Earth Systems and Environment (2018) 2:193–211
https://doi.org/10.1007/s41748-018-0069-3

REVIEW ARTICLE

A Review of Surrogate Models and Their Ensembles to Develop Saltwater Intrusion Management Strategies in Coastal Aquifers

Dilip Kumar Roy[1] · Bithin Datta[1]

Received: 11 March 2018 / Accepted: 6 August 2018 / Published online: 21 August 2018

Abstract

The complex physical processes in a typical coastal aquifer system with transient inputs to numerical simulation models (NSM) result in substantial computational burden in a coupled simulation–optimization (S/O) approach. In such situations, an approximate emulator of the complex physical processes provides a computationally efficient alternative to the NSM. The reliability of these surrogate models (SM) within the coupled S/O approach depends on how accurately they capture the trend of the underlying physical processes. Moreover, these SMs are often associated with prediction uncertainties, which hinder optimality of the solution of the coupled S/O methodology. In this review article, we summarize ensemble approach of combining data-driven SMs to address this prediction uncertainty. Different techniques of ensemble formation as well as their relative advantages and disadvantages are also discussed. Although a wide range of data-driven SMs have been used to approximate associated physical processes of coastal aquifers, the use of ensemble SMs is quite limited. Moreover, these ensemble-based modelling approaches are based on manipulating the training data set, i.e., using different realizations of training data set to train individual SMs within the ensemble. Although ensemble formation by combining multiple SMs based on different algorithms can be found in other application domains, the application of ensemble SMs in the prediction of saltwater intrusion processes has not been developed yet. In addition, more advanced ensemble surrogate-modelling approaches are yet to be established in the context of developing regional scale saltwater intrusion management models.

Keywords Coastal aquifer · Ensemble · Surrogate model · Coupled simulation–optimization · Saltwater intrusion · Management model

1 Introduction

Groundwater flow and transport processes in a typical coastal aquifer system are dynamic, non-linear, and complex in nature. Simulation of these complex physical processes to predict future scenarios of saltwater intrusion requires three-dimensional (3D) density-dependent coupled flow and solute transport numerical simulation models (NSM). These simulation models need to incorporate real-world aquifer processes in the form of initial and boundary conditions, model parameters, and transient model inputs. Density-dependent saltwater intrusion problems are typically solved by utilizing finite-difference approximation for the variable-density groundwater flow and transport equations, e.g., SEAWAT (Langevin et al. 2007) or by a finite-element-based density-dependent coupled flow and salt transport equations such as FEMWATER (Lin et al. 1997) and HydroGeoSphere (Therrien et al. 2010). HydroGeoSphere provides an additional feature of coupling surface and subsurface domain of the model boundary. Complexity of the associated physical processes and the simulation model itself determines the number of model parameters and inputs, which significantly increases computational time required to run the model. A typical study area consisting of a portion of a coastal aquifer with spatial pumping and saltwater concentration monitoring locations is shown in Fig. 1 (after Roy and Datta (2017c)).

Model complexity and longer runtimes constrain the utilization of NSMs in applications, where repetitive use of such models is necessary, e.g., in a coupled simulation–optimization (S/O) approach to develop a regional scale saltwater

✉ Dilip Kumar Roy
dilip.roy@my.jcu.edu.au

Bithin Datta
bithin.datta@jcu.edu.au

[1] Discipline of Civil Engineering, College of Science and Engineering, James Cook University, Townsville, QLD 4811, Australia

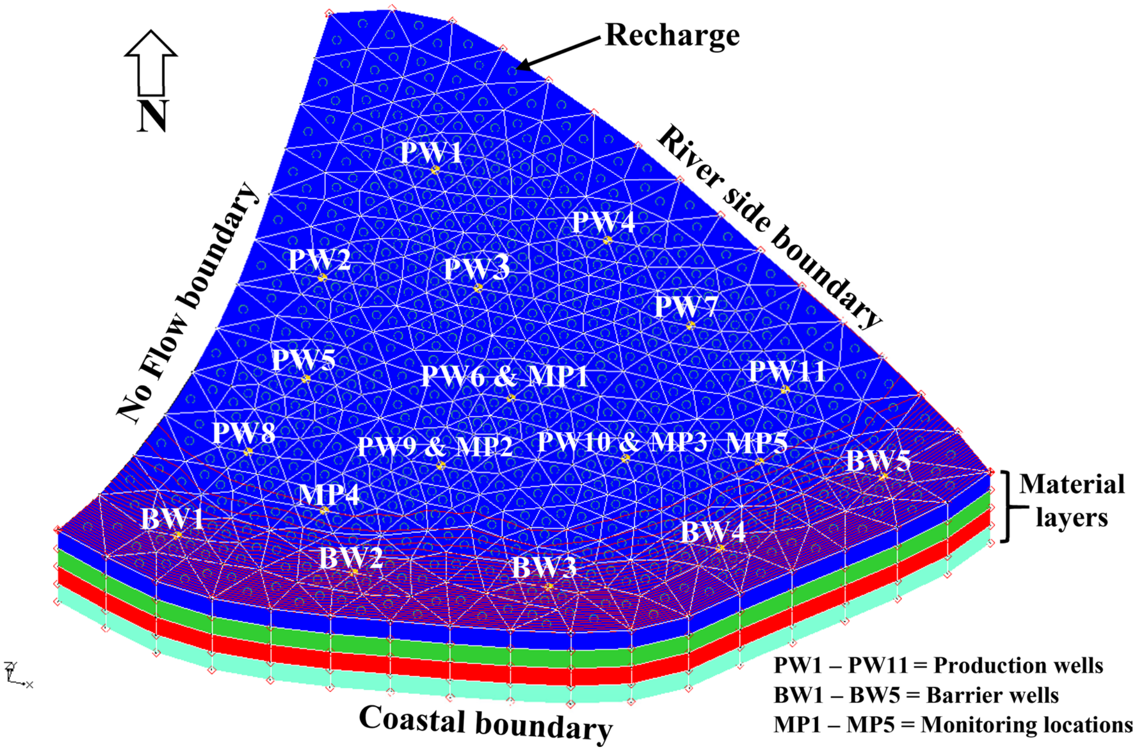

Fig. 1 Illustrative study area showing model boundaries and well locations (after Roy and Datta (2017c))

intrusion management model. In a coupled S/O approach, the optimization algorithm calls numerical models' several thousand times to derive an optimal solution. Therefore, the use of NSM in a coupled S/O approach is constrained by huge computational burden because of this multiple calls of the complex NSM within the optimization framework (Mantoglou and Papantoniou 2008; Dhar and Datta 2009). For instance, a coupled S/O approach for a small 3D coastal aquifer may require as large as 30-day computer run time (Dhar and Datta 2009). Therefore, replacing the original NSM by a reasonably accurate surrogate model (SM), which is trained and tested using solution results obtained from an NSM, is a promising approach that has been used to achieve computational efficiency in the design and optimization of computationally intensive problems (Goel et al. 2007). In a coupled S/O approach, SMs serve as computationally efficient proxies or emulators of the complex physical processes to replace time consuming and memory intensive NSMs (Blanning 1975). The previous studies of saltwater intrusion management modelling utilized different SMs as computationally efficient substitutes of complex NSMs in optimization formulations. The most commonly used SMs are based on artificial neural network (ANN) (Bhattacharjya and Datta 2005; Bhattacharjya et al. 2007; Kourakos and Mantoglou 2009; Sreekanth and Datta 2011c), genetic programming (GP) (Sreekanth and Datta 2010, 2011a, b), multivariate adaptive regression splines (MARS) (Roy and Datta 2017c, 2018b), and adaptive neuro fuzzy inference system (ANFIS) (Roy and Datta 2017d).

However, these SMs are always associated with certain amount of prediction uncertainty that may propagate to the optimization routine, and may affect optimality of solutions (optimal groundwater extraction strategies). Rajabi and Ketabchi (2017) proposed a probabilistic emulator, gaussian process regression (GPR) to address epistemic and aleatory uncertainty in developing management strategy in coastal groundwater for single-objective problem settings. Although using GPR alone to handle uncertainty in surrogate modelling is well justified (Rajabi and Ketabchi 2017; Sun et al. 2014), investigating ensembles of GPR surrogates to further reduce the prediction uncertainty is still important. Moreover, a standalone SM often fails to extract true trends in the data from the total decision space. An ensemble of SMs can be an effective tool for extracting accurate trends in the data, and may protect against single SM having inadequate generalization capabilities (Goel et al. 2007). The ensemble approach essentially serves as an effective way of accounting for model uncertainty (Sreekanth and Datta 2011b). Ensemble is an approach to integrate two or more similar or unalike algorithms or base learners, often called SMs or meta-models. The idea is to develop a more reliable and robust system that incorporates each individual learners' unique feature to predict the future scenario. Each individual member of the ensemble has different input–output mapping functions based on the understanding of the associated physical processes of the system. Therefore, these individual SMs are supposed to provide varied predictions on the response variable based on their own mapping functions. The final prediction obtained from the ensemble is likely to be less biased and more robust, reliable, and accurate than any of the individual members of the ensemble. However, care should be taken to maintain adequate accuracy and sufficient diversity of the individual models.

Accuracy of the individual SMs is achieved using the right choice of models and by adjusting the associated parameters of the selected model. Diversity of the individual SMs within the ensemble is maintained by integrating a wide range of SMs or using a single SM in which training data realizations can be obtained by any suitable random sampling strategy. Combination of multiple SMs based on different algorithms to construct an ensemble can be found in other water resources' management problems (Jovanović et al. 2015; Melin et al. 2012). In saltwater intrusion management problems, the diversity of individual members of the ensemble is obtained through different realizations of training data set either by nonparametric bootstrap sampling (Sreekanth and Datta 2011b) or by random sampling without replacement (Roy and Datta 2017c, d) technique.

Ensemble-based surrogate-modelling approach is no doubt associated with an increased computational cost (Sun et al. 2014) as the ensemble is formed by integrating two or more individual models. Each individual member of the ensemble is often coupled externally to the optimization algorithm as a binding constraint to derive a meaningful global Pareto optimal groundwater extraction strategy to control saltwater intrusion (Roy and Datta 2017c, d; Sreekanth and Datta 2011b). For multi-objective formulations, computational time requirement becomes an issue, where the problem has reasonably higher numbers of decision variables. In addition, an adequate number of SMs within the ensemble have to be used as binding constraints to calculate the output within the optimization formulation. Such situations demand further efficiency in computation time that can be achieved by utilizing a parallel-processing strategy (Ketabchi and Ataie-Ashtiani 2015a) or any other high-performing computing techniques. In parallel-processing strategy, the objective functions and constraints of the multi-objective saltwater intrusion problem are distributed to a parallel pool of workers to reduce the computational burden. This can be achieved by using the physical cores of a single PC or by a network of PCs. The previous studies on saltwater intrusion management problems utilized physical cores of a standard PC for parallel computation of objective functions and all constraints of the management problem (Ketabchi and Ataie-Ashtiani 2015a; Roy and Datta 2017b, c, d).

Recent literature addresses several criteria to analyze ensemble surrogate-modelling approaches in coupled S/O-based saltwater intrusion management problems. These include quantifying computational gain, measuring nearness to represent original simulation model by capturing the relevant processes accurately, and addressing uncertainty of prediction induced by SMs. Computational efficiency is one of the most important issues that can be assessed by quantifying the number of input–output patterns (obtained through running complex simulation model) required to train and validate the individual SMs within the ensemble. Gaining computational efficiency can also be assessed by comparing efforts required to direct linking of simulation model vs SMs within the optimization algorithm. In this review article, commonly used data-driven SMs used to approximate density-dependent coupled flow and salt transport processes in coastal aquifers are discussed. In particular, ensemble of individual SMs to achieve more reliable and robust prediction of the underlying processes as well as to address prediction uncertainty is also highlighted. Efforts to reduce the computational burden of the ensemble-based surrogate-modelling approaches in saltwater intrusion management problems are also discussed.

2 Data-Driven SMs

Surrogate models are intended to replace complex NSMs within a coupled S/O approach to reduce the computational burden during the search process of global optimal solution in saltwater intrusion management problems for coastal aquifers. SMs are especially advantageous in situations, where the use of NSMs is infeasible, and where an approximate and reliable representation of the NSM is sufficient. In general, a typical data-driven SM can be represented by

$$Y_{\mathrm{output}}(x) = y_{\mathrm{output}}(x) + \varepsilon, \tag{1}$$

where $Y_{\mathrm{output}}(x)$ represents output from the NSM at point x, $y_{\mathrm{output}}(x)$ denotes predicted output from the SM, and ε is an error term that represents error between NSM outputs and SM predictions. The aim is to minimize this error term as much as possible by selecting appropriate SMs for specific problems at hand and by obtaining optimum sets of parameters for the selected SMs. In this section, commonly used SMs and their application in saltwater intrusion management problems are discussed.

2.1 Adaptive Neuro Fuzzy Inference System (ANFIS)

An adaptive neuro fuzzy inference system (ANFIS) is a multiple layer adaptive fuzzy inference system (FIS) that incorporates the concepts of both an ANN and fuzzy set theory (Jang 1993). ANFIS has attained considerable attention in recent years as an effective computational tool for application in multi-dimensional fields (Jang et al. 1997).

ANFIS-based SMs are able to map non-linear relationships between the predictor and response variables, and are better suited for modelling non-linear processes (Sugeno and Yasukawa 1993; Takagi and Sugeno 1985). A Sugeno-type ANFIS is simple in model structure and has better learning capabilities compared to other types of ANFIS structures (Jang et al. 1997). Fuzzy if–then rules for an ANFIS structure based on the principle of first order Sugeno FIS can be expressed as

$$\text{Rule 1}: \text{ If } \alpha \text{ is } P_1 \text{ and } \beta \text{ is } Q_1 \text{ then } f_1 = p_1\alpha + q_1\beta + r_1 \tag{2}$$

$$\text{Rule1}: \text{ If } \alpha \text{ is } P_2 \text{ and } \beta \text{ is } Q_2 \text{ then } f_2 = p_2\alpha + q_2\beta + r_2, \tag{3}$$

where α and β are two inputs.

A typical ANFIS structure has five layers. These are a fuzzy layer, a product layer, a normalized layer, a defuzzification layer, and a total output layer. Each layer contains several nodes described by the corresponding node functions. Inputs are fuzzified in layer 1 whereas outputs are defuzzified in layer 4. Parameters in layer 1 are referred to as premise parameters whereas the parameters in layer 4 are referred to as consequent parameters of the rules. In layer 5, the single node computes the final output as the summation of all incoming signals. Interested readers are directed to Jang et al. (1997) for detailed description of each of these layers and the associated node functions.

2.2 Artificial Neural Network (ANN)

An artificial neural network (ANN) approximates physical processes by mapping predictor–response relationships through a learning approach that mimics human-reasoning process. The network consists of several synaptically connected artificial neurons. A set of optimal weights for these neurons are obtained through learning process. This ability to learn from prior knowledge using predictor–response training patterns makes it possible to apply ANN-based SMs for prediction of saltwater intrusion processes (Bhattacharjya et al. 2007) as well as for developing saltwater intrusion management models (Bhattacharjya and Datta 2009; Sreekanth and Datta 2011c) in coastal aquifers. The predictor–response mapping is encoded in a properly trained and validated ANN structure that can be used later to predict future scenarios with an entirely unseen set of testing data. Feedforward backpropagation neural network (FFNN) and modular neural network (MNN) are two of the commonly used ANN models used to develop saltwater intrusion management models in coastal aquifers (Sreekanth and Datta 2010; Bhattacharjya and Datta 2009). A typical ANN architecture consists of an input layer, an output layer, and one or more hidden layers of neurons. The neurons in subsequent layers are entirely interconnected through adaptable weighted connections. During training, the backpropagation algorithm use gradient descent technique to minimize the error between the actual responses and the ANN predictions. However, utilizing a weight matrix for training and larger complexity of the "black-box" ANN model structure are the main disadvantages of ANN-based modelling approaches (Rezania et al. 2008).

2.3 Evolutionary Polynomial Regression (EPR)

Evolutionary polynomial regression (EPR) (Giustolisi and Savic 2006) is a hybrid data-driven-based SM that combines the features of traditional regression techniques with genetic programming and symbolic regression approach. Unlike "black-box" data-driven models, e.g., ANN, the EPR-based SMs are simple mathematical formulations. Despite simplicity in the model formulation, EPR-based models demonstrated better performance when compared to GP and ANN-based modelling techniques (El-Baroudy et al. 2010). EPR was used as a computationally efficient substitute of the complex NSM within the coupled S/O methodology to control saltwater intrusion in coastal aquifers (Hussain et al. 2015). The study utilized three management scenarios comprising of simultaneous use of water abstraction and aquifer recharge for controlling saltwater intrusion in a coastal aquifer system. The performance of each management scenario was evaluated through a multi-objective formulation of the management problem. The two objectives considered were minimizing the cost of management and minimizing the salinity levels in the aquifer.

2.4 Fuzzy Inference System (FIS)

Fuzzy inference system (FIS) (Jang et al. 1997) based on fuzzy set theory has received considerable attention, and it is recognized as a successful computing framework due to its capability of application in multi-dimensional fields. FIS is capable of capturing non-linear relationships between predictor and response variables, and is an effective tool to model non-linear systems (Sugeno and Yasukawa 1993; Takagi and Sugeno 1985). A Sugeno-type FIS, also known as Takagi–Sugeno–Kang model (Sugeno 1985), is ideal for this non-linear mapping of predictor–response relationships. The computational framework of a Sugeno FIS follows the theory of fuzzy logic that combines fuzzy set theory, fuzzy if–then rules, and fuzzy reasoning. The basic structure of the FIS is composed of three components: (1) a rule base consisting of fuzzy if–then rules; (2) a database that determines the type, size, and number of membership functions (MF) used in the fuzzy rules; and (3) a reasoning mechanism that accomplishes the inference process (Jang et al. 1997). FISs can be used for non-linear mapping of predictor and response spaces utilizing a number of these fuzzy if–then rules.

2.5 Gaussian Process Regression (GPR)

Gaussian process regression (GPR) (Rasmussen and Williams 2005) is a flexible, nonparametric, and probability-based stochastic approach used to build non-linear approximation models that are able to provide probabilistic information on prediction. In GPR approach, the way a machine learns is formulated within a Bayesian framework in which model variables are considered as random variables drawn from a Gaussian distribution (Bazi et al. 2012). GPR is a nonparametric modelling approach, i.e., no assumption is made about the shape of the function to estimate. GPR provides a "principled, practical, and probabilistic approach to learning in kernel machines" (Rasmussen and Williams 2005). As a popular artificial intelligence tool, GPR has been successfully applied in many engineering problems (Forrester et al. 2008). GPR provides a flexible Bayesian framework to identify a non-linear predictor–response mapping from a set of training data (Sun et al. 2014). In GPR-based approach, the response, Y, is related to the predictors, $X(k)$, such that $Y = f(X(k)) + \varepsilon$, where ε is a Gaussian noise with variance σ_n^2 (Bishop 2006). A Gaussian process is entirely indicated by its mean and covariance functions. For a real function $f(x)$, the mean and covariance functions are defined as

$$\text{Mean function:} m\left(x_i\right) = E\left[f\left(x_i\right)\right] \tag{4}$$

$$\text{Covariance function : } k\left(x_i, x_j\right) = E\left[\left(f\left(x_i\right) - m\left(x_i\right)\right) \times \left(f\left(x_j\right) - m\left(x_j\right)\right)\right]. \tag{5}$$

Finally, the Gaussian process can be written as

$$f(x) \sim gp\left(m\left(x_i\right), k\left(x_i, x_j\right)\right). \tag{6}$$

The mean function provides a description of the expected value of the function at any particular point within the input space. On the other hand, covariance function defines proximity (nearness) or resemblance (similarity) between the predictor values x_i and the response (target) value y_i (Rasmussen and Williams 2005). The covariance function is considered as the most important and influential element of GPR models. The parameters related to the mean and covariance functions are called free parameters or hyperparameters. The properties of predictive probability distribution are defined by these hyperparameters, the values of which are obtained by maximizing log-likelihood function of the training data (Rasmussen and Williams 2005).

In coastal groundwater management problems, Rajabi and Ketabchi (2017) proposed the use of GPR-based SMs to replace the complex NSM to achieve computational efficiency of an uncertainty-based single-objective optimization formulation. The potential applicability of this probability and uncertainty-based SM to achieve computational efficiency and reliability in multi-objective saltwater intrusion management problems in coastal aquifers is yet to be investigated.

2.6 Genetic Programming

Genetic programming (GP) (Koza 1994) models are genetic algorithm (GA)-based computer programs evolved from Darwinian principle of natural selection. GPs are intended to perform specific tasks through a search technique that applies GA to computer programming (Koza 1994). The working principle of GP is similar to GA as GP also starts with an initial population that compounds the randomly generated chromosomes (Wang et al. 2009a). GPs are used to obtain the best-fit computer programs that can be employed to predict future scenarios when presented with a set of predictors. GPs have successfully been utilized as computationally efficient substitutes of complex NSMs in a coupled S/O approach for developing saltwater intrusion management problems (Sreekanth and Datta 2011b).

2.7 Multivariate Adaptive Regression Spline (MARS)

Multivariate adaptive regression spline (MARS) (Friedman 1991) utilizes an adaptive search space methodology in which the search space is modified through integrating both a forward and a backward stepwise procedure. Initially, MARS builds a relatively complex model through incorporating modeller-specified number of basis functions, and using all the predictors. Later, the optimum number of basis functions and the most influential predictors are selected parsimoniously to eliminate irrelevant predictors in determining the response (SPM 2016). This backward step keeps the developed MARS model as simple as is required, and rules out the possibility of model overfitting. MARS is a nonparametric adaptive regression technique that is considered to be a rapid, flexible, and accurate artificial intelligence technique suitable for predicting both continuous and binary responses (SPM 2016). Learning process of MARS is associated with dividing the entire solution space into different intervals of predictors. Afterwards, the MARS-based prediction models are developed by fitting individual splines or basis functions to each interval (Bera et al. 2006). MARSs are nonparametric and adaptive emulators in which no prior assumption is made for the functional relationship between the predictors and responses, rather this relationship is built in an adaptive manner (Friedman 1991). A set of coefficients and basis functions determined by the training data is used to develop this functional relationship. MARS produces simple and easy-to-interpret models by capturing the predictor–response mapping from a high-dimensional data pattern (Zhang and Goh 2016). MARS is parsimonious in selecting the most influential predictors based on the relative importance of predictors in determining the response.

The predictor–response mapping of a typical MARS model can be expressed as (Roy and Datta 2017c)

$$\begin{aligned}\mathrm{BF}_i(X) &= \max\left(0, X_j - p\right) \text{ OR}\\ &= \max\left(0, p - X_j\right)\end{aligned} \tag{7}$$

$$Y = f(X) = q \pm r_\mathrm{k} \times \mathrm{BF}_i(X), \tag{8}$$

where i and j are the indices for basis functions and input variables (groundwater extraction), respectively.

2.8 Radial Basis Function (RBF)

Radial basis function (RBF) uses linear combinations of m radially symmetric functions $h(x)$ to approximate response functions as

$$y_i = \sum_{i=1}^{m} w_i h_i(x) + \varepsilon_i, \tag{9}$$

where w is the coefficient of the linear combinations, h is the radial basis functions, and ε_i is independent errors with equal variance σ^2. Radial basis functions are a special class of functions, whose main feature is that the response decreases (or increases) monotonically with distance from a central point. The centre, the distance scale, and the precise shape of the radial function are the parameters of the model.

A typical radial function is the Gaussian, which is (in the case of a scalar input):

$$h(x) = \exp\left(-\frac{(x-c)^2}{r^2}\right). \tag{10}$$

An RBF model can be expressed as

$$y = Hw + \varepsilon V(\varepsilon) = \sigma^2. \tag{11}$$

3 Application of Different Data-Driven SMs in Saltwater Intrusion Management

This section provides a comparative evaluation of the performances among different SMs used in developing saltwater intrusion management models in coastal aquifers. ANNs are the most widely used emulators of complex physical processes to reduce computational complexity of a coupled S/O approach. ANN-based emulators were first introduced to derive optimal pumping management strategies within a coupled S/O approach in which the ANN-based emulator was coupled to a non-linear optimization algorithm (Rogers et al. 1995). Yan and Minsker (2006) proposed a dynamic modelling approach using ANN within a GA-based optimization algorithm and demonstrated that the proposed approach was able to save 85–90% of the simulation model calls while maintaining adequate accuracy of the optimal solutions for a single-objective optimization formulation. Bhattacharjya and Datta (2009) applied ANN as an approximate simulator of the density-dependent coupled flow and salt transport processes in a coastal aquifer. In their study, an ANN-based emulator was coupled with GA within a coupled S/O approach for a multiple objective problem setting. However, in situations, where the number of decision variables is quite large, the resulting ANN structure might be very complex and would be difficult to train. To solve this issue, Kourakos and Mantoglou (2009) proposed modular neural network (MNN) as an approximate emulator of the NSM in a pumping optimization problem in coastal aquifers. Dhar and Datta (2009) and Sreekanth and Datta (2011c) also demonstrated the capability of ANN-based emulators to achieve computational efficiency in achieving non-dominated Pareto optimal front for multiple objective saltwater intrusion management of coastal aquifers. However, ANN models suffer from different limitations despite achieving computational efficiency in a coupled S/O approach and making the surrogate-based coupled S/O approach for saltwater intrusion management problem feasible. These drawbacks of ANN models include proneness to premature convergence in local minima, the "black-box" nature of the models, higher computational burden, and susceptibility to model overfitting (Holman et al. 2014). In addition, ANN models have stability issues for smaller number of input–output training data sets (Hsieh and Tang 1998). Moreover, the architecture of the ANN has to be fixed a priori, i.e., the number of hidden layers and nodes need to be chosen before ANN training starts with the training data set (Hussain et al. 2015).

Considering the drawbacks of ANN models, researchers have been continuing to search for a better SM that can provide a reliable and global Pareto optimal solution within the coupled S/O approach. Consequently, GP was found to be more reliable, flexible, and accurate to approximate saltwater intrusion processes prediction (Sreekanth and Datta 2011a). A couple of research works demonstrated the potential applicability GP to develop multiple objective saltwater intrusion management problems in coastal aquifers. These include demonstration of the superiority of GP models over ANN (Sreekanth and Datta 2011a) and MNN (Sreekanth and Datta 2010)-based SMs. GPs are able to identify the relative importance of input variables in determining the output variable. This parsimonious selection process of input variables helps develop better and efficient SMs (Sreekanth and Datta 2010). GP, an explicit mathematical formulation (Shiri and

Kişi 2011), produces simple regression models (Sreekanth and Datta 2011a) that can be coupled easily within an optimization algorithm to achieve computational efficiency in coupled S/O methodology. However, GP requires extensive training time for evaluating millions of model structures before finding the optimal structure (Sreekanth and Datta 2011a). Besides, GP suffers from being trapped in local minima (Pillay 2004). In searching for the best-fit expression of a m-dimensional function F of input X as a set of m input parameters, a set of constants represented by θ are generated as non-adjustable constants and do not essentially represent optimal values. As a consequence, the so-called global search process of GP can be unable to develop good structures of F for predicting the output, Y represented by $Y = F(X, \theta)$ (Hussain et al. 2015).

To overcome some of the limitations of ANN and GP, Giustolisi and Savic (2006) proposed a hybrid data-driven approach EPR that integrates traditional numerical regression techniques, GP, and symbolic regression techniques in a general framework. Hussain et al. (2015) utilized EPR-based SMs to approximate the physical processes of non-linear and computationally complex coastal aquifer systems subjected to seawater intrusion. A multi-objective saltwater intrusion management model was developed by integrating the developed EPR model within a coupled S/O approach to evaluate the performance of different combinations of hydraulic barriers in controlling saltwater intrusion. They demonstrated a satisfactory non-linear mapping capability of EPR models to approximate complex physical processes of coastal aquifers and to develop multi-objective saltwater intrusion management models. However, in general, polynomial regression has stability problems when the polynomial order is high for polynomial fits. In addition, individual observations of the training data sets can have an unexpected influence on remote parts of the curve in polynomial regression (Green and Silverman 1993).

RBF is quite simple in formulation (Sóbester et al. 2014), and easy to implement in any number of dimensions with a reasonable accuracy for certain types of radial functions (Piret 2007). Christelis et al. (2017) used a cubic RBF model augmented with a linear polynomial tail as an SM for a single-objective pumping optimization problem of coastal aquifers under limited computational budgets. They showed that cubic RBF-based optimization provides better sample means when compared to direct optimization with HydroGeoSphere under specified computational budgets. Christelis and Mantoglou (2016) also utilized a cubic RBF as a computationally efficient substitute of HydroGeoSphere to develop a single-objective pumping optimization problem in coastal aquifers. The stability of RBF-based surrogate-modelling approach is an issue that relates to the distances of the points in the training data set. However, the computational cost is not critical particularly for nonparametric formulations, e.g., thin plate spline or cubic RBF.

Fuzzy logic-based approximate simulators, FIS are useful tools for non-linear mapping of predictor–response relationships of complex physical processes (Jang 1993). This non-linear mapping is performed by utilizing fuzzy IF–THEN rules that incorporate human expert knowledge or common sense to address the problem (Cherkassky 1998). FISs provide accurate predictions with less computational requirements (MATLAB 2017) and are suitable for predicting multiple outputs using a single-global FIS architecture for multiple output problems (Roy and Datta 2017a). A Sugeno-type FIS was successfully utilized to approximate density-dependent coupled flow and solute transport processes in a coastal aquifer system (Roy and Datta 2017a). Based on the computational efficiency and the accuracy of prediction, the authors recommend this FIS model as a good candidate for linking within a coupled S/O approach to develop regional scale saltwater intrusion management model. Later, Roy and Datta (2018a) investigated the quality of the Pareto optimal groundwater extraction patterns produced by the FIS model within the optimization routine.

Adaptive SMs have recently been used to develop saltwater intrusion management models in coastal aquifers to prescribe optimal groundwater extraction patterns to control saltwater intrusion. ANFIS (Jang et al. 1997)-based surrogate-modelling approach was used to develop optimal groundwater pumping strategy in coastal aquifers using a coupled S/O approach in which ANFIS model replaces the complex NSM (Roy and Datta 2017d). The application of another adaptive surrogate model, MARS (Friedman 1991), can also be found in the recent literature of saltwater intrusion management problem in a multi-layered coastal aquifer system (Roy and Datta 2017c, 2018b). An ensemble of MARS (En-MARS) was proposed by Roy and Datta (2017c) to reduce the prediction uncertainty of surrogate modelling. While providing a very good prediction accuracy, MARS and En-MARS models are also very efficient in terms of computational requirement, i.e., in searching for global Pareto optimal solution after 3990,401 function evaluations, the optimization routine using En-MARS took only 76 min (Roy and Datta 2017c). Although the computational time required to obtain the global Pareto optimal solution using En-ANFIS model is not provided in Roy and Datta (2017d), the complex nature of ANFIS-based modelling approaches suggests a relatively better prediction accuracy with the cost of additional computational requirements.

Gaussian process-based SM has recently been applied in a coastal groundwater management problem utilizing an uncertainty-based coupled S/O approach (Rajabi and Ketabchi 2017). The authors demonstrated that GPR-based emulators significantly reduce the computational time of the coupled S/O approach while providing a reliable and acceptable prediction accuracy with no bias and with low statistical dispersion.

4 Computational Effort Needed to Generate the Number of Training Patterns Required to Train Data-Driven SMs

Surrogate models require a sufficiently large set of predictor–response data sets usually obtained either from the real-field data or from a physically-based NSM. The latter is usually used in situations, where obtaining real-field data is difficult, i.e., in developing a coastal groundwater pumping management strategy utilizing a coupled S/O methodology. In this methodology, input data set comprises spatially and temporally variable groundwater extraction patterns obtained from a set of production bores and barrier extraction wells or recharge wells. However, it may be quite time intensive and computationally inefficient or even unrealistic to generate hundreds to thousands of predictor–response patterns by simulating the original model to develop a reasonably accurate SM (Sreekanth and Datta 2015). These large numbers of training patterns are associated with enormous computational requirement especially for simulating 3D density-dependent coupled flow and solute transport processes in coastal aquifers with spatially and temporally variable groundwater extraction patterns as inputs to the NSM.

The required number of predictor–response array depends on the number of predictors as well as on the considered SM. Therefore, for a large-dimensional problem, the right choice of SMs is the crucial first step of developing a saltwater intrusion management model. After deciding on the appropriate SM for a specific problem, the usual practice in saltwater intrusion modelling is to generate a sufficiently large number of transient input patterns by any suitable sampling strategy. For instance, LHS (Pebesma and Heuvelink 1999) technique was utilized to generate spatially and temporarily varying groundwater extraction patterns within the practical limit of 0–1300 m^3/day from a set of production bores and barrier extraction wells (Sreekanth and Datta 2011c; Roy and Datta 2017a, b, c; Bhattacharjya and Datta 2009). These input patterns were then fed to the NSM to obtain the corresponding saltwater concentrations as outputs. A sufficiently large number of input–output data pairs are usually generated, and the training of SM starts with a certain number of such patterns and adding the input–output patterns incrementally until no significant improvement is achieved by the addition of further data pairs (Sreekanth and Datta 2011c; Roy and Datta 2017a). Although this technique demands a significant computational time, the SMs thus obtained are able to capture the nearly true trend of the input–output relationships from the entire decision space. These SMs when coupled to an optimization algorithm allow obtaining optimal groundwater extraction values from the entire decision space of the input variables.

Recently, researchers have focused on reducing the required training patterns to train the SM adaptively using a relatively small number of training patterns (Sreekanth and Datta 2014b; Kourakos and Mantoglou 2009; Papadopoulou et al. 2010; Christelis and Mantoglou 2016; Sreekanth and Datta 2010). Adaptive training of SMs is able to reduce the number of training patterns significantly in developing a reasonably accurate approximate emulator of the density-dependent coupled flow and salt transport processes (Sreekanth and Datta 2010; Christelis and Mantoglou 2016). For adaptive training of MNN and GP-based SMs, Sreekanth and Datta (2010) utilized an expanding set method in which more and more training patterns generated by the optimization strategy are added to the initial training pattern based on the direction of search. Initially, trained SMs with the limited training pattern are used in conjunction with an optimization algorithm to find the near optimal solution. In this approach, initial approximate optimal solutions are obtained utilizing the SM based on the limited training data covering the entire feasible range of solution. However, once an approximate optimal solution is obtained using the SM, additional training data can be generated near this approximate optimal solution. These new training data are utilized for retraining the SM near the approximate optimal solution for improved accuracy. Then, the optimization routine is re-run to obtain an optimal solution that is more accurate. The process is continued until the desired level of accuracy of the SM is achieved.

Christelis and Mantoglou (2016) developed an online-training scheme of RBF-based SM that is embedded within an optimization algorithm. Their approach was also associated with adding infill points to the initial sampling plan using the current best solutions found by the RBF model during the optimization operations. This infill strategy favours a fast improvement of the RBF model at the region of the current optimum (local exploitation). However, it neglects the global improvement of the SM and might fail to identify the region of the global optimum (Forrester et al. 2008). In this approach, SMs are based on the local optimal solutions of the optimization process, and provide more accurate predictions at limited regions of the total decision space of the input variables. Moreover, this approach does not ignore going back to the NSM repeatedly to evaluate the current best solutions. While trying to achieve the computational efficiency associated with generating required input–output patterns for SM training, this approach ignores the accuracy and uncertainty of the SM predictions.

Reducing dimension of the optimization problem utilizing zonation approach (Ataie-Ashtiani et al. 2014) is also evidenced to reduce the required number of training patterns to develop accurate SMs. Nevertheless, the number of required input–output training patterns depends on the complexity of the system and the number of decision variables. For a relatively complex problem, large number of input–output training patterns may be necessary to develop a sufficiently reliable and accurate SM.

5 Ensemble of Surrogate Models to Improve Prediction Accuracy and Reduce Prediction Uncertainty

Many researches on coupled S/O approach emphasized the use of computationally efficient SMs to achieve global optimal solution within a reasonable computational cost. Most of the saltwater intrusion management models utilized a single SM coupled to an optimization algorithm. These include utilization of a single SM either for a single-objective problem formulation (Ataie-Ashtiani et al. 2014; Christelis and Mantoglou 2016; Christelis et al. 2017; Kurtulus and Razack 2010; Rajabi and Ketabchi 2017) or for a multiple objective saltwater intrusion management problems (Bhattacharjya and Datta 2009; Dhar and Datta 2009; Roy and Datta 2017b, c, d; Sreekanth and Datta 2010, 2011b; Hussain et al. 2015). However, such an SM when coupled within the optimization algorithm may provide a misleading global optimal solution (Hou et al. 2017). An ensemble of such SMs may provide better accuracy in prediction, and help in providing better optimal solutions in a coupled S/O methodology (Sreekanth and Datta 2011b; Roy and Datta 2017c). The reason why an ensemble of SMs is preferred is that a standalone SM often fails to capture the underlying predictor–response relationships within the feasible regions of the input space. However, very little work has been conducted on the use of ensemble of SMs to address prediction uncertainty of surrogate modelling as well as to achieve a reliable global Pareto optimal solution (Roy and Datta 2017c, d; Sreekanth and Datta 2011b, 2014a, b). A saltwater intrusion management model utilizing ensemble-based coupled S/O approach is represented schematically, as shown in Fig. 2.

Ensemble of SMs can improve robustness of the predictions by extracting true trends in the data while protecting against single wrong SM by reducing the impact of poor predictions by the model (Goel et al. 2007). An ensemble surrogate-modelling approach is supposed to provide better prediction capability than individual models, because the ensemble is built by integrating the outputs from all individual models of the ensemble (Jovanović et al. 2015; Roy and Datta 2017c; Sreekanth and Datta 2014b). However, an ensemble should be developed from individual models that are adequately diverse and reasonably accurate in their own prediction capabilities. Accuracy of the individual SMs within the ensemble can be achieved by selecting the appropriate surrogate models. Nevertheless, selecting different SMs based on various available algorithms (e.g., ANN, ANFIS, RBF, MARS, etc.) for a specific problem may also require substantial time involving exploration and evaluation of performances. On the other hand, diversity

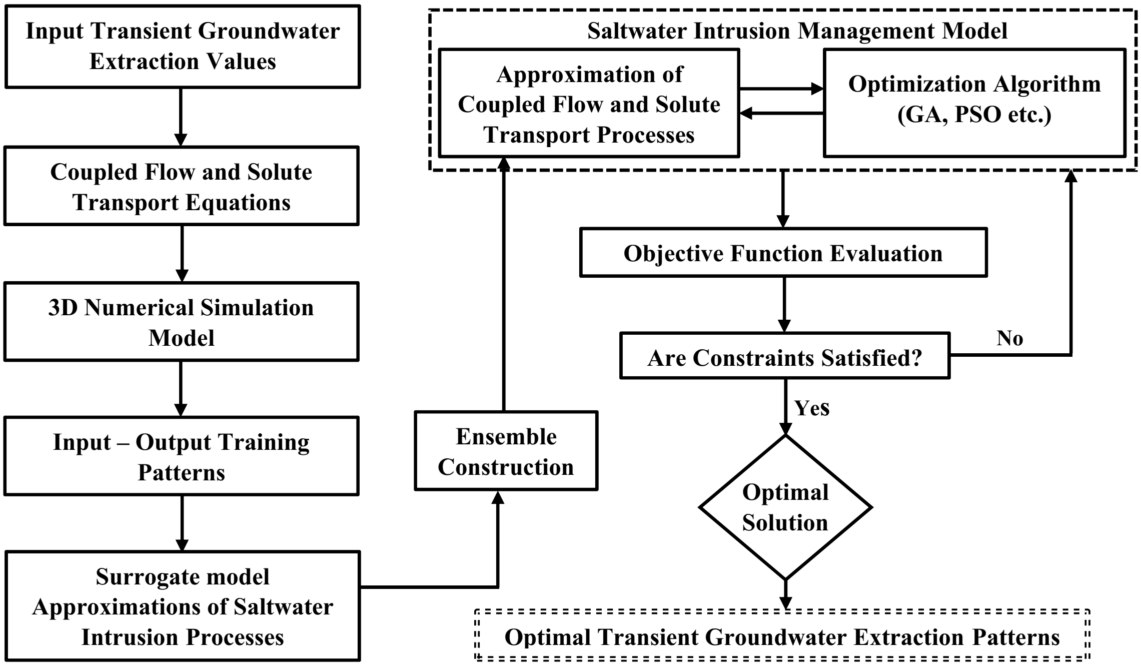
Input Transient Groundwater Extraction Values
Coupled Flow and Solute Transport Equations
3D Numerical Simulation Model
Input – Output Training Patterns
Surrogate model Approximations of Saltwater Intrusion Processes
Ensemble Construction
Saltwater Intrusion Management Model
Approximation of Coupled Flow and Solute Transport Processes
Optimization Algorithm (GA, PSO etc.)
Objective Function Evaluation
Are Constraints Satisfied?
No
Yes
Optimal Solution
Optimal Transient Groundwater Extraction Patterns

Fig. 2 Flow diagram of the ensemble-based coupled S/O approach in saltwater intrusion management models

can be maintained using different SMs (different training algorithms), different architectures of the same SM, or by manipulating the training data set utilizing different realizations to obtain different structures of the same SM (Shu and Ouarda 2007). An ensemble-modelling approach utilizes the distinctive feature of individual models to capture different patterns of the predictor–response relationships from the entire decision space.

The ensemble approach essentially serves as an effective way of accounting for model uncertainty (Goel et al. 2007). An ensemble of SMs is used to locate the regions of huge uncertainty by calculating the standard deviation of predictions at a certain design point. Higher standard deviations of predictions will be obtained in regions, where predictions of SMs vary significantly. The higher the standard deviation, the more will be the prediction uncertainty of any SM. However, lower values of the standard deviations do not ensure greater accuracy of prediction, because it may be possible that all SMs within the ensemble predict similar responses that produce low standard deviations (Goel et al. 2007).

5.1 Individual Surrogate Models Within the Ensemble

Each member of the ensemble is obtained either from an individual training algorithm or from a set of different training algorithms, often called a "heterogeneous ensemble". A heterogeneous ensemble may be obtained through combining two or more of the SMs described in Sect. 2. The best model structure of the individual models is obtained by optimizing model parameters with the same training data. An ensemble can also be formed by combining different architectures of the single best SM or by integrating different model realizations obtained from various realizations of training data sets.

In saltwater intrusion management problems, ensemble was formed utilizing a suitable SM but with different architectures of this SM by manipulating the training data either by bootstrap sampling (Sreekanth and Datta 2011b) or by random sampling without replacement (Roy and Datta 2017c) technique. In the ensemble formation approach by manipulating the training data set to obtain different architectures of the same SM, the best model is selected from a set of available SMs that are tested on the data set (Hou et al. 2017). Once the appropriate SM is selected, different realizations of this SM are obtained through different realizations of the training data sets. These training data realizations can be obtained through nonparametric bootstrap sampling (repeated random sampling with replacement) (Parasuraman and Elshorbagy 2008), random sampling without replacement technique (Hastie et al. 2008), bagging (Breiman 1996), or boosting (Schapire 1990; Freund and Schapire 1996). Each SM is different from each other within the ensemble as the surrogates are trained using different realizations of the training data set.

In bootstrap sampling, each data set is different from each other due to the repetition of some predictor–response training patterns and elimination of some patterns form the original training data set. Differential weighting of the bootstrap sampled training data sets enables the developed SMs to capture different trends of the data sets. Therefore, the resulting individual SMs predict future scenarios in different regions of the decision space depending on the associated differential weights of the different realizations of the training data sets from each region (Sreekanth and Datta 2011b). The COV of the RMSE of prediction of individual models can be used to quantify the predictive uncertainty of the models (Sreekanth and Datta 2011b). On the other hand, random sampling without replacement (Hastie et al. 2008) produces more diverse data sets for individual SM training. Each pair of the predictor–response training patterns is sufficiently different from each other, because each realization is obtained from permutation of the original data set. Therefore, the developed SMs are sufficiently different from each other with respect to their prediction capability, because these surrogates represent different regions of the decision space (Roy and Datta 2017c). The model parameters slightly differ among the different SMs trained on different sets of training data as the final values of the model parameters depend on the training data set.

Boosting algorithms (Schapire 1990; Freund and Schapire 1996) generate different distributions of the original training data to train a set of SMs. The training procedure starts with training the first SM with the original train data, and then, the training data sets for the subsequent SMs are resampled based on the predictive performance of the previous SMs. The training patterns, whose predicted outputs acquired from the preceding SM vary significantly from their actual values, are adjusted with higher probability of being sampled. Therefore, these training patterns are more likely to appear in the new training data sets for training of the next SM. On the other hand, the training patterns for which the preceding SM predictions are more accurate are less likely to appear in the new training data set. Thus, different SMs are supposed to provide better predictions in different parts of the decision space. The boosting algorithm comprises of developing a set of SMs in such a way that the subsequent SMs emphasize on training a set of scenarios that are not well captured by the previous SM (Schapire 1990). Boosting often provides better results compared to both bagging and randomization techniques, while the performances of both the bagging and randomization are almost similar (Dietterich 2000).

5.2 Optimum Number of Individual SMs in the Ensemble

Optimal number of individual SMs within the ensemble is decided based on the prediction accuracy and reduction of uncertainty in prediction. The optimum number of SMs within the ensemble is determined by incrementally adding the surrogates and checking the resulting root mean square error (RMSE) of the ensemble prediction (Roy and Datta 2017c) or by checking the resulting coefficient of variation of the resulting RMSE of ensemble prediction (Sreekanth and Datta 2011b). Sequential addition of individual surrogates in the ensemble is continued until it gives the smallest RMSE value (Roy and Datta 2017c), or until, there is no significant change in the uncertainty of the ensemble with further addition of SMs (Sreekanth and Datta 2011b). A more direct approach to determine the optimum number of individual members of the ensemble is to observe the resulting prediction error after sequential addition of individual SMs (Roy and Datta 2017c, d). In this approach, optimal size of the ensemble is one for which the resulting prediction error is minimum, and the addition of any further model increases the error. The downside of using this technique is that a sufficiently large number of individual models must be developed initially. Another drawback is that the cumulative prediction error may decrease again after subsequent increases depending on the added models' prediction accuracy. Therefore, the modeller must be careful in selecting the optimal ensemble size, and there is a trade-off between the prediction accuracy and model complexity. The integration of output from different SMs (based on different algorithms) as an ensemble-modelling approach is still lacking either in saltwater intrusion processes prediction or in developing saltwater intrusion management models.

5.3 Integration of Individual Surrogate Models

In ensemble-based surrogate-modelling approach, selection of the most suitable SM(s) is the most challenging task. The next challenge is to determine the contribution of the individual SMs' output in determining the final output of the ensemble. While some researchers prefer to use simple averaging approach (Roy and Datta 2017c; Sreekanth and Datta 2011b), the others proposed to assign some weights to individual SMs based on their performance on test data set (Goel et al. 2007; Zerpa et al. 2005). An ensemble of different SMs can be formed using weighted average (Goel et al. 2007; Zerpa et al. 2005), set pair analysis (SPA) (Hou et al. 2017), or by Dempster–Shafer theory (DST) (Müller and Piché 2011).

5.3.1 Weighted Average

The most commonly used integration technique to combine individual members of the ensemble is the simple averaging technique. Examples of using simple averaging method to combine individual SMs to form an ensemble for developing saltwater intrusion management model include the work of Roy and Datta (2017c) and Sreekanth and Datta (2011b). In simple averaging technique, outputs from the individual SMs are combined through assigning uniform weights to each individual SMs within the ensemble to provide equal priorities to all considered SMs. This is done by adding the outputs generated by each individual learner model divided by the optimum number of learner models within the ensemble, mathematically given by

$$\text{OutEnsemble} = \sum_{i=1}^{n} \frac{\text{OutIndividualLearner}_{\text{i}}}{n}, \tag{12}$$

where $\text{OutIndividualLearner}_i$ is the output from the ith individual model, and n is the number of individual models within the ensemble.

The other suggested way of using ensemble of surrogates is to construct weighted average ensemble SM. It was observed that the weighted average SM yielded the best correlation between the actual and predicted responses for different test problems and showed relatively low sensitivity to the choice of design of experiments (DOE) (Goel et al. 2007). In the concept of weighted average, more accurate SMs are assigned larger weights and vice versa. The sum of weights given to each SM within the ensemble must be equal to one. Nevertheless, the right choice of weights is still an art of selection. Weighted average surrogate or optimal weighted surrogate (Viana et al. 2009) is a technique to construct ensembles in which performance of worse SMs within the ensemble is equilibrated by assigning particular weights to individual surrogates. Weighted average of individual surrogates, known as ensemble of surrogates, provides lower variance of prediction compared to that obtained using standalone SMs (Bishop 1995). Individual SMs developed using different algorithms (e.g., ANN, ANFIS, GP, MARS, GPR, etc.) can be combined together by assigning different weights to individual models. Mathematically, it can be represented by the form of a deterministic function as

$$Y_{\text{WA}}(X) = \sum_{i=1}^{n} \omega_i(X) \times Ys_i(X), \tag{13}$$

where Y_{WA} is the summed prediction of the ensemble model based on weighted average, Ys_i represents the prediction from the ith SM, ω_i denotes the weight given to ith SM, and n is the number of SMs within the ensemble. As the weights are a function of X, the resulting ensemble can be regarded as adaptive in nature (Zerpa et al. 2005). To achieve minimal variance in the unbiased weighted average ensemble models, predictions from the individual SMs should be unbiased and independent. In that situation, the unbiased weight can be calculated by

$$\omega_i = \frac{\frac{1}{V(i)}}{\sum_{i=1}^{n} \frac{1}{V(i)}}, \tag{14}$$

where $V(i)$ represents the variance of prediction from the ith SM. This method of ensemble construction is based on the SMs' prediction variance reduction (Zerpa et al. 2005).

Selection of weights need to be chosen carefully, so that they should provide confidence in the particular SM, and weights should rule out the adverse effects of surrogate modelling related to an SM that performs worse on unseen data. Goel et al. (2007) described a global weights' selection scheme in which three weighting methods are implemented based on global database measure of goodness. The first one deals with relative magnitude of errors, and the weights are functions of these errors. The amount of weights given to the best SM in this approach depends on the number of SMs within the ensemble. For large number of SMs used to construct the ensemble, weights given to the best SM become unreasonably low. Some SMs may perform well in sampled data set, but provide poor predictions on unseen data. This weighting approach minimizes the ensemble prediction errors by protecting against errors produced by these types of SMs. The second weighting approach is referred to as best PRESS model in which the best model (in terms of prediction accuracy based on the sampled data) within the ensemble is assigned a weight of one whereas all other SMs within the ensemble are given a zero weight. This strategy is nothing but picking the best model among all considered SMs within the ensemble. The downside of this approach is that the selected best model may perform well in the sampled data, but may be quite incapable of showing good performance on the new unseen data. In the third strategy, weights are selected in a way that they address drawbacks of the above-mentioned strategies. This approach introduces two parameters to control simultaneously the relative importance of individual SMs and the importance of averaging.

5.3.2 Set Pair Analysis (SPA)

"SPA sets the certainty and uncertainty as a mutually associated, mutually restricted, and mutually penetrated system" (Hou et al. 2017). SPA utilizes two different sets of responses, one obtained from the NSM whereas the other obtained from a number of SMs on the test predictor–response data samples. Then, the connection degree between the NSM and a certain number of SMs as well as the set pair weight of the SMs are calculated. Suppose, a set pair $S(A, B)$ is constructed from a set A and a relative set B. The nth terms in set A and B express the characteristics of set A and B. Three classifiers, e.g., identity, discrepancy, and contradistinction (Wang et al. 2009b), can be used to describe the relationships between the sets A and B to obtain a three-element connection degree as follows:

$$\mu^m = \frac{s^m}{n} + \frac{d^m}{n} i + \frac{c^m}{n} j, \tag{15}$$

in which μ^m denotes the connection weights of the mth SM; n is the total number of characteristics.s, d, and c represent the numbers of identical, discrepant, and contradictory characteristics, respectively; i and j are the coefficients of discrepancy and contradistinction, respectively.

Percentage absolute error (PAE) between the NSM and the SM predictions is usually used to classify the set pair. For instance, Hou et al. (2017) used four categories: identical (PAE < 0.3%), mild discrepancy (PAE between 0.3 and 0.6%), severe discrepancy (PAE between 0.6 and 1%), and contradistinction (PAE > 1%).

Set pair weight of the mth SM is then calculated by

$$\omega_m = U_m / \sum_{1}^{k} U_m \tag{16}$$

$$U_m = 0.5 \times \mu_m + 0.5. \tag{17}$$

In this approach, a number of different SMs are developed separately and their set pair weights are calculated based on their prediction errors on test data sets. The outputs from the ensemble are the weighted sum of the outputs from the individual SMs within the ensemble.

5.3.3 Dempster–Shafer Theory (DST)

The choice of the appropriate models and their integration can be accomplished by utilizing DST (Dempster 1968; Shafer 1976), which provides a way of integrating information from various sources to build a degree of belief. The DST permits integration of conflicting information that may be imprecise, vague, and uncertain. Three functions, namely, belief, plausibility, and pignistic probability, are used to calculate the reliability of a given hypothesis. These functions are analogous to model characteristics [correlation coefficients or mean square error (MSE)] for SM-modelling approaches (Müller and Piché 2011). Sometimes, a certain SM may provide conflicting performance in terms of statistical performance indices, i.e., the performance may be good (high correlation coefficients) and bad (high MSE) at the same time. The DST is an effective way of addressing this conflict by calculating pignistic probabilities for all considered SMs. Based on the values of this pignistic probabilities, it is possible to decide on which of all considered models is the best, or whether an ensemble model is suitable for this problem, which weight should be assigned to each individual SMs (Müller and Piché 2011). This approach can be utilized in ensemble SM-based saltwater intrusion management models.

6 Reduction of Computational Burden in Ensemble-Based Management Models

Computational requirement is an important issue in coupled S/O-based saltwater intrusion management models especially when an ensemble of surrogates is coupled to an optimization algorithm. Despite providing better prediction accuracy and reliability of prediction by addressing prediction uncertainty, an ensemble of surrogates inevitably adds an extra computational burden to the coupled S/O approach. An ensemble consists of two or more surrogates, which need to be individually coupled to an optimization algorithm (Roy and Datta 2017c; Sreekanth and Datta 2011b) to calculate the responses, and the responses are combined to check the constraint satisfaction. These additional steps essentially require more computational requirements when compared to single surrogate-based coupled S/O approach. Nevertheless, researchers have put several efforts to achieve computational efficiency in the ensemble-based management models. The following sub-sections briefly outline some of the techniques used to achieve this goal.

6.1 Choice of Individual SM Structure

One of the most effective ways of reducing computational burden is to choose simple SM structure that is sufficiently accurate and adequately reliable to capture the coupled flow and salt transport processes in coastal aquifers. However, not all SMs are equally suitable for all physical systems, i.e., models' performance in terms of prediction accuracy is problem specific. Therefore, it should always be a wise decision to select an appropriate SM for the specific problem at hand. For example, Christelis and Mantoglou (2016) have obtained a 96% reduction in computational cost using RBF model-based pumping optimization strategy when compared to NSM-based optimization strategy for a single-objective problem setting. Therefore, RBF could be an ideal candidate for ensemble construction either by combining with other SMs or using different realizations of RBF models by manipulating the training data. However, the performance of an ensemble of RBF models within a coupled S/O methodology for multi-objective problem settings needs to be evaluated.

For multi-objective saltwater intrusion management problems, ensembles of MARS (Roy and Datta 2017c), GP (Sreekanth and Datta 2011b), and ANFIS (Roy and Datta 2017d) have been evaluated as an accurate and reliable ensemble-based surrogate-modelling approaches to obtain global Pareto optimal pumping management strategies within a reasonable computational cost. As MARS and GP model structures are relatively simple, their ensembles do not incur significant computational cost within the optimization framework. On the other hand, ANFIS model structures are comparatively complex, and as such, their ensemble may require more computational requirements compared to simpler ones, i.e., GP and MARS-based ensembles. For instance, to solve a multi-objective saltwater intrusion management model of medium complexity, MARS-based ensemble SMs (139 MARS models were coupled as binding constraints) took only 76 min to evaluate 3,990,401 functions in achieving the global Pareto optimal pumping management strategies (Roy and Datta 2017c). Roy and Datta (2017d) used an ensemble of 111 ANFIS models to develop a saltwater intrusion management model in a multi-layered coastal aquifer system. The computational time required to solve this optimization problem is not provided and needs to be evaluated in future. While computational cost is an important issue, the trade-offs between the prediction accuracies and computational burden must be addressed carefully. Accuracy and associated uncertainties in prediction should not be overshadowed by computational efficiency attainment.

For another multi-objective saltwater intrusion management model, EPR achieves a significant computational saving when compared to direct linking with the simulation model in a coupled S/O methodology (Hussain et al. 2015). The overall average time required by EPR to solve the optimization problem (including the time required to generate 500 sets of predictor–response training patterns for EPR training) is less than 10% of the time required to solve the same optimization problem using the NSM. Therefore, based on the less computational time requirement, EPR-based SMs can be used to form ensembles either with other types of computationally cheap surrogate models or with different realizations of EPR using different realizations of training data sets. An ensemble of EPR with other computationally efficient surrogate models, i.e., GP and MARS utilizing a suitable integration technique (simple or weighted average) to develop multiple objective pumping optimization in coastal aquifers would be an interesting topic for future research.

Another promising approach to keep the model structure simple while maintaining the adequate accuracy especially in fuzzy logic-based modelling approaches is to divide the large-dimensional input space into identical clusters using FCM or other clustering algorithms. For saltwater intrusion problems, (Roy and Datta 2017a, b, d) utilized FCM algorithm to divide large-dimensional input space into identical clusters. Choice of the optimum number of clusters was made by conducting numerical experiments using different numbers of clusters and observing the resulting prediction errors of the SMs. The optimum number of clusters is one that produces minimum prediction error as well as minimum variation of training and testing errors. As the number of clusters determines the amount of model complexity in terms of linear and non-linear parameters of fuzzy logic-based FIS structures, a judicial decision considering the trade-offs between prediction accuracy and model complexity is of practical importance.

6.2 Method of Ensemble Formation

The way that the ensembles are formed largely influence the accuracy, computational feasibility, and efficiency of modelling. As mentioned in Sect. 5.3, ensembles can be constructed by integrating two or more different SMs, often called a heterogeneous ensemble. In this approach, right choice of the SMs that produce reliable prediction accuracies should be chosen based on the specific problems. Afterwards, based on the prediction accuracies of individual SMs, and consequently on the specific weights given to individual surrogates, the total number of SMs within the ensemble can be kept minimum. If one feels that a particular SM provides far better prediction accuracies than others for the problem in hand, then ensemble can be constructed by varying different parameters of the same SM and combining the outputs by any suitable integration technique. For instance, ensembles of ANFIS can be developed by varying different membership functions (Melin et al. 2012). In their study, three ANFIS structures were developed using triangular, sigmoid, and π-shaped builtin membership functions, respectively. The outputs from these three ANFIS models were then integrated to predict a chaotic time series (the Mackey–Glass, Dow Jones, and Mexican stock exchange time series). In another study, Jovanović et al. (2015) combined the outputs from three different network structures, e.g., Feed Forward Neural Network (FFNN), Radial Basis Function Network (RBFN), and ANFIS to predict heating energy consumption.

However, in all situations, the motivation should be to minimize the number of models within the ensemble by observing the prediction accuracies of individual models and specific weight given to each model. Varying training data set realizations (Roy and Datta 2017c; Sreekanth and Datta 2011b) for training a set of SMs based on a relatively simple, adequately reliable, and sufficiently accurate SM algorithm can achieve a considerable computational efficiency in a coupled S/O-based saltwater intrusion management model. In this approach, optimal number of SM realizations to construct an ensemble can be selected by sequentially adding surrogates and checking the resulting RMSE values (Roy and Datta 2017c, d) or by checking standard deviation (Goel et al. 2007) or by checking COV of RMSE values (Sreekanth and Datta 2011b). A comparative evaluation of the computational requirements of different SMs to evolve global Pareto optimal solution for coupled S/O-based saltwater intrusion management model for a particular problem would be beneficial to sort the accurate and computationally efficient SMs. Nevertheless, the effectiveness of the accurate and computationally cheap SMs in ensemble formation needs to be evaluated.

6.3 Right Choice of Computationally Cheap Optimization Algorithms

Selecting an appropriate optimization algorithm (OA) for specific purposes is one of the most important decisions to achieve global optimal solutions with a reasonably small computational effort. The performance of an OA largely depends on the specific purposes, and it is impossible to have an OA that is superior to others for all situations and suitable for all purposes (Wolpert and Macready 1997; Pham and Castellani 2014). GA and its variants are most commonly used OAs for saltwater intrusion management problems, especially in multi-objective optimization formulations (Kourakos and Mantoglou 2011; Dhar and Datta 2009; Roy and Datta 2017b, c, d; Sreekanth and Datta 2011b). Other types of OAs used in saltwater intrusion management problems were limited to solve single-objective optimization formulations. Ataie-Ashtiani and Ketabchi (2011) utilized Continuous Ant Colony Optimization (CACO) to solve a single-objective coastal aquifer management problem assuming a sharp interface solution for steady-state problem. Differential Evolution (DE) was used to solve a single-objective coastal subsurface water management problem subjected to environmental criteria (Karterakis et al. 2007). For another coastal aquifer management problem, CACO and Particle-Swarm Optimization (PSO) were recommended for solving single-objective coastal groundwater management problems (Ketabchi and Ataie-Ashtiani 2015b). The authors demonstrated that CACO provides 17% better solutions than Artificial Bee Colony Optimization (ABC). For this single-objective optimization formulation, the performance of the OAs based on computational time requirement is: CACO > PSO > GA (Ketabchi and Ataie-Ashtiani 2015b). However, the performance of these OAs was not evaluated for multiple objective saltwater intrusion management problems, and is still a challenge and opportunity for future researches. In addition, a comparison of different OAs to solve multi-objective saltwater intrusion management models to evaluate the computational requirements could be the topic of future research.

The previous literature revealed that genetic algorithm and its variants are utilized in most of the multi-objective optimization formulation of coastal groundwater management problems. These include both hypothetical study areas (Park and Aral 2004; Bhattacharjya and Datta 2009; Dhar and Datta 2009; Sreekanth and Datta 2010; Kourakos and Mantoglou 2011; Abd-Elhamid and Javadi 2011; Javadi et al. 2012) and real-world case studies (Qahman et al. 2009; Abd-Elhamid and Javadi 2011; Ataie-Ashtiani et al. 2014). However, this review article specifically focuses on the optimization algorithm used in ensemble surrogate-based saltwater intrusion management problems. In this problem domain, Sreekanth and Datta (2011b) externally coupled GP-based ensemble SMs within a non-controlled elitist multi-objective genetic algorithm (Deb et al. 2000). On the other hand, controlled elitist multi-objective genetic algorithm (CEMGA) (Deb and Goel 2001) has been successfully applied to attain global Pareto optimal solutions in saltwater intrusion management problems utilizing MARS (Roy and Datta 2017c) and ANFIS (Roy and Datta 2017d)-based ensemble SMs. Fast converging optimization algorithms, CACO and PSO in single-objective test problems (Ketabchi and Ataie-Ashtiani 2015b), should be tested for multi-objective problem settings of coastal aquifer management problems. This will further verify their performance on multi-objective constrained problem settings in which the SMs are also used as binding constraints.

6.4 Pre-specified Constraints on Computational Cost

Applying pre-specified restrictions on the overall computational cost (Christelis et al. 2017) are applied to achieve computational efficiency in the surrogate-based optimization approach of coastal aquifer management problems. However, Razavi et al. (2012) questioned about the advantages of SMs for complex optimization problems under imposed limited computational budgets. For a single-objective pumping optimization problem, Christelis et al. (2017) demonstrated the performance of SMs for different dimensionalities of the decision variable space under imposed restrictions on the number of numerical model simulations. This approach can be utilized in ensemble surrogate-based multi-objective pumping optimization problems in coastal aquifers. However, care should be taken to ensure that any attempt implemented with a desire to attain computational efficiency should not compromise with the accuracy and reliability of the management models. Therefore, the accuracy of the management model with and without computational constraint should be justified.

6.5 Parallel Computation

Ensemble-based saltwater intrusion management models demand further efficiency in computation time requirement that can be achieved utilizing a parallel-processing strategy (Ketabchi and Ataie-Ashtiani 2015a; Roy and Datta 2017b, c, d). Parallel and distributed computing not only speeds up the search process of optimization algorithms, but also improves the obtained global Pareto optimal fronts and robustness of the acquired optimal solutions for large-scale problems (Talbi et al. 2008). The advent of multi-core processor computers facilitates the use of shared memory parallelism approaches (McLaughlin 2008; Tang et al. 2010), which provides fast computations by distributing the tasks among associated worker machines (cores of a multi-core computer). Ketabchi and Ataie-Ashtiani (2015a) obtained a speedup ratio of up to 3.53 on an eight-core processor by implementing the parallel-processing strategy compared to when implemented in serial. The authors concluded that implementation of parallel strategy significantly increases computational performance of optimization algorithm for single-objective coastal groundwater management problems.

Multi-objective PSO (MOPSO) within the framework of parallel computing has been found in literature for coupled S/O of bridge maintenance planning (Yang et al. 2012). However, the use of MOPSO has neither been used for single surrogate-based coastal aquifer management problems nor ensemble surrogate-based saltwater intrusion management problems. For a multi-objective problem setting, Roy and Datta (2017c) reported a considerable amount of computational efficiency using parallel-processing capability through distributing the objective functions and constraints among the four physical cores of a seven core PC. The multi-objective saltwater intrusion management model was developed using the ensemble SM-based coupled S/O approach. The optimization routine took 1.09 and 2.57 h to obtain the global Pareto optimal groundwater extraction values with and without parallel-processing implementation.

In some cases, computationally intensive SMs need to be utilized to achieve better prediction accuracy and to accurately represent the complex and non-linear flow and salt transport processes. An ensemble of such SMs to address prediction uncertainty and to achieve even better prediction accuracy may demand considerable amount of computational efforts in a coupled S/O approach. Moreover, heterogeneous ensembles in which several complex-structured SMs are integrated also require a significant computational demand when linked to an optimization algorithm within the framework of coupled S/O approach. In these situations, parallel computing strategy using single–multi-core PCs may not be adequate, and other high-performing computations are recommended.

7 Conclusions and Recommendations for Further Research

Development of saltwater intrusion management models depends solely on coupled S/O methodology in which the simulation part is usually replaced by a properly trained and validated SM. These approximate surrogates are essential components of the coupled S/O methodology as direct linking of the original NSM is impractical due to huge computational requirement. However, SM predictions are always associated with a certain amount of uncertainty that can be addressed by forming an ensemble of such SMs. Ensembles can be formed by integrating different realizations of the same SM through manipulation of the training data set. Another approach of ensemble formation utilizes different SMs using the so-called heterogeneous ensemble approach. Individual SM within the ensemble is integrated either by simple averaging or by weighted averaging approach. Because multiple SMs are integrated in an ensemble, linking each contributing individual SM to the optimization framework inevitably adds an extra computational burden to the coupled S/O approach. Parallel-processing framework is believed to achieve additional computational efficiency in the ensemble-based coupled S/O approach for saltwater intrusion management problems. Parallel processing is performed either using multiple cores of the same PC or by combining multiple cores from several PCs or by implementing any other high-performing computation technique. Obviously, the choice of parallel-processing technique depends on the complexity of the problem itself, structure of the SMs used, complexity of the optimization formulation in terms of the number of linear and non-linear constraints, and the nature of optimization formulations (single or multiple objective).

A number of areas require specific attention that could be the interesting topics for future research. Only a few studies has focused on ensemble approach for achieving better prediction accuracy and for reducing prediction uncertainty in coupled S/O-based saltwater intrusion management problems. These studies are based on generating different realizations of a single SM by manipulating the training data and combing them by simple averaging technique (Roy and Datta 2017c, d; Sreekanth and Datta 2011b). Therefore, future researches need to explore the development and implementation of weighted average approaches of combining multiple SMs (based on different algorithms) to develop saltwater intrusion management models in coastal aquifers. In particular, there is a need to utilize unique features of several SMs within a general framework to address prediction uncertainty of surrogate-based modelling approaches. Quantification of uncertainty reduction utilizing ensemble-based surrogate-modelling approaches demands more rigorous investigations using different approaches. There is a lack of established measure of surrogate induced prediction uncertainty reduction by implementing an ensemble of two or more surrogates. Furthermore, a comparison between the quality of the global Pareto optimal solutions provided by a single SM and an ensemble of SMs (especially heterogeneous surrogates) also demand further research. Quantification of uncertainty using single SM-based coupled S/O approach and that utilizing an ensemble of SMs warrant further research.

Compliance with Ethical Standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Abd-Elhamid HF, Javadi AA (2011) A cost-effective method to control seawater intrusion in coastal aquifers. Water Resour Manage 25(11):2755–2780. https://doi.org/10.1007/s11269-011-9837-7

Ataie-Ashtiani B, Ketabchi H (2011) Elitist continuous ant colony optimization algorithm for optimal management of coastal aquifers. Water Resour Manage 25(1):165–190. https://doi.org/10.1007/s11269-010-9693-x

Ataie-Ashtiani B, Ketabchi H, Rajabi MM (2014) Optimal management of a freshwater lens in a small island using surrogate models and evolutionary algorithms. J Hydrol Eng 19(2):339–354. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000809

Bazi Y, Alajlan N, Melgani F (2012) Improved estimation of water chlorophyll concentration with semisupervised gaussian process regression. IEEE Geosci Remote Sens Lett 50(7):2733–2743. https://doi.org/10.1109/TGRS.2011.2174246

Bera P, Prasher SO, Patel RM, Madani A, Lacroix R, Gaynor JD, Tan CS, Kim SH (2006) Application of MARS in simulating pesticide concentrations in soil. T Asabe 49(1):297–307

Bhattacharjya RK, Datta B (2005) Optimal management of coastal aquifers using linked simulation optimization approach. Water Resour Manage 19(3):295–320. https://doi.org/10.1007/s11269-005-3180-9

Bhattacharjya RK, Datta B (2009) ANN-GA-based model for multiple objective management of coastal aquifers. J Water Resour Plan Manage 135(5):314–322. https://doi.org/10.1061/(ASCE)0733-9496(2009)135:5(314)

Bhattacharjya RK, Datta B, Satish MG (2007) Artificial neural networks approximation of density dependent saltwater intrusion process in coastal aquifers. J Hydrol Eng 12(3):273–282. https://doi.org/10.1061/(ASCE)1084-0699(2007)12:3(273)

Bishop CM (1995) Neural networks for pattern recognition. Oxford University Press Inc, Oxford

Bishop C (2006) Pattern recognition and machine learning. Springer-Verlag, New York

Blanning RW (1975) The construction and implementation of metamodels. Simulation 24:177–184

Breiman L (1996) Bagging predictors. Mach Learn 24(2):123–140. https://doi.org/10.1023/a:1018054314350

Cherkassky V (1998) Fuzzy inference systems: a critical review. In: Kaynak O, Zadeh LA, Türkşen B, Rudas IJ (eds) Computational intelligence: soft computing and fuzzy-neuro integration with applications. Springer, Berlin, pp 177–197. https://doi.org/10.1007/978-3-642-58930-0_10

Christelis V, Mantoglou A (2016) Pumping optimization of coastal aquifers assisted by adaptive metamodelling methods and radial basis functions. Water Resour Manage 30(15):5845–5859. https://doi.org/10.1007/s11269-016-1337-3

Christelis V, Regis RG, Mantoglou A (2017) Surrogate-based pumping optimization of coastal aquifers under limited computational budgets. J Hydroinform. https://doi.org/10.2166/hydro.2017.063

Deb K, Goel T (2001) Controlled elitist non-dominated sorting genetic algorithms for better convergence. In: Zitzler E, Thiele L, Deb K, Coello Coello CA, Corne D (eds) Evolutionary multi-criterion optimization: first international conference, EMO 2001 Zurich, Switzerland, March 7–9, 2001 proceedings. Springer, Berlin, pp 67–81. https://doi.org/10.1007/3-540-44719-9_5

Deb K, Agrawal S, Pratap A, Meyarivan T (2000) A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: NSGA-II. In: Schoenauer M, Deb K, Rudolph G et al (eds) Parallel problem solving from nature PPSN VI: 6th international conference Paris, France, September 18–20, 2000 proceedings. Springer, Berlin, pp 849–858. https://doi.org/10.1007/3-540-45356-3_83

Dempster AP (1968) A generalization of bayesian inference. J R Stat Soc Ser B (Methodol) 30(2):205–247

Dhar A, Datta B (2009) Saltwater intrusion management of coastal aquifers. I: linked simulation-optimization. J Hydrol Eng 14(12):1263–1272. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000097

Dietterich TG (2000) An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. Mach Learn 40(2):139–157. https://doi.org/10.1023/a:1007607513941

El-Baroudy I, Elshorbagy A, Carey SK, Giustolisi O, Savic D (2010) Comparison of three data-driven techniques in modelling the evapotranspiration process. J Hydroinform 12(4):365–379. https://doi.org/10.2166/hydro.2010.029

Forrester AIJ, Sóbester A, Keane AJ (2008) Constructing a surrogate. Engineering design via surrogate modelling. Wiley, Oxford, pp 33–76. https://doi.org/10.1002/9780470770801.ch2

Freund Y, Schapire RE (1996) Experiments with a new boosting algorithm. Paper presented at the proceedings of the thirteenth international conference on international conference on machine learning, Bari, Italy

Friedman JH (1991) Multivariate adaptive regression splines (with Discussion). Ann Stat 19(1):1–67. https://doi.org/10.1214/aos/1176347963

Giustolisi O, Savic DA (2006) A symbolic data-driven technique based on evolutionary polynomial regression. J Hydroinform 8(3):207–222. https://doi.org/10.2166/hydro.2006.020

Goel T, Haftka RT, Shyy W, Queipo NV (2007) Ensemble of surrogates. Struct Multidiscip O 33(3):199–216. https://doi.org/10.1007/s00158-006-0051-9

Green PJ, Silverman BW (1993) Nonparametric regression and generalized linear models: a roughness penalty approach. Taylor & Francis, Abingdon

Hastie T, Tibshirani R, Friedman J (2008) The elements of statistical learning, 2nd edn. Springer, New York

Holman D, Sridharan M, Gowda P, Porter D, Marek T, Howell T, Moorhead J (2014) Gaussian process models for reference ET estimation from alternative meteorological data sources. J Hydrol 517:28–35. https://doi.org/10.1016/j.jhydrol.2014.05.001

Hou Z, Lu W, Xue H, Lin J (2017) A comparative research of different ensemble surrogate models based on set pair analysis for the DNAPL-contaminated aquifer remediation strategy optimization. J Contam Hydrol 203:28–37. https://doi.org/10.1016/j.jconhyd.2017.06.003

Hsieh WW, Tang B (1998) Applying neural network models to prediction and data analysis in meteorology and oceanography. Bull Am Meteorol Soc 79(9):1855–1870. https://doi.org/10.1175/1520-0477(1998)079%3c1855:annmtp%3e2.0.co;2

Hussain MS, Javadi AA, Ahangar-Asr A, Farmani R (2015) A surrogate model for simulation–optimization of aquifer systems subjected to seawater intrusion. J Hydrol 523:542–554. https://doi.org/10.1016/j.jhydrol.2015.01.079

Jang J-SR (1993) ANFIS: adaptive-network-based fuzzy inference system. IEEE Trans Syst Man Cybern 23(3):665–685. https://doi.org/10.1109/21.256541

Jang J-SR, Sun CT, Mizutani E (1997) Neuro-fuzzy and soft computing: a computational approach to learning and machine intelligence. Prentice Hall, Upper Saddle River

Javadi AA, Abd-Elhamid HF, Farmani R (2012) A simulation-optimization model to control seawater intrusion in coastal aquifers using abstraction/recharge wells. Int J Numer Anal Methods Geomech 36(16):1757–1779. https://doi.org/10.1002/nag.1068

Jovanović RŽ, Sretenović AA, Živković BD (2015) Ensemble of various neural networks for prediction of heating energy consumption. Energ Build 94:189–199. https://doi.org/10.1016/j.enbuild.2015.02.052

Karterakis SM, Karatzas GP, Nikolos IK, Papadopoulou MP (2007) Application of linear programming and differential evolutionary optimization methodologies for the solution of coastal subsurface water management problems subject to environmental criteria. J Hydrol 342(3):270–282. https://doi.org/10.1016/j.jhydrol.2007.05.027

Ketabchi H, Ataie-Ashtiani B (2015a) Assessment of a parallel evolutionary optimization approach for efficient management of coastal aquifers. Environ Model Softw 74:21–38. https://doi.org/10.1016/j.envsoft.2015.09.002

Ketabchi H, Ataie-Ashtiani B (2015b) Evolutionary algorithms for the optimal management of coastal groundwater: a comparative study toward future challenges. J Hydrol 520:193–213. https://doi.org/10.1016/j.jhydrol.2014.11.043

Kourakos G, Mantoglou A (2009) Pumping optimization of coastal aquifers based on evolutionary algorithms and surrogate modular neural network models. Adv Water Resour 32(4):507–521. https://doi.org/10.1016/j.advwatres.2009.01.001

Kourakos G, Mantoglou A (2011) Simulation and multi-objective management of coastal aquifers in semi-arid regions. Water Resour Manage 25(4):1063–1074. https://doi.org/10.1007/s11269-010-9677-x

Koza JR (1994) Genetic programming as a means for programming computers by natural selection. Stat Comput 4(2):87–112. https://doi.org/10.1007/bf00175355

Kurtulus B, Razack M (2010) Modeling daily discharge responses of a large karstic aquifer using soft computing methods: artificial neural network and neuro-fuzzy. J Hydrol 381(1–2):101–111. https://doi.org/10.1016/j.jhydrol.2009.11.029

Langevin CD, Thorne DTJ, Dausman AM, Sukop MC, Guo W (2007) SEAWAT Version 4: a computer program for simulation of multi-species solute and heat transport. US Geol Surv Tech Methods Book 6(Chapter A22):39

Lin H-CJ, Rechards DR, Talbot CA, Yeh G-T, Cheng J-R, Cheng H-P, Jones NL (1997) FEMWATER: a three-dimensional finite element computer model for simulating density-dependent flow and transport in variable saturated media. Technical Rep No CHL-97-12 Vicksburg, MS: US Army Engineer Waterways Experiment Station Coastal and Hydraulics Laboratory

Mantoglou A, Papantoniou M (2008) Optimal design of pumping networks in coastal aquifers using sharp interface models. J Hydrol 361(1):52–63. https://doi.org/10.1016/j.jhydrol.2008.07.022

MATLAB (2017) What is sugeno-type fuzzy inference? MATLAB documentation http://aumathworks.com/help/fuzzy/what-is-sugeno-type-fuzzy-inference.html Accessed 12 Dec 2017

McLaughlin JD (2008) Parallel processing of reactive transport models using OpenMP. M.Sc. Thesis, Brigham Young University, Provo, UT, USA

Melin P, Soto J, Castillo O, Soria J (2012) A new approach for time series prediction using ensembles of ANFIS models. Expert Syst Appl 39(3):3494–3506. https://doi.org/10.1016/j.eswa.2011.09.040

Müller J, Piché R (2011) Mixture surrogate models based on Dempster-Shafer theory for global optimization problems. J Glob Optim 51(1):79–104. https://doi.org/10.1007/s10898-010-9620-y

Papadopoulou MP, Nikolos IK, Karatzas GP (2010) Computational benefits using artificial intelligent methodologies for the solution of an environmental design problem: saltwater intrusion. Water Sci Technol 62(7):1479–1490

Parasuraman K, Elshorbagy A (2008) Toward improving the reliability of hydrologic prediction: model structure uncertainty and its quantification using ensemble-based genetic programming framework. Water Resour Res 44(12):W12406. https://doi.org/10.1029/2007WR006451

Park CH, Aral MM (2004) Multi-objective optimization of pumping rates and well placement in coastal aquifers. J Hydrol 290(1):80–99. https://doi.org/10.1016/j.jhydrol.2003.11.025

Pebesma EJ, Heuvelink GBM (1999) Latin hypercube sampling of gaussian random fields. Technometrics 41(4):303–312. https://doi.org/10.2307/1271347

Pham DT, Castellani M (2014) Benchmarking and comparison of nature-inspired population-based continuous optimisation algorithms. Soft Comput 18(5):871–903. https://doi.org/10.1007/s00500-013-1104-9

Pillay N (2004) An investigation into the use of genetic programming for the induction of noviceprocedural programming solution algorithms in intelligent programming tutors. Dissertation, University of KwaZulu-Natal, Durban

Piret C (2007) Analytical and numerical advances in radial basis functions. PhD dissertation, Univ of Colorado, Boulder, CO

Qahman K, Larabi A, Ouazar D, Naji A, Cheng AHD (2009) Optimal extraction of groundwater in Gaza coastal aquifer. J Water Resour Protect 4:249–259

Rajabi MM, Ketabchi H (2017) Uncertainty-based simulation-optimization using Gaussian process emulation: Application to coastal groundwater management. J Hydrol 555(Supplement C):518–534. https://doi.org/10.1016/j.jhydrol.2017.10.041

Rasmussen CE, Williams CKI (2005) Gaussian processes for machine learning. The MIT, Cambridge

Razavi S, Tolson BA, Burn DH (2012) Review of surrogate modeling in water resources. Water Resour Res 48(7):W07401. https://doi.org/10.1029/2011WR011527

Rezania M, Javadi AA, Giustolisi O (2008) An evolutionary-based data mining technique for assessment of civil engineering systems. Eng Comput 25(6):500–517. https://doi.org/10.1108/02644400810891526

Rogers LL, Dowla FU, Johnson VM (1995) Optimal field-scale groundwater remediation using neural networks and the genetic algorithm. Environ Sci Technol 29(5):1145–1155

Roy DK, Datta B (2017a) Fuzzy c-mean clustering based inference system for saltwater intrusion processes prediction in coastal aquifers. Water Resour Manage 31(1):355–376. https://doi.org/10.1007/s11269-016-1531-3

Roy DK, Datta B (2017b) Genetic algorithm tuned fuzzy inference system to evolve optimal groundwater extraction strategies to control saltwater intrusion in multi-layered coastal aquifers under parameter uncertainty. Model Earth Syst Environ. https://doi.org/10.1007/s40808-017-0398-5

Roy DK, Datta B (2017c) Multivariate adaptive regression spline ensembles for management of multilayered coastal aquifers. J Hydrol Eng 22(9):04017031

Roy DK, Datta B (2017d) Optimal management of groundwater extraction to control saltwater intrusion in multi-layered coastal aquifers using ensembles of adaptive neuro-fuzzy inference system. In: World Environmental and Water Resources Congress 2017, American society of civil engineers, May 21–25, 2017 Sacramento, California, USA, pp 139–150

Roy DK, Datta B (2018a) Comparative efficiency of different artificial intelligence based models for predicting density dependent saltwater intrusion processes in coastal aquifers and saltwater intrusion management utilizing the best performing model. Desalin Water Treat 105:160–180

Roy DK, Datta B (2018b) A surrogate based multi-objective management model to control saltwater intrusion in multi-layered coastal aquifer systems. Civ Eng Environ Syst. https://doi.org/10.1080/10286608.2018.1431777

Schapire RE (1990) The strength of weak learnability. Mach Learn 5(2):197–227. https://doi.org/10.1007/bf00116037

Shafer G (1976) A mathematical theory of evidence. Princeton University Press, Princeton

Shiri J, Kişi Ö (2011) Comparison of genetic programming with neuro-fuzzy systems for predicting short-term water table depth fluctuations. Comput Geosci 37(10):1692–1701. https://doi.org/10.1016/j.cageo.2010.11.010

Shu C, Ouarda TBMJ (2007) Flood frequency analysis at ungauged sites using artificial neural networks in canonical correlation analysis physiographic space. Water Resour Res 43(7):W07438. https://doi.org/10.1029/2006WR005142

Sóbester A, Forrester AIJ, Toal DJJ, Tresidder E, Tucker S (2014) Engineering design applications of surrogate-assisted optimization techniques. Optim Eng 15(1):243–265. https://doi.org/10.1007/s11081-012-9199-x

SPM (2016) SPM® (version 8.2), Salford predictive modeller. Salford Systems, San Diego

Sreekanth J, Datta B (2010) Multi-objective management of saltwater intrusion in coastal aquifers using genetic programming and modular neural network based surrogate models. J Hydrol 393(3–4):245–256. https://doi.org/10.1016/j.jhydrol.2010.08.023

Sreekanth J, Datta B (2011a) Comparative evaluation of genetic programming and neural network as potential surrogate models for coastal aquifer management. Water Resour Manage 25(13):3201–3218. https://doi.org/10.1007/s11269-011-9852-8

Sreekanth J, Datta B (2011b) Coupled simulation-optimization model for coastal aquifer management using genetic programming-based ensemble surrogate models and multiple-realization optimization. Water Resour Res 47(4):W04516. https://doi.org/10.1029/2010WR009683

Sreekanth J, Datta B (2011c) Optimal combined operation of production and barrier wells for the control of saltwater intrusion in coastal groundwater well fields. Desalin Water Treat 32(1–3):72–78. https://doi.org/10.5004/dwt.2011.2680

Sreekanth J, Datta B (2014a) Design of an optimal compliance monitoring network and feedback information for adaptive management of saltwater intrusion in coastal aquifers. J Water Resour Plan Manage 140(10):04014026. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000406

Sreekanth J, Datta B (2014b) Stochastic and robust multi-objective optimal management of pumping from coastal aquifers under parameter uncertainty. Water Resour Manage 28(7):2005–2019. https://doi.org/10.1007/s11269-014-0591-5

Sreekanth J, Datta B (2015) Review: simulation-optimization models for the management and monitoring of coastal aquifers. Hydrogeol J 23(6):1155–1166. https://doi.org/10.1007/s10040-015-1272-z

Sugeno M (1985) Industrial applications of fuzzy control. Elsevier Science Inc., New York City

Sugeno M, Yasukawa T (1993) A fuzzy-logic-based approach to qualitative modeling. IEEE Trans Fuzzy Syst 1(1):7. https://doi.org/10.1109/tfuzz.1993.390281

Sun AY, Wang D, Xu X (2014) Monthly streamflow forecasting using Gaussian process regression. J Hydrol 511:72–81. https://doi.org/10.1016/j.jhydrol.2014.01.023

Takagi T, Sugeno M (1985) Fuzzy identification of systems and its applications to modeling and control. IEEE Trans Syst Man Cybern SMC 15(1):116–132. https://doi.org/10.1109/tsmc.1985.6313399

Talbi E-G, Mostaghim S, Okabe T, Ishibuchi H, Rudolph G, Coello Coello CA (2008) Parallel approaches for multiobjective optimization. In: Branke J, Deb K, Miettinen K, Słowiński R (eds) Multiobjective optimization: interactive and evolutionary approaches. Springer, Berlin, pp 349–372. https://doi.org/10.1007/978-3-540-88908-3_13

Tang G, D'Azevedo EF, Zhang F, Parker JC, Watson DB, Jardine PM (2010) Application of a hybrid MPI/OpenMP approach for parallel groundwater model calibration using multi-core computers. Comput Geosci 36(11):1451–1460. https://doi.org/10.1016/j.cageo.2010.04.013

Therrien R, McLaren R, Sudicky E, Panday S (2010) HydroGeoSphere—a three-dimensional numerical model describing fully-integrated subsurface and surface flow and solute transport. University of Waterloo and Université Laval, Canada

Viana FAC, Haftka RT, Steffen V (2009) Multiple surrogates: how cross-validation errors can help us to obtain the best predictor. Struct Multidiscip O 39(4):439–457. https://doi.org/10.1007/s00158-008-0338-0

Wang W-C, Chau K-W, Cheng C-T, Qiu L (2009a) A comparison of performance of several artificial intelligence methods for forecasting monthly discharge time series. J Hydrol 374(3):294–306. https://doi.org/10.1016/j.jhydrol.2009.06.019

Wang WS, Jin JL, Ding J, Li YQ (2009b) A new approach to water resources system assessment—set pair analysis method. Sci China Ser E Technol Sci 52(10):3017–3023

Wolpert DH, Macready WG (1997) No free lunch theorems for optimization. IEEE Trans Evol Comput 1(1):67–82. https://doi.org/10.1109/4235.585893

Yan S, Minsker B (2006) Optimal groundwater remediation design using an adaptive neural network genetic algorithm. Water Resour Res 42(5):W05407. https://doi.org/10.1029/2005WR004303

Yang I-T, Hsieh Y-M, Kung L-O (2012) Parallel computing platform for multiobjective simulation optimization of bridge maintenance planning. J Constr Eng M 138(2):215–226. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000421

Zerpa LE, Queipo NV, Pintos S, Salager J-L (2005) An optimization methodology of alkaline–surfactant–polymer flooding processes using field scale numerical simulation and multiple surrogates. J Petrol Sci Eng 47(3):197–208. https://doi.org/10.1016/j.petrol.2005.03.002

Zhang W, Goh ATC (2016) Multivariate adaptive regression splines and neural network models for prediction of pile drivability. Geosci Front 7(1):45–52. https://doi.org/10.1016/j.gsf.2014.10.003

Dilip Kumar Roy was a Ph.D. student in the discipline of Civil Engineering, College of Science and Engineering, James Cook University, Australia. Currently, he is working as a Senior Scientist in the Irrigation and Water Management Division of Bangladesh Agricultural Research Institute. He completed his bachelor degree in Agricultural Engineering and Master's degree in Irrigation and Water Management from Bangladesh Agricultural University, Bangladesh. Then, he completed another Master's degree in Environmental Technology and Engineering from Ghent University, Belgium. He has authored a number of technical articles in wellknown and reputed peer reviewed international journals and conference proceedings. Recently, he has submitted his Ph.D. dissertation on "Development of a sustainable groundwater management strategy and sequential compliance monitoring to control saltwater intrusion in coastal aquifers" in the College of Science and Engineering, James Cook University, Australia under the guidance of Dr. Bithin Datta.

Dr. Bithin Datta is currently associated with the Discipline of Civil Engineering at James Cook University, Australia. He graduated with a B.Tech. (Hons.) and then Masters degrees in Civil Engineering from internationally renowned IIT Kharagpur in India. He obtained his Ph.D. from the School of Civil Engineering at Purdue University, U.S.A. He worked in different positions in a number of reputed Universities in U.S.A, including University of California, Davis, and University of Washington, Seattle. He then joined the Civil Engineering Department at IIT Kanpur in India, and served in the position of Professor and Senior Professor. He also held the position of Professor and Head of Civil Engineering, and Head of the Environmental Engineering and Management Program. He also served as Visiting Professor at Dalhousie University, Halifax, Canada, Denmark Technical University, Copenhagen, and Asian Institute of Technology, Bangkok. He has guided many Master's and Ph.D. theses. He has authored many Technical Articles in very reputed international Journals, Book Chapters, and numerous Conference papers. He is also internationally well known for his research work on Optimal Identification of Unknown Pollution Sources in Groundwater Systems, design of optimal monitoring networks incorporating uncertainties, management of coastal aquifers for saltwater intrusion management, and real-time operation of reservoir systems.