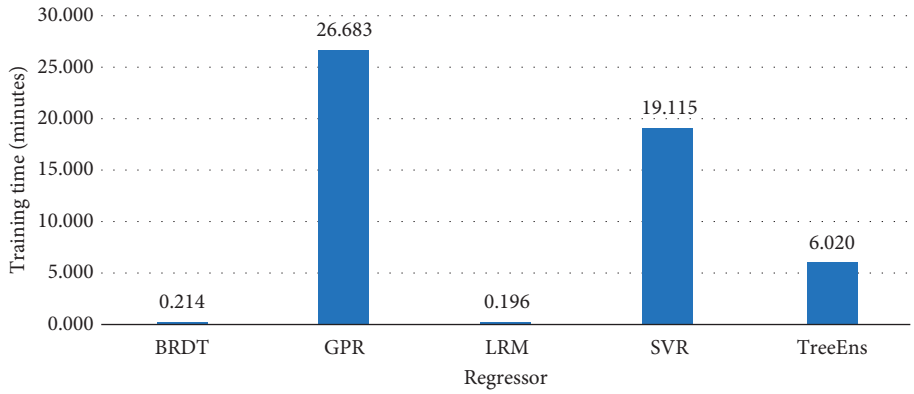

Figure 5: Optimization time for the regressors' hyperparameter estimation.

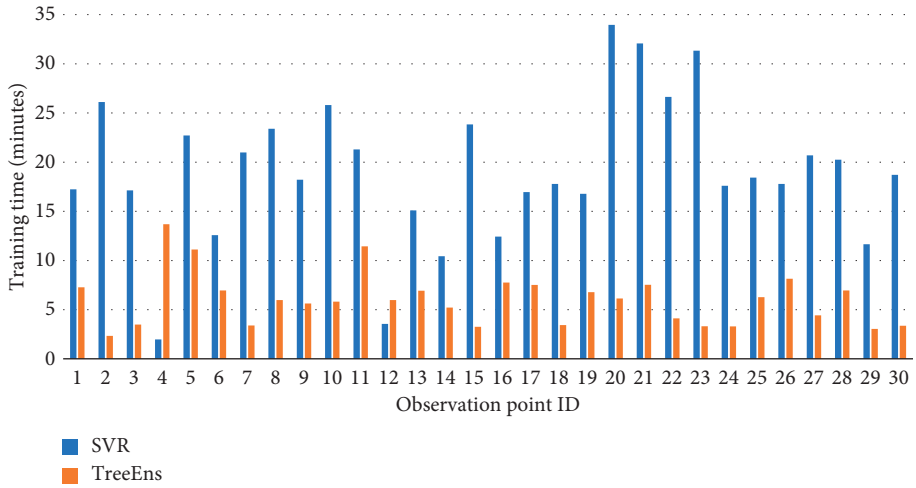

Figure 6: Optimization time variance, depending on the observation point data used for training.

5.2.3. Measuring Actual SI Toe Location Estimation Error. The statistical errors provide various information regarding the model performance. However, in our case, actual errors in the SI toe location estimations, measured in millimeters (mm), provide a deeper understanding of the regressors' performance. Figure 8 presents the discrepancies between

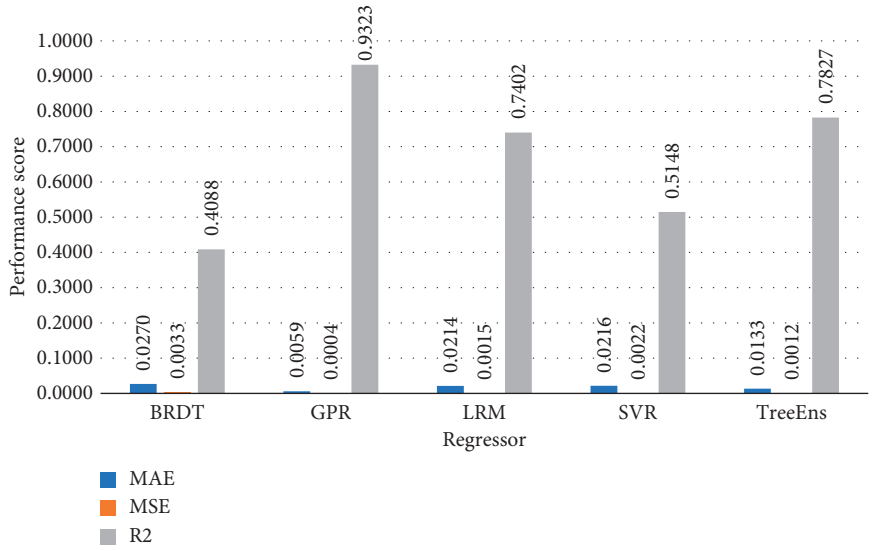

Figure 7: Average performance scores for each of the proposed regressors. GPR performs all other approaches in both error scores MAE and RMSE (lower the better) and coefficient of determination values (higher the better).

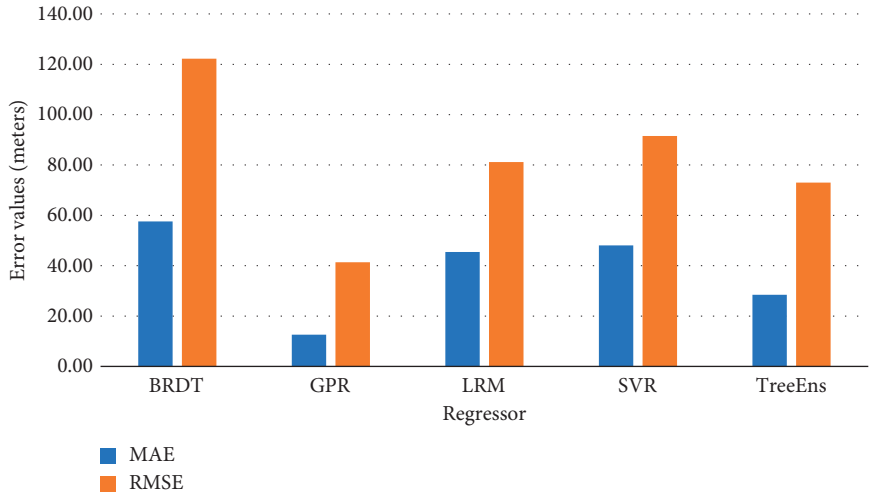

Figure 8: Discrepancy between actual and estimated distances for all observation points and regressors.

the actual and the estimated location of the SI toe, on average, for the five examined regressors. The comparative results indicate that the GPR method is overall more efficient and achieves more accurate prediction of the SI.

However, average scores fail to indicate the performance for each of the observation points, using each of the proposed regressors. Table 3 provides a further insight into the error values (meters) for each of the observation points.

5.2.4. Analysis of Variance. To obtain further insights into the results and the relative performance of the different algorithms, we conducted an analysis of variance (ANOVA) on the distance between the SI toe and the observation points score results for the test samples. The MAE score, in meters, represents a significant amount of information about the overall performance. Using this method, we can study the effects that the main design factors have [52].

Table 4 displays the outcomes of ANOVA. In this table, the "Source" column corresponds to the source of variation in data (i.e., the regressors and the observation points). Sum and mean sq. correspond to mean measurements between the m groups and the grand mean; it is a means of quantification of the variability among the groups of interest. For the degrees of freedom (d.f.), it holds that d.f. $= m - 1$. The F metric corresponds to the "average" intergroup variability divided by the "average" intragroup variability. The last column includes the p value, which is derived by comparing the F-statistic to an F-distribution with $m - 1$ numerator degrees of freedom and $n - m$ denominator degrees of freedom, for the total set of n observations.

As can be seen in Table 4, both regressors and observation points have a crucial role in explaining variations in RMSE score, given the fact that the respective p value is approximately zero. The Tukey's honest significant

Table 3: Detailed error values (in meters) per observation point.

	BRDT		GPR		LRM		SVR		TreeEns	
Point ID	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
1	58.73	124.27	10.40	45.88	41.80	82.70	14.43	76.07	31.93	80.75
2	57.45	110.07	11.10	47.58	41.46	81.62	16.58	55.14	29.48	95.61
3	59.78	132.30	12.46	52.10	42.30	84.68	12.02	52.18	22.15	49.13
4	58.55	127.25	14.78	48.58	46.34	82.30	41.16	90.72	26.19	82.46
5	65.54	135.78	19.10	57.30	51.83	87.72	23.05	56.86	46.93	89.04
6	67.87	146.06	15.01	54.75	47.48	91.85	41.46	101.26	29.38	102.24
7	62.90	125.48	13.70	51.09	44.34	95.78	24.12	98.33	23.69	81.39
8	60.40	130.50	10.43	47.40	42.82	84.43	91.60	127.27	27.11	54.66
9	62.58	136.19	13.33	44.25	43.71	84.97	12.60	53.97	25.24	69.97
10	55.96	118.68	14.24	41.08	44.15	78.52	30.35	42.57	25.07	84.73
11	59.04	124.85	10.21	44.13	46.02	82.87	14.18	52.86	25.44	83.98
12	56.08	112.88	10.95	43.10	46.64	81.33	38.15	86.49	28.04	56.87
13	59.74	113.97	11.75	39.19	47.93	77.56	36.76	90.79	23.91	71.20
14	63.33	126.04	14.21	45.23	52.04	82.21	38.14	90.45	31.00	72.95
15	61.31	121.07	21.54	45.34	56.52	83.13	43.26	94.37	31.78	57.03
16	63.08	126.62	12.80	36.78	51.33	78.68	15.42	43.84	26.24	59.00
17	61.16	122.98	16.78	38.38	48.91	78.25	37.91	86.90	24.08	66.75
18	61.55	124.98	10.25	34.19	47.96	77.53	31.53	56.39	22.88	51.39
19	56.87	126.95	10.47	33.29	47.13	81.79	48.32	82.92	39.96	70.75
20	55.55	117.58	11.32	35.46	47.19	84.37	15.19	42.14	27.20	85.57
21	56.78	129.45	11.99	37.63	46.57	82.85	32.17	96.65	29.58	68.91
22	58.32	134.78	13.21	40.35	46.35	82.16	68.38	86.52	35.45	75.93
23	54.20	116.94	12.55	38.86	45.98	83.04	89.78	153.93	31.69	90.52
24	54.40	118.41	12.00	35.88	43.58	77.67	71.91	149.64	35.01	77.75
25	51.40	115.55	10.96	33.83	41.34	75.98	311.01	314.78	32.07	83.58
26	50.83	109.44	10.92	35.07	40.83	74.10	81.77	139.24	21.12	58.01
27	47.07	105.99	10.59	33.14	40.99	75.23	13.93	40.47	32.87	68.18
28	48.54	109.17	10.39	33.64	40.74	75.62	81.09	139.40	23.28	65.71
29	49.81	113.71	9.85	31.42	39.57	72.40	53.63	77.97	24.38	71.77
30	48.81	109.37	10.30	36.14	40.33	73.53	12.15	65.30	20.26	64.56

Table 4: ANOVA outcomes.

Source	Sum sq.	d.f.	Mean sq.	F	p Value
'Regressor'	306492.5	4	76623.113	122442.294	0.00
'ObservationPoint'	140395.9	29	4841.237	7736.206	0.00
'HoldoutSet'	33.7	7	4.814	7.693	0.00
'Regressor ∗ ObservationPoint'	591604.8	116	5100.041	8149.770	0.00
'Regressor ∗ HoldoutSet'	16.8	28	0.599	0.958	0.52
'ObservationPoint ∗ HoldoutSet'	1297.4	203	6.391	10.213	0.00
'Error'	508.1	812	0.626		
'Total'	1040349.2	1199			

difference (HSD) post hoc test is also employed to identify sampling schemes and classifiers that provide the best results, while taking into consideration the statistical significance of the differences between the results.

Figure 9 indicates that GPR by far surpasses the other regressors' MAE score. Mean scores for each regressor are shown as 'o'. The average scores from the subgroups in the experiment are also provided, in the form of a horizontal line. Because there is no overlap between the RMSE values for the GPR type compared with the other regressors, GPR scores are clearly statistically better than the others [53, 54].

Figure 10 indicates that the best observation point is point no. 30. A slight overlap in MAE subgroups' scores with point 27 (2^{nd} best observation point) is observed.

6. Conclusion

The present study performs a comparative analysis of four different surrogate models for the variable density approach of seawater intrusion, in particular Gaussian process regression, binary regression decision tree method, ensemble tree learners, and the support vector machine regression models. Emphasis was given on the optimization of the examined techniques. To this end, a Bayesian optimization procedure of the surrogate models hyperparameters is used. The evaluation results indicate that the GPR method surpasses the other regressors in terms of the mean absolute error (MAE), the root mean square error (RMSE), and the coefficient of determination (R^2). It should be noted that GPR is significantly more time

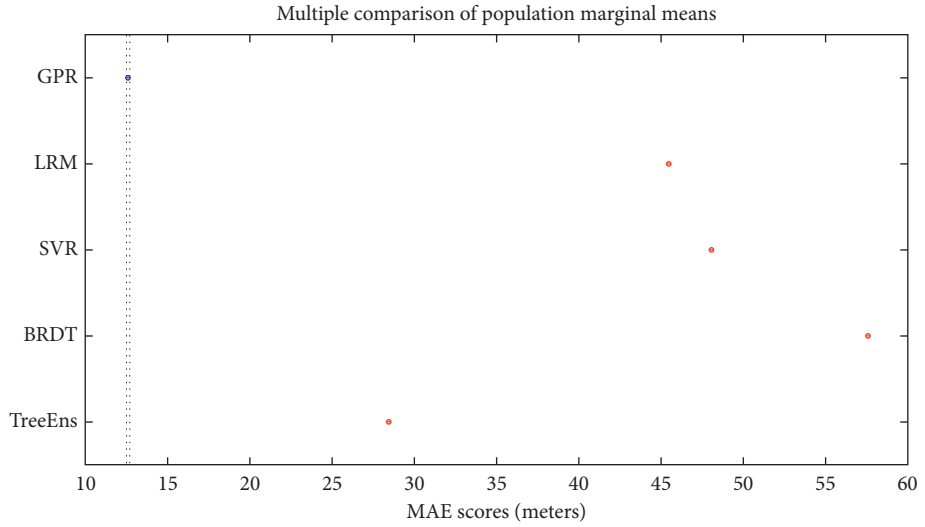

Figure 9: MAE scores for different types of regressors.

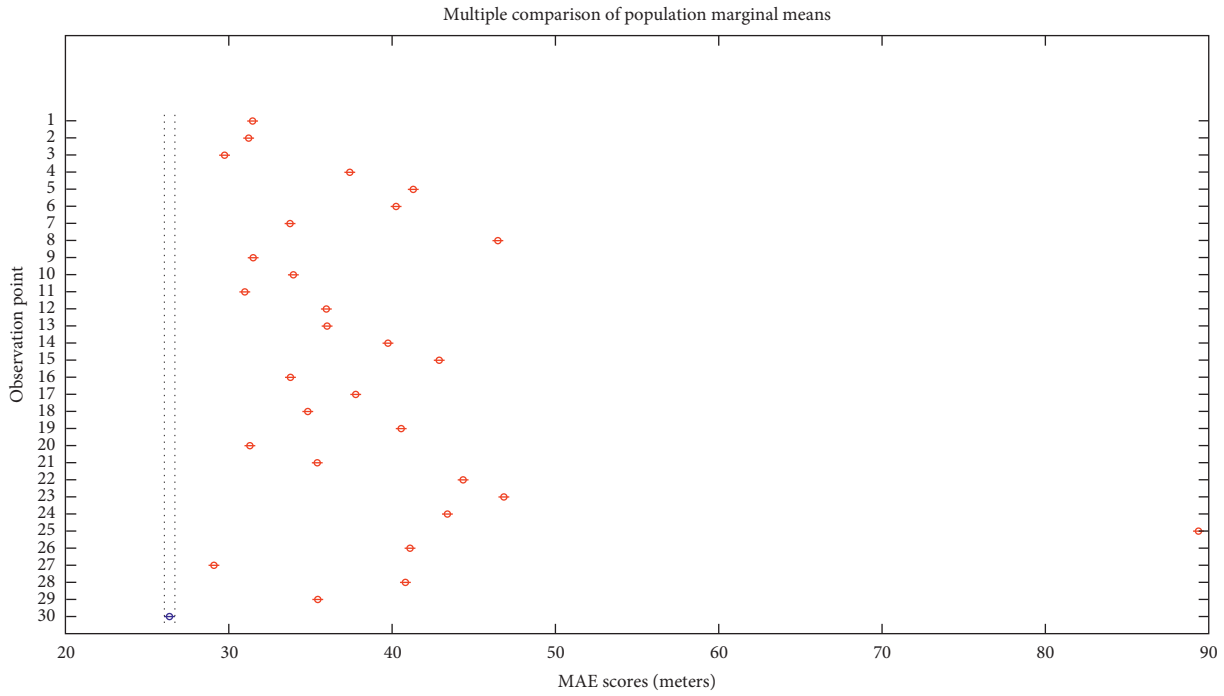

Figure 10: MAE scores for different observation points' positions.

consuming. In summary, the GPR method is a reliable and accurate surrogate model for SI and could be incorporated in a pumping optimization framework in coastal aquifers. Future research will focus on further scrutinizing the effectiveness of the GPR method for saltwater intrusion prediction, compared with other well-established surrogate models.

Data Availability

Data are not publicly available at this point because of intellectual property restrictions, but they can be provided to anyone interested on request. The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The research leading to these results has received funding from the European Union's Horizon 2020 Research and

Innovation Programme under grant agreement no. 740610 (STOP-IT project).

References

[1] C. T. Simmons, "Variable density groundwater flow: from current challenges to future possibilities," *Hydrogeology Journal*, vol. 13, no. 1, pp. 116–119, 2005.

[2] M. J. Asher, B. F. W. Croke, A. J. Jakeman, and L. J. M. Peeters, "A review of surrogate models and their application to groundwater modeling," *Water Resources Research*, vol. 51, no. 8, pp. 5957–5973, 2015.

[3] T. W. Simpson, J. D. Poplinski, P. N. Koch, and J. K. Allen, "Metamodels for computer-based engineering design: survey and recommendations," *Engineering with Computers*, vol. 17, no. 2, pp. 129–150, 2014.

[4] A. I. J. Forrester, A. Sóbester, and A. J. Keane, *Engineering Design via Surrogate Modelling–A Practical Guide*, Wiley, New York, NY, USA, 2008.

[5] S. Razavi, B. A. Tolson, and D. H. Burn, "Review of surrogate modelling in water resources," *Water Resources Research*, vol. 48, no. 7, 2012.

[6] M. M. Rajabi and H. Ketabchi, "Uncertainty-based simulation-optimization using Gaussian process emulation: application to coastal groundwater management," *Journal of Hydrology*, vol. 555, pp. 518–534, 2017.

[7] M. C. Kennedy and A. O'Hagan, "Bayesian calibration of computer models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 63, no. 3, pp. 425–464, 2001.

[8] C. E. Rasmussen, "Gaussian processes in machine learning," in *Advanced Lectures on Machine Learning*, pp. 63–71, Springer, Berlin, Heidelberg, Germany, 2004.

[9] P. Moonen and J. Allegrini, "Employing statistical model emulation as a surrogate for CFD," *Environmental Modelling and Software*, vol. 72, pp. 77–91, 2015.

[10] F. Liu and M. West, "A dynamic modelling strategy for Bayesian computer model emulation," *Bayesian Analysis*, vol. 4, no. 2, pp. 393–411, 2009.

[11] N. E. Owen, P. Challenor, P. P. Menon, and S. Bennani, "Comparison of surrogate-based uncertainty quantification methods for computationally expensive simulators," *SIAM/ASA Journal on Uncertainty Quantification*, vol. 5, no. 1, pp. 403–435, 2017.

[12] L. S. Bastos and A. O'Hagan, "Diagnostics for Gaussian process emulators," *Technometrics*, vol. 51, no. 4, pp. 425–438, 2009.

[13] V. Christelis, V. Bellos, and G. Tsakiris, "Employing surrogate modelling for the calibration of a 2D flood simulation model," in *Sustainable Hydraulics in the Era of Global Change: Proceedings of 4th IAHR Congress*, pp. 27–29, Liege, Belgium, 2016.

[14] V. Christelis and A. G. Hughes, "Metamodel-assisted analysis of an integrated model composition: an example using linked surface water–groundwater models," *Environmental Modelling and Software*, vol. 107, pp. 298–306, 2018.

[15] R. K. Bhattacharjya, B. Datta, and M. G. Satish, "Artificial neural networks approximation of density dependent saltwater intrusion process in coastal aquifers," *Journal of Hydrologic Engineering*, vol. 12, pp. 273–282, 2007.

[16] D. K. Roy and B. Datta, "Fuzzy C-mean clustering based inference system for saltwater intrusion processes prediction in coastal aquifers," *Water Resources Management*, vol. 31, no. 1, pp. 355–376, 2016.

[17] A. Lal and B. Datta, "Development and implementation of support vector machine regression surrogate models for predicting groundwater pumping-induced saltwater intrusion into coastal aquifers," *Water Resources Management*, pp. 1–15, 2018.

[18] A. Singh, "Optimization modelling for seawater intrusion management," *Journal of Hydrology*, vol. 508, pp. 43–52, 2014.

[19] R. K. Bhattacharjya and B. Datta, "ANN-GA-based model for multiple objective management of coastal aquifers," *Journal of Water Resources Planning and Management*, vol. 135, no. 5, pp. 314–322, 2009.

[20] S. V. N. Rao, V. Sreenivasulu, S. M. Bhallamudi, B. S. Thandaveswara, and K. P. Sudheer, "Planning groundwater development in coastal aquifers/planification du développement de la ressource en eau souterraine des aquifères côtiers," *Hydrological Sciences Journal*, vol. 49, no. 1, pp. 155–170, 2011.

[21] G. Kourakos and A. Mantoglou, "Pumping optimization of coastal aquifers using 3-d density models and approximations with neural networks," in *Proceedings of XVI International Conference on Computational Methods in Water Resources*, Copenhagen, Denmark, June 2006.

[22] G. Kourakos and A. Mantoglou, "Pumping optimization of coastal aquifers based on evolutionary algorithms and surrogate modular neural network models," *Advances in Water Resources*, vol. 32, no. 4, pp. 507–521, 2009.

[23] B. Ataie-Ashtiani, H. Ketabchi, and M. M. Rajabi, "Optimal management of freshwater lens in a small island using surrogate models and evolutionary algorithms," *Journal of Hydrologic Engineering*, vol. 19, no. 2, 2014.

[24] V. Christelis and A. Mantoglou, "Pumping optimization of coastal aquifers assisted by adaptive metamodelling methods and radial Basis functions," *Water Resources Management*, vol. 30, no. 15, pp. 5845–5859, 2016.

[25] V. Christelis, R. G. Regis, and A. Mantoglou, "Surrogate-based pumping optimization of coastal aquifers under limited computational budgets," *Journal of Hydroinformatics*, vol. 20, no. 1, pp. 164–176, 2017.

[26] J. Sreekanth and B. Datta, "Multi-objective management of saltwater intrusion in coastal aquifers using genetic programming and modular neural network based surrogate models," *Journal of Hydrology*, vol. 393, no. 3-4, pp. 245–256, 2010.

[27] J. Sreekanth and B. Datta, "Comparative evaluation of genetic programming and neural network as potential surrogate models for coastal aquifer management," *Water Resources Management*, vol. 25, no. 13, pp. 3201–3218, 2011.

[28] J. Sreekanth and B. Datta, ""Coupled simulation-optimization model for coastal aquifer management using genetic programming-based ensemble surrogate models and multiple-realization optimization"," *Water Resources Research*, vol. 47, no. 4, 2011.

[29] D. K. Roy and B. Datta, "Multivariate adaptive regression spline ensembles for management of multilayered coastal aquifers," *Journal of Hydrologic Engineering*, vol. 22, no. 9, 2017.

[30] D. K. Roy and B. Datta, "A surrogate based multi-objective management model to control saltwater intrusion in multi-layered coastal aquifer systems," *Civil Engineering and Environmental Systems*, vol. 34, no. 3-4, pp. 238–263, 2018.

[31] D. K. Roy and B. Datta, "Optimal management of groundwater extraction to control saltwater intrusion in multi-layered coastal aquifers using ensembles of adaptive neuro-fuzzy inference system," in *Proceedings of World*

Environmental and Water Resources Congress, pp. 139–150, Sacramento, CA, USA, May 2017.

[32] D. K. Roy and B. Datta, "A review of surrogate models and their ensembles to develop saltwater intrusion management strategies in coastal aquifers," *Earth Systems and Environment*, vol. 2, pp. 193–211, 2018.

[33] N. Stone, "Gaussian process emulators for uncertainty analysis in groundwater flow," , Doctoral dissertation, University of Nottingham2011.

[34] J. Zhang, W. Li, L. Zeng, and L. Wu, "An adaptive Gaussian process-based method for efficient Bayesian experimental design in groundwater contaminant source identification problems," *Water Resources Research*, vol. 52, no. 8, pp. 5971–5984, 2016.

[35] D. Crevillén-García, R. D. Wilkinson, A. A. Shah, and H. Power, "Gaussian process modelling for uncertainty quantification in convectively-enhanced dissolution processes in porous media," *Advances in water resources*, vol. 99, pp. 1–14, 2017.

[36] N. S. Raghavendra and P. C. Deka, "Multistep ahead groundwater level time-series forecasting using Gaussian Process Regression and ANFIS," in *Advanced Computing and Systems for Security, Advances in Intelligent Systems and Computing*, R. Chaki, A. Cortesi, K. Saeed, and N. Chaki, Eds., Vol. 396, Springer, New Delhi, India, 2016.

[37] D. K. Roy and B. Datta, "Influence of sea level rise on multiobjective management of saltwater intrusion in coastal aquifers," *Journal of Hydrologic Engineering*, vol. 23, no. 8, 2018.

[38] D. K. Roy and B. Datta, "Trained meta-models and evolutionary algorithm based multi-objective management of coastal aquifers under parameter uncertainty," *Journal of Hydroinformatics*, vol. 20, pp. 1247–1267, 2018.

[39] W. Guo and C. D. Langevin, "User's guide to SEWAT: a computer program for simulation of three-dimensional variable-density ground-water flow," Book 6, chapter A7, Techniques of Water–Resources Investigations of the U.S. Geological Survey, pp. 7–18, 2002.

[40] A. Mantoglou, M. Papantoniou, and P. Giannoulopoulos, "Management of coastal aquifers based on nonlinear optimization and evolutionary algorithms," *Journal of Hydrology*, vol. 297, no. 1, pp. 209–228, 2004.

[41] A. Mantoglou and M. Papantoniou, "Optimal design of pumping networks in coastal aquifers using sharp interface models," *Journal of Hydrology*, vol. 361, pp. 52–63, 2008.

[42] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[43] C. B. Do, *Gaussian Processes*, H. Lee, Ed., Stanford University, Stanford, CA, USA, 2008.

[44] L. Breiman, *Classification and Regression Trees*, Routledge, Abingdon, UK, 2017.

[45] D. Basak, S. Pal, and D. C. Patranabis, "Support vector regression," *Neural Information Processing-Letters and Reviews*, vol. 11, no. 10, pp. 203–224, 2007.

[46] E. Protopapadakis, A. Voulodimos, and N. Doulamis, "An investigation on multi-objective optimization of feedforward neural network topology," in *Proceedings of 2017 8th International Conference on Information, Intelligence, Systems & Applications (IISA)*, pp. 1–6, Larnaca, Cyprus, August 2017.

[47] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep learning for computer vision: a brief review," *Computational Intelligence and Neuroscience*, vol. 2018, Article ID 7068349, 13 pages, 2018.

[48] A. Doulamis, "Adaptable neural networks for objects' tracking Re-initialization," in *Artificial Neural Networks–ICANN 2009, Lecture Notes in Computer Science*, C. Alippi, M. Polycarpou, C. Panayiotou, and G. Ellinas, Eds., Vol. 5769, Springer, Berlin, Heidelberg, 2009.

[49] A. D. Bull, "Convergence rates of efficient global optimization algorithms," *Journal of Machine Learning Research*, vol. 12, pp. 2879–2904, 2011.

[50] M. A. Gelbart, J. Snoek, and R. P. Adams, *Bayesian Optimization with Unknown Constraints*, arXiv preprint arXiv: 1403.5607, 2014.

[51] J. Snoek, H. Larochelle, and R. P. Adams, "Practical Bayesian optimization of machine learning algorithms," in *Advances in Neural Information Processing Systems*, pp. 2951–2959, MIT Press, Cambridge, MA, USA, 2012.

[52] E. Protopapadakis, A. Voulodimos, A. Doulamis, S. Camarinopoulos, N. Doulamis, and G. Miaoulis, "Dance pose identification from motion capture data: a comparison of classifiers," *Technologies*, vol. 6, no. 1, p. 31, 2018.

[53] V. Christelis and A. Mantoglou, "Pumping optimization of coastal aquifers using seawater intrusion models of variable-fidelity and evolutionary algorithms," *Water Resources Management*, pp. 1–14, 2018.

[54] G. S. Atsalakis, E. E. Protopapadakis, and K. P. Valavanis, "Stock trend forecasting in turbulent market periods using neuro-fuzzy systems," *Operational Research*, vol. 16, no. 2, pp. 245–269, 2015.