Development of surrogate for unsaturated flow: model calibration with the monitored infiltration test

R.G.S. Gomes
Pontifical Catholic University of Rio de Janeiro, Rio de Janeiro, Brazil, ruan_gomes93@hotmail.com

G.J.C. Gomes
Federal University of Ouro Preto, Minas Gerais, Brazil, guilhermejcg@yahoo.com.br

E.A. Vargas[1], F.R. Alves[2], R.Q. Velloso[3]
Pontifical Catholic University of Rio de Janeiro, Rio de Janeiro, Brazil, vargas@puc-rio.br[1], felipealvesrosa@gmail.com[2], raquelvelloso@puc-rio.br[3]

ABSTRACT: Water flow analyses under transient soil hydraulic conditions require knowledge of the soil hydraulic properties. These constitutive relationships, named soil-water characteristic curve (SWCC) and hydraulic conductivity function (HCF) generally have several parameters that must be calibrated against collected data. Many of the parameters in SWCC and HCF models cannot be directly measured in field or laboratory but can only be meaningfully inferred from collected data and inverse modeling. In this paper we introduce Evolutionary Polynomial Regression (EPR) as a tool to develop surrogate models of the physically-based unsaturated flow. A rich dataset of soil hydraulic parameters is used to calibrate our surrogate model, and real-world data are then utilized to validate our methodology. Our results demonstrate that the EPR model predicts accurately the observed pressure head data. The model simulations are shown to be in good agreement with the Hydrus software package.

Keywords: unsaturated flow; inverse modeling; surrogate model; evolutionary polynomial regression

1. Introduction

Unsaturated soils are filled with water and air [1]. In such environments, infiltration processes depend critically on the geological properties. The processes determine groundwater flow, subsurface saturation, run-off generation, variably saturated water flow, the shape of the hydrograph and slope stability [2-6]. Thus, a model that describes adequately the water movement in the unsaturated zone is a prerequisite to characterize many different earth surface phenomena.

The water movement in unsaturated soils is generally described by Richards' equation [7]. This water flow model solves the pressure head distribution of the soil, yet it is a highly nonlinear partial differential equation. This requires a detailed computational analysis since analytical solutions are very difficult to derive. Water flow analyses under transient soil hydraulic conditions require knowledge of the soil hydraulic properties. These constitutive relationships, coined soil-water characteristic curve (SWCC) and hydraulic conductivity function (HCF), generally have several parameters that must be calibrated against collected data [8-9].

Field and laboratory measurement methods allow to collect data that can be used to model water flow movement. Such methods include a wide variety of instruments and techniques. Examples of laboratory methods include pressure plates, dew point and filter paper. A common field method is the Guelph permeameter. These latter four techniques make it possible to provide one constitutive relationship (SWCC or HCF), yet inversion methods are required to interpret the observations of hydraulic head. Much effort is required to use collected data to characterize transient behavior of unsaturated soils [10].

Many of the parameters in SWCC and HCF models cannot be directly measured in field or laboratory but can only be meaningfully inferred from collected data. Parameter estimation by inverse modeling provides a simple way to merge observed data and models. The inverse modeling approach has been widely used for variably saturated parameter estimation [11-16]. This body of work has led Velloso [17] to propose a new field test herein called Monitored Infiltration (MI) test. The test consists on measuring pressure head at a specific soil depth during an infiltration process. Observed data can then be inverted in local and global optimization algorithms. The MI test has been used in numerical [18-20] and field studies [19] in Rio de Janeiro, Brazil. While common local optimization algorithms, such as Levenberg-Marquardt [21], might eventually search only around the local minimum [18-20], other state-of-art Bayesian algorithm, such as the Differential Evolution Adaptive Metropolis (DREAM) [22] generally finds global optimal results. However, the Bayesian paradigm needs to run the direct water flow model several times. This can complicate its application to nonlinear, time consuming, numerical models as Richards' equation. Instead, one may give preference to simulate surrogate models, which mimic the output of Richards' model. Surrogate models are simple polynomials that enable inverse modeling techniques to run optimization algorithms at their full potential [20].

In recent years, several computer pattern recognition and data-driven approaches have emerged and developed. Although there are many data-driven techniques, artificial neural networks (ANN) and genetic programming (GP) are most widely used pattern recognition

methods to model complex engineering problems and capture nonlinear interactions between various parameters in a system [23]. In ANN, the user can only analyze inputs and outputs of the simulations. Other disadvantages are related to pre-processing of the data, identification of the optimum structure of the network, large parameterization and overfitting problems [24]. GP algorithms on the other hand are grey-boxes techniques as they provide an analytical expression to represent the system response. However, the principle of parsimony must be critically controlled by including a measure of trade-off between the quality of fit and the model complexity [25-26]. A new data driven technique, Evolutionary Polynomial Regression (EPR) is promising as it can develop surrogate models avoiding the main shortcomings of ANN and GP.

EPR is a grey-box conceptual model [26], with a mathematical structure, derived empirically from physical phenomena. These models are usually polynomials that require parameter estimation during data modeling. The EPR framework merges input and output data to develop transparent and well-structured models. Examples of EPR models in geotechnical engineering include settlement of shallow foundations [27], hydraulic conductivity [28], air permeability of the soil [39], lateral bearing capacity [30], stability of soil and rock slopes [31], pedotransfer functions [32], stress-strain relationships [33-34], mechanical behavior of unsaturated soils [35], constitutive modeling in finite element analyses [23,33-34], optimization of aquifers subjected to sea water intrusion [36] and modeling of soil water characteristic curves [37]. Whereas much work has been made on development and use of EPR models for prediction of geotechnical behavior of soils and rocks, little attention has been given to the unsaturated water flow.

In this paper, we built on the ideas of Giustolisi and Savic [27] and introduce an EPR model for unsaturated flow. We use synthetic and real-world field data to illustrate our method. The proposed EPR modelling framework uses incremental approach [23, 34-35] coupled with Differential Evolution [38] to adequately characterize the highly nonlinear behavior of Richards' equation. The model is calibrated against a rich dataset of soil hydraulic parameters derived using the Rosetta program [39]. Predictions of pressure head with the Hydrus program are used to simulate the infiltration process.

2. Evolutionary Polynomial Regression

The Evolutionary Polynomial Regression is a data-driven method based on evolutionary computing. A general EPR model structure, that can have m different terms, is written as follows:

$$\boldsymbol{y} = a_0 + \sum_{j=1}^{m} F\left(\boldsymbol{X}, f(\boldsymbol{X}), a_j\right) \quad (1)$$

where $\boldsymbol{y}$ stores the estimated vector of target values; a_0 signifies bias term, a_j is an adjustable parameter for the j[th] term; F is a function constructed by the process; $\boldsymbol{X}$ is the matrix of input independent variables; f represents an optional function given by the user. The main goal of EPR is to search for the best model structure denoted by Equation 1. This process involves calibration against observed data, $\tilde{\boldsymbol{y}}$.

In the classical EPR procedure [26], genetic algorithm is used to find feasible structures of Equation 1 while the adjustable parameters are computed by means of the linear least squares. In such technique, the sum of squared errors (SSE) is minimized as cost function, given as:

$$\text{SSE} = \frac{\sum_{i=1}^{N}(y_i - \tilde{y}_i)^2}{N} \quad (2)$$

where N is the number of data points. A detailed description of EPR appears in [26] and thus will not be repeated herein.

In this paper, we derive EPR using Differential Evolution (DE) [38] for model structure global exploration. The developed DE algorithm, quickly converges to the global optimum, is relatively easy to use and have few control variables [40]. Details on the developed DE-based EPR algorithm are outside the scope of this work and will be discussed in a further publication. We now present the basic building blocks of our methodology.

3. Methodology

This work methodology uses the iterative research cycle approach [22]. In order to understand the estimation of unsaturated parameters all five components of our framework, Fig. (1), must evolve at the same rate. In other words, the more accurate the field data and model assumptions are, it is to be expected that good quality estimations are obtained. Aditionaly, a good estimation is obtained if an adequate optimization algorithm is use, and hence if an adequate surrogate model is used (if the optimization strategy requires one). Therefore, this approach lets the user question which of the steps might need atention. This research aimes primarly in applying EPR procedure to model unsaturated flow, consequently, the main concern herein is to adequadly provide good informations from itens shown in Fig. (1.a) until Fig. (1.d).

3.1. Data measurements

The first step is to collect data information. The Monitored Infiltration test, Fig. (1.a), is used to collect the system response. The test consists of a simple geometry of a open circular pit, 20 cm deep and 16 cm wide. A tensiometer instrument measures the pressure head changes as the test is on course of action. This device should be placed right on the axis of symmetry. Also, to obtain the desired response it is necessary to have on hand a Mariotte type bottle device to keep the hydraulic head constant during the test. Instantly after applying the hydraulic head, the water will start the infiltration process. When the infiltration front reaches the tensiometer ceramic, the saturation process will start, until the pressure head reaches constant values. When no more significant changes in pressure heads are observed, the test can be considered as finished. After gathering field information, it is necessary to collect disturbed and undisturbed soil samples in order to characterize

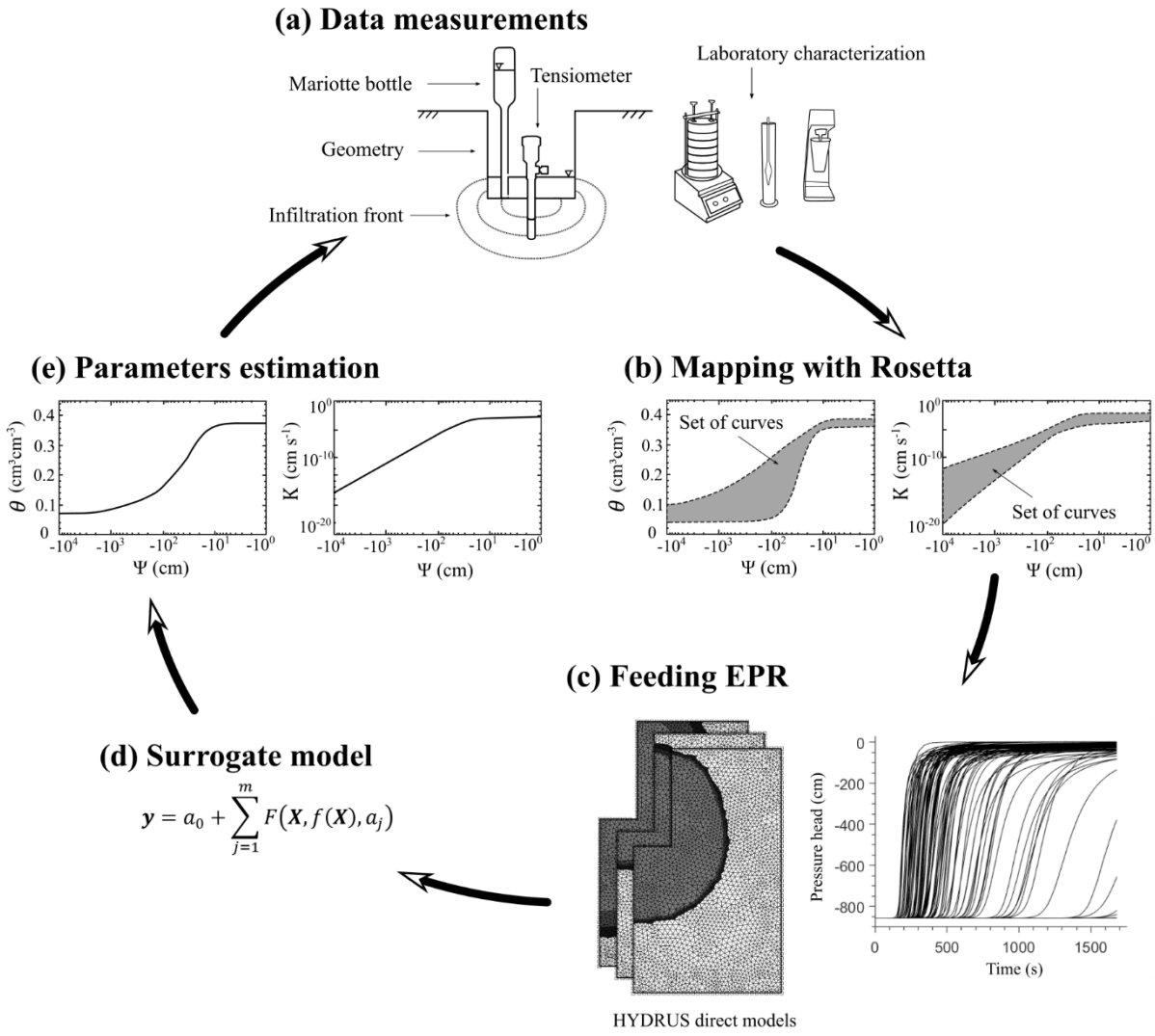

Figure 1. Methodology used to calibrate a surrogate model with the monitored infiltration test. (a) Data collection through field and laboratory measurements; (b) Parameters mapping through laboratory results and Rosetta prediction; (c) A series of forward models to generate a predetermined number of pressure head versus time curves; (d) The EPR algorithm uses the synthetic curves to calibrate a model; (e) Parameters estimation through inverse modelling.

its physical properties. With the soil specifications made through laboratory tests it is possible to generate, through the developed Rosetta program [39], different groups of hydraulic parameter that maps and represents candidate parameters.

3.2. Mapping soil parameters with Rosetta

Step two shown in Fig. (1.b), consists on mapping the unsaturated hydraulic parameters that best honors the soil characteristic. This process is achieved by defining ranges of textural contents which will feed the Rosetta pedotransfer function. The final product of this step provides a range of predetermined set number of the van Genuchten parameters, as in Eq. (2). This set should assume that the optimum soil parameters is placed within its range.

$$\boldsymbol{x}_{ix5} = [\boldsymbol{\theta_r}\ \boldsymbol{\theta_s}\boldsymbol{\alpha}\ \boldsymbol{n}\ \boldsymbol{K_{sat}}] \quad (2)$$

where $\boldsymbol{x}$ is a matrix of the $i = 1,2,\dots,n$ predetermined groups of parameters; $\boldsymbol{\theta_r}$ [cm³ cm^{-3}] is a column vector of different residual water content; $\boldsymbol{\theta_s}$ [cm³ cm^{-3}] is a vector of predicted saturated water content; $\boldsymbol{\alpha}$ [cm^{-1}] and $\boldsymbol{n}$ [-] are vectors of the model's adjustable parameters; and $\boldsymbol{K_{sat}}$ [cm s^{-1}] is a vector of the hydraulic conductivity parameter at saturation.

3.3. Feeding EPR with HYDRUS

Step three, Fig. (1.c), uses HYDRUS 2D/3D comercial finite element program developed by PC-Progress company to generate curves of synthetic data. The HYDRUS program, together with Rosetta predictions, will feed the EPR algorithm into developing the surrogate model. This procedure is done by introducing each group of parameter and numerical specifications (i.e mesh, observation point, initial and boundary conditions), to HYDRUS direct mode calculations. The output of this process will provide synthetic pressure head versus time response associated to each group of parameters. The numerical specifications for the MI test, shown in Fig. (2), will be described. The type of geometry is a 2D axissymetric vertical flow with element size corresponding to 1 cm for the hole domain. Boundary

conditions are: constant head placed inside the open pit, atmospheric boundary on top of the geometry and no flux condition in the remaining of the geometry. Although the no flux condition may not maintain faithfullness of field conditions, the geometry was fixed to a size were computational effort would not be costly and the infiltration front would not reach the impervious contour. If the infiltration front reaches the contour, pressure head acummulation might be encoutered, and therefore the boundary must be modified to overcome this deficiency.

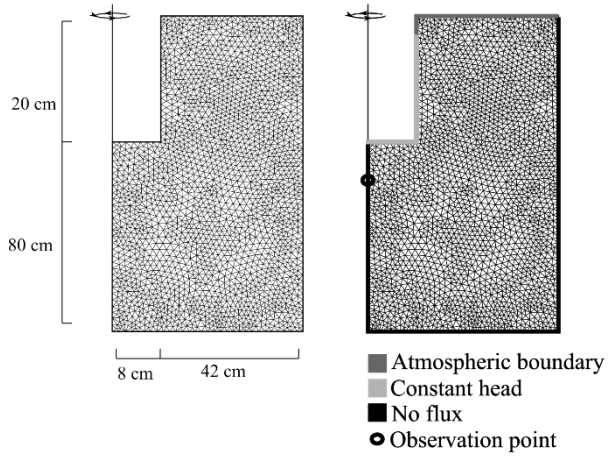

Figure 2. Numerical settings: type and size of geometry, element sizes and boundary condition (Out of scale).

3.4. Surrogate model development

Step four, Fig. (1.d), aimes to introduce a vector of all the sinthetic data and a matrix of the corresponding independent variables to the evolutionary process. With this information and EPR settings this technique will approximate an analytical equation through what is known as symbolic regression.

Before starting the symbolic regression, there are a number of constraints to take into account, these are: select the optimization strategy, type of function *f*, select independent variables, range of exponents, number of terms, number or populations and generations.

The single-objective optimization strategy guides the algorithm towards the best fitting function by comparing the observation data with prediction data. It is the simplest strategy to implement. However, all additional model information should be chosen before applying the regression. By applying multi-objective optimizations [41], the number of terms, noise control and choice of the best function, based on a fitness function, can be chosen as the EPR evolution proceeds. This procedure, is considered as an advantage by the fact that the global optimum structure may reveal itself. However, due to the simpler EPR implementation, a single-objetive optimization was explored in this study.

Although a general equation would be more suitable for researchers interested in applying the surrogate model for any field condition, the computational effort did not match the capability of i7-7700 CPU @ 3.6GHz and 16.GB of memory machine used herein. Therefore, the choice of the independent variables is based on computational performance. It is worth meationing that previous research using incremental approaches [23, 34-35] have shown that this technique is suitable for applications with non linear behavior and allows a point-by-point construction of the entire curve. In this approach the model is calibrated with past information of the dependent variable, which ables the algorithm to capture the curve's shape. Rezania [23] shows that if the optimum mathematical structure is unknown, a pure polynomial structure shows fitting values equal or better than pseudo-polynomial functions. In addition, a pure polynomial structure acelerates the speed of the regression. Therefore, no function was used in the present work. Equation (3) presents the selected polynomial structure for the present work.

$$\boldsymbol{Y} = a_0 + \sum_{j=1}^{m} a_j \cdot (\boldsymbol{X}_1)^{\boldsymbol{ES}(j,k)} \cdot \ldots \cdot (\boldsymbol{X}_k)^{\boldsymbol{ES}(j,k)} \qquad (3)$$

where $\boldsymbol{Y}$ is the estimate target value; m is the number of terms of the target expression; a_0 is the bias term and a_j is the adjustable parameter for the j[th] term; $\boldsymbol{X}$ represents the matrix of input for k independent variables $\boldsymbol{X} = [\boldsymbol{X}_1 \cdot \ldots \cdot \boldsymbol{X}_k]$; $\boldsymbol{ES}_{mxk}$ is the matrix of exponents whose elements can assume values within user-define-bounds. The target value in this work is the pressure head $\boldsymbol{\Psi}_{i+1}$ (cm) referred to the subsequent time step prediction for $i = 0,1,2,\ldots,N$ records. The adopted independent matrix is $\boldsymbol{X} = [\boldsymbol{t}_i\ \boldsymbol{\theta}_r\ \boldsymbol{\theta}_s\ \boldsymbol{\alpha}\ \boldsymbol{n}\ \boldsymbol{K}_{\boldsymbol{sat}}\ \boldsymbol{\Psi}_i]$ where $\boldsymbol{t}_i$ (s) is the current time step; $\boldsymbol{\theta}_r$, $\boldsymbol{\theta}_s$, $\boldsymbol{\alpha}$, $\boldsymbol{n}$ and $\boldsymbol{K}_{\boldsymbol{sat}}$ are the van Genuchten parameters and $\boldsymbol{\Psi}_i$ is the current pressure head (cm). Equation (4) shows the input matrix for one pressure head versus time curve, which in order to perform the regression must include all i curves. The exponent values where defined between $[-3; -2; -1; 0; 1; 2; 3]$. No decimal exponents, such as 0.5 and 1.5, where considered. This avoids equations with complex number (the association between a negative pressure head, Ψ_i and decimal exponent).

$$\boldsymbol{X} = \begin{bmatrix} t_1 & \theta_r & \theta_s & \cdots & \Psi_0 \\ t_2 & \theta_r & \theta_s & \cdots & \Psi_1 \\ t_3 & \theta_r & \theta_s & \cdots & \Psi_2 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ t_i & \theta_r & \theta_s & \cdots & \Psi_i \end{bmatrix}$$

$$= [\ \boldsymbol{t}_i \quad \boldsymbol{\theta}_r \quad \boldsymbol{\theta}_s \quad \cdots \quad \boldsymbol{\Psi}_i\ ] \qquad (4)$$

After adjusting the EPR setting, the algorithm proceeds to the regression. The data is divided into training (calibration and evaluation) and validation. On the training data, 75% of data calibrates the model, the remaining 25 % of the data evaluates the model. The validation set, corresponds to one selected curve that will appraise the generalization capability of the model to an unseen simulation. Both evaluation and validation sets are in the range of the calibration data, because EPR is good at predicting data from interpolation, but not so effective for extrapolation of model's range[34].

As the EPR procedure starts, Rezania [23] has shown that by increasing the number of evolutions the equations gradually picks up the different participanting parameters in order to form equation representing the constitutive

model. The level of accuracy at each evolution is measured using sum of squared errors until the stop criterion is reached, in this case, a predetermined number of generations.

After revealing the best equation, just as in EPR incremental approaches [23, 34-35], a point-by-point construction of the entire curve is made through a loop. Figure. (3) demonstrates this procedure. For this approach the initial condition $\Psi_{i=0}$ for the initial time $t_{i=1}$ is known, consequently by introducing this value in the equation the next pressure head is calculated through each loop, until the last time step is computed.

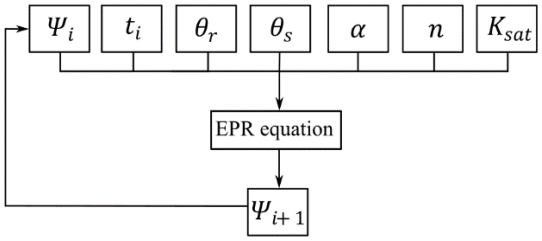

Figure 3. Schematic sketch of equation loop.

3.5. Inverse analyses

The last step, number five is showed in Fig. (1.e). It consists on applying the inverse modelling to observed data. This step consideres two approaches, that are illustrated in Fig. (4).

First, the EPR equation goes through inverse analyses with the Levenberg-Marquardt optimization algorithm in order to estimate the parameters. This will be executed with the Matlab 2019.a software package and the Lsqnonlin function. This procedure consists on supplying the Lsqnonlin function with the observation data, EPR equation, initial condition $\Psi_{i=0}$, and a initial guess for the parameters. The Lsqnonlin will use the Levenberg-Marquardt algorithm to estimate the parameters that produces the best fit between the observed data and the EPR predictions. In Matlab, the Levenberg algorithm does not handle boundary constraints. In other words, the user cannot limit the search space. Therefore, the initial estimate must be as accurate as possible, so that the algorithm, from the phenomenon point of view is able to produce acceptable parameters. After inverse analysis, the estimated parameters will be introduced to HYDRUS 2D direct model to best represent the prediction.

For comparison pourposes, the available HYDRUS 2D inverse option will also be executed. For this numerical optimization, the program allows boundary constraints, so it is possible to be not so strict regarding initial estimates. The numerical solution of the inverse problem requires field observations and numerical informations: geometry, boundary conditions, initial conditions and initial estimatives, which must be identical to assay setting in order to correctly represent field observations. Before the last parameters estimation, as a subrotine procedure, the Hydrus performs the direct model to compare the observed and predicted data. This information, along with the estimated parameters allows the user to visualize how much the algorithm has been able to adjust.

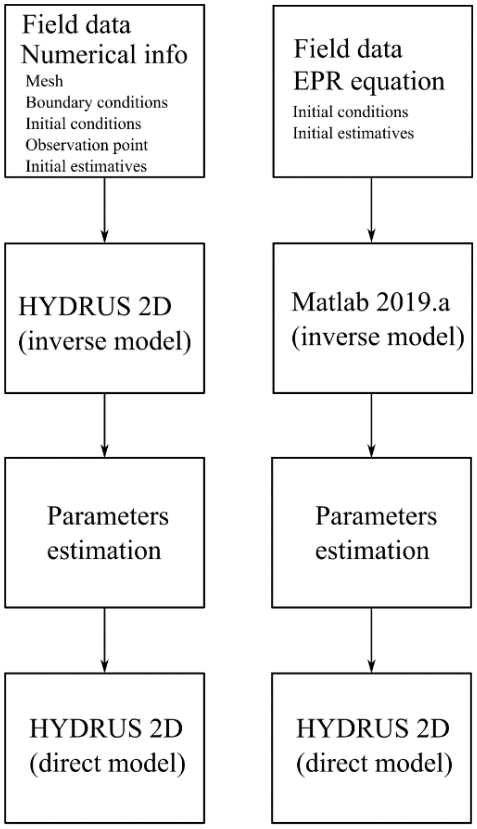

Figure 4. Flow diagram of inverse analyses.

Even though the interpretation of the estimates are a crutial part of the inverse analyses, they are outside the scope of this work.

4. Results

4.1. Synthetic data

The purpose of this example is to apply the metodology described in section 3 and study the EPR modeling and optimization capacity in a controlled scenario. Where the optimum parameters are known. In this perspective, a group of hydraulic parameters was randomly chosen from a Rosetta prediction for sandy soils to represent a hypothetical measurement.

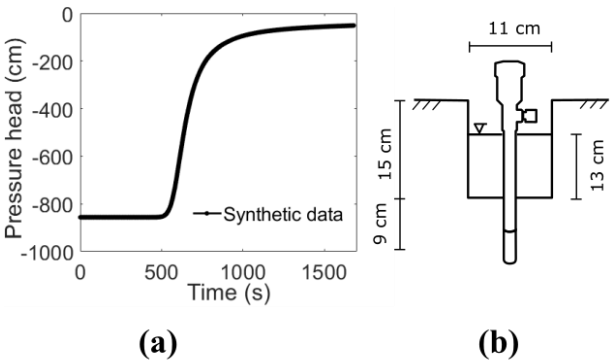

Figure 5. Synthetic data whose unsaturated hydraulic parameters are known. (a) MItest response. (b) MItest field specifications.

Figure. (5.a) illustrates the synthetic field data. The MI test setting is presented in Fig. (5.b). It took approximately 500 seconds for the infiltration front to reach the observation point and from there 1180 seconds for field saturation. It can be seen that at the end of the assay there is a residual pressure head, which suggests that the boundary condition was not able to remove all the air fluid from the pores.

In order to map the hydraulic predictions, the Rosetta program searched for 100 hydraulic parameter sets for ranges between 80% and 100% of sand content and 0 to 20% for silt content. After the predictions, all the parameters where introduced to HYDRUS 2D forward model to generate pressure head versus time curves. Figure. (6) demonstrates that the synthetic data is placed within the predicted range, consequently EPR will interpolate and not extrapolate the optimum parameters.

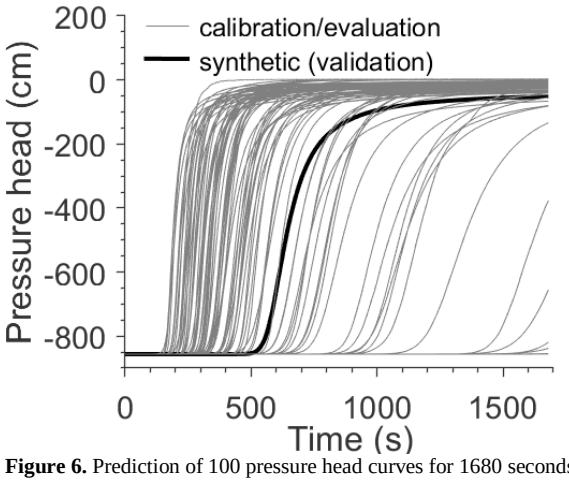

Figure 6. Prediction of 100 pressure head curves for 1680 seconds time-lapse.

The input data from the previous step was introduced to the EPR. The matrix of inputs was identical to Eq. (4). All curves considers a total time of 1680 seconds divided in equal time steps of 1 second. This regression was setted to: 3 polynomial terms; a population of 90 potential fitting exponents whose values where described in subsection 3.4. The stop criteria (number of generations) considers one hundred evolutions. Table 1 presents the statistical and computational performance of the regression. It can be seen by the statistics and sum of squared errors (SSE) that the EPR adjusts a equation with great accuracy. The computational performance also shows great performance for this study case. We present in Eq. (5) the generated equation. It is worth meationing that the independent variable, pressure head Ψ_i, was divided by 1000 because the equation can present convergence problems while executing the loop.

$$\Psi_{i+1} = 0.3487 - 53.5158 \cdot \frac{t \cdot n \cdot {K_{sat}}^2 \cdot (\Psi_i/1000)}{{\theta_s}^2 \cdot \alpha} - 1.0204 \cdot 10^{-9} \cdot \frac{t^2}{{\theta_s}^2 \cdot \alpha} + 1000.7922 \cdot (\Psi_i/1000) \quad (5)$$

Table 1. EPR performance for synthetic data.

Data set	Metric	Values
Calibration	R^2	0.99
	RMSE (cm)	1.29
Validation	R^2	0.99
	RMSE (cm)	0.90
Sum of Squared Erros (SSE)		1.67
Computational time (min)		18.23

Equation (5) was introduced to Matlab for inverse procedure. Velloso [12] recomends, for numerical inverse procedure, that saturated and residual moisture contents are to be fixed during inversion, the reason being that this two parameter are highly correlated in the van Genuchten model. As a result, several combinations between them satisfies the best fit, thus the optimum parameters might never be reached. Therefore, HYDRUS optimization throught the Levenberg-Marquardt algorithm and van Genuchten model, estimates only three parameters. The same approach is adopted for EPR optimization. Figure (7) presents the graphical results of the inverse analyses. By the fact that the synthetic data and the inverse procedure comes from the same model, i.e HYDRUS, the inverse model presents very good fit between the synthetic data and the estimated data. However, it can be seen from the second column of table (2), which shows the estimated parameters, that the optimum parameters where not found. This fact can be assigned to the correlation between the parameters, therefore optimum parameters might never be reached. The EPR on the other hand did not show the best fit between observed and estimated datasets. However, some conclusions can be drawn. First of all, the arrival of the infiltration front was correctly estimated. In second place, the time it took for the tensiometers ceramic capsule to saturate was correcly estimated and finally the EPR overestimated the residual pressure head. Even though, the best fit was not reached a good agreement was found between observed and estimated patterns.

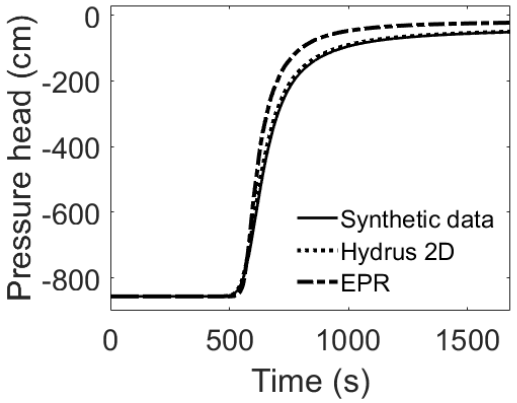

Figure 7. Graphical results of inverse analyses for synthetic data.

Table 2 presents the estimated parameters. As it can be seen the HYDRUS presented the best parameters because the synthetic data and the inversion procedure comes from the same governing model solution. However, the optimum parameters where not found. The EPR overestimated all the parameters, but capture the desired phenomenon. In the computational aspect, table

(2) reinforces the EPR capability of outperforming the HYDRUS inversion procedure.

Table 2. Synthetic inversion results.

Parameter	Data	HYDRUS	EPR
θ_r	0.0379	-	-
θ_s	0.3901	-	-
α	0.0134	0.0139	0.0303
n	2.3870	2.6035	3.6183
K_{sat}	0.000243	0.000236	0.000308
Computational time (min)		8	0.04

4.2. Field data

Figure. (8.a) illustrates real field mesurement whose MI test setting is presented in Fig. (8.b). The soil at site was classified as residual soil of high strenght. This factor made it difficult to drill the open pit and to introduce the tensiometer. Thus the test settings where adapted to match field conditions. It took approximatly 30 seconds for the infiltration front to reach the observation point and from there 140 seconds for field saturation. It can be seen that at the end of the assay there is no residual pressure head, which suggests that full saturation was reached.

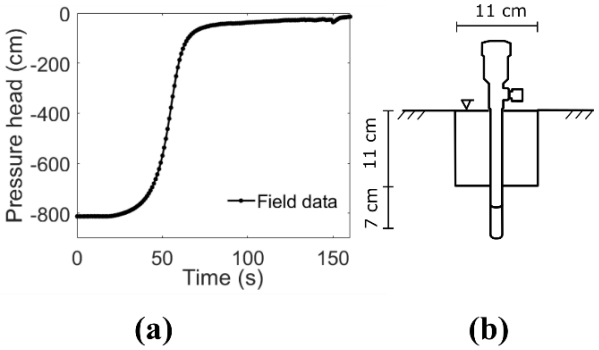

Figure 8. MItest measurement on a residual soil test site whose un-saturated hydraulic parameters are unknown (a) MItest response (b) MItest field specifications

Through disturbed and undisturbed soil samples it was possible to obtain six of the most important soil characteristics for MI test inverse analyses. Table (3) shows the soil textural contents and physical indices, where is the ρ_d bulk density, $\emptyset$ is the porosity and ω the soil moisture.

Table 3. Textural contents and physical indices for the field data

Textural contents			Physical indices		
Sand (%)	Silt (%)	Clay (%)	ρ_d (g/cm³)	$\emptyset$ (%)	ω (%)
65.5	14.3	20.2	1.75	33.50	6.94

Next step consisted on mapping the soil hydraulic parameters through Rosetta. The textural contents were introduced to the developed Rosetta. After running the HYDRUS curves, through the numerical foward model, the field data fell out of predicted range. Therefore textural content alone were not enough to map the parameters in this case. Khoshkroudi et.al [32], shows that better predictions are obtained with addition of bulk density. Therefore, HYDRUS was used for preliminary mapping of the parameter, because it has a coupled Rosetta program that allows predictions with more variables than textural contents. The textural classes along with the bulk density were introduced to the HYDRUS-Rosetta. Results from predictions showed that the porosity (0.33), measured in laboratory, and the predicted saturated moisture content (0.33) were identical. It is known that at saturation, both values are theoretically the same. Consequently it could be assumed that better predictions are obtained through PTFs with textural classes and Bulk density. In this sense, with the HYDRUS-Rosetta prediction, it enables to set a range of search, shown in table (4). This range was introduced to the Rosetta developed in this research, were it mapped 100 sets of parameters. Figure (9) demonstrates that this procedure showed reliable predictions even though some adjustments where needed.

Table 4. Using HYDRUS Rosetta to overcome the developed Rosetta limitations.

Rosetta type		θ_r	θ_s	α	n	K_{sat}
HYDRUS		0.05	0.33	0.03	1.24	0.002
Devel -oped	Max	0.08	0.36	0.05	3.00	0.010
	Min	0.01	0.30	0.01	1.10	0.001

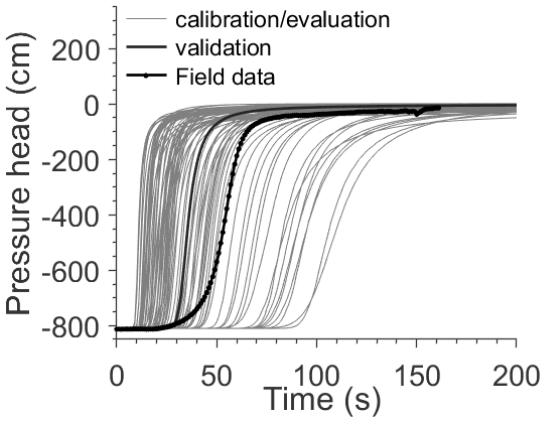

Figure 9. Prediction of 100 pressure head curves for 200 seconds time-lapse.

The input data from the previous step was introduced into EPR. The regression settings were similar as for the sythetic data. The differences included: a polynomial length of five terms. Table 5 presents the statistical R^2 and RMSE parameters and computational performance of the regression. It can be seen by the statistics and SSE that the EPR adjusts a equation with good accuracy. The validation set shows that the EPR prioritized for curves that began to saturate from 10 seconds until 50 seconds, where the highest amount of data is. The computational performance shows excelent performance for this study case. The elapsed time (2.37 min) proved to be much faster than for the synthetic case (18.23 min) because each curve on the synthetic dataset had approximatly 10

times more datapoints. In other word, the synthetic data had 100 curves and 1680 datapoints for each curve (a total of 168,000 data points) whereas the field dataset had 100 curve and 200 seconds for each curve (a total of 20,000 data points). Therefore, it took longer for algorithm to perform the regression for the synthetic data even though it had smaller number of polynomial terms. Equation (6) presents the generated equation, were the variable Ψ_i was divided by 1000 to avoid convergency problems during the loop.

$$\Psi_{i+1} = 2.1219 - 0.6005 \cdot \frac{t \cdot K_{sat} \cdot n^2 \cdot (\Psi_i/1000)}{\theta_s \cdot \alpha} + 0.0277 \cdot \frac{{\theta_r}^2 \cdot {\theta_s}^2 \cdot n \cdot (\Psi_i/1000)}{{K_{sat}}^2} - 2.0930 \cdot 10^{-4} \cdot \frac{\alpha^2 \cdot {K_{sat}}^2}{\theta_r \cdot \theta_s \cdot n^3 \cdot (\Psi_i/1000)^2} + 1003.4633 \cdot (\Psi_i/1000) \quad (5)$$

Table 5. EPR performance for field data

Data set	Metric	Values
Calibration	R^2	0.99
	RMSE (cm)	14.17
Validation	R^2	0.99
	RMSE (cm)	9.20
Sum of Squared Erros (SSE)		157.67
Computational time (min)		2.37

Next step was to execute both inversions. Both inversion included fixed saturated and residuals water contents, which were estimated by the HYDRUS-Rosetta which is shown in table (4). Figure (10) presents the graphical results of the inverse analyses. EPR shows that the predictions underestimated the arrival of the infiltration front and pressure head at saturation, however it estimated correctly the time it took for the tensiometer's ceramic capsule to saturate. HYDRUS on the other hand, had similar results, which underestimated the infiltration front arrival but it estimated correclty the time for the ceramic to saturate. The HYDRUS inversion estimated the residual pressure head with better accuracy if compared to the EPR estimation.

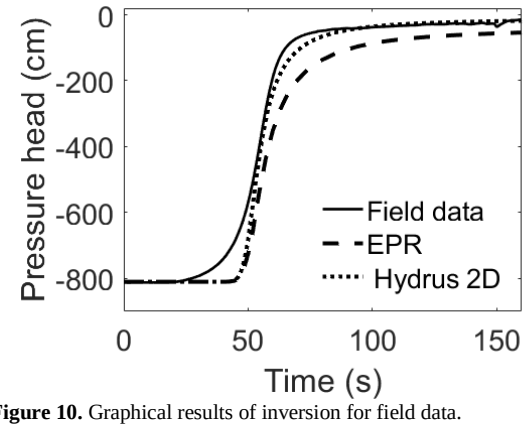

Figure 10. Graphical results of inversion for field data.

Table 6 presents the parameters estimations. EPR inversion showed smaller hydraulic conductivity coeficient K_{sat} and smaller α coeficient. Even though it provides smaller hydraulic conductivity and larger air entry values if compared to HYDRUS prediction, no realiable conclusions can be drawn from their direct comparison and therefore it would be more appropriate to compare this results with laboratory measurements. In the computational aspect, table (6) reinforcers that EPR inversion outperforms the HYDRUS inversion procedure.

Table 6. Field data inversion results, the fixed parameters and values were: θ_r=0.0501 and θ_s=0.3329.

Parameter	HYDRUS	EPR
θ_r	-	-
θ_s	-	-
α	0.0325	0.0132
n	2.6901	3.3020
K_{sat}	0.0026	0.0013
Computational time (min)	5	0.02

5. Conclusions

We have introduced a metodology to calibrate surrogate models for unsaturated flow. Our methodology uses field (MI test) and laboratory measurements (characterization tests) to map hydraulic parameters through Rosetta pedotransfer functions. This enables to build a surrogate model which the optimum parameters are within the models range. Therefore inversion procedure doesn't have to extrapolate the model's range.

A surrogate model was explored through EPR methodology. The EPR shows great potential at modelling unsaturated flow and therefore providing an alternative solution to overcome computational cost raised by bayesian inverse procedures. From the modelling point of view the incremental approach reveals to be suitable for this application because it provides an independent variable (Ψ_i) which captures the curve's behaviour. This enables a point-by-point construction of the entire pressure head path from an imbibition condition. From a computational prespective, calibrating a surrogate for each test ensures great computational performance. In this case both analyses used a low number of curves (100 sets), evolutionary population (90), generations (100) and number of terms, which ensured good performance. Therefore, a general equation might be computationaly time-consuming because of the number of datas needed. For instance, if a general equation would be considered with: 100 sets of parameters, 20 initial conditions, 10 boundary conditions, 10 observations points and a total time of 1000 seconds for each curve, the data set length would be of 20,000,000 points, which excedes by far what has been applied in the present work.

We also have introduced inversion analyses. The results shows that EPR model captures the underlying phenomenon reasonably. In both applications, EPR surrogates, have captured the infiltration front arrival and saturation time reasonably well. However it overestimated the residual pressure head. The computational time showed great performance, as it was expected.

For future researches we highlight three recommendations for EPR procedure: use of a multi-objective approach in order to estimate optimal strutures [41], evaluate the impact of pseudo-polynomials on the model accuracy and extend the model to capture spatial variables. Finally we recommend other data-driven techniques in order to capture the desired phenomenon.

Acknowledgement

The project presented in this article is supported by the Brazilian National Council for Scientific and Technological Development, CNPq (*Conselho Nacional de Desenvolvimento Científico e Tecnológico*).

References

[1] Freeze, R. A., Cherry, J.A. "Groundwater", 1st ed., Prentice-Hall, New Jersey, United States, 1979.

[2] Todd, D. K., Mays, L.W. "Groundwater Hydrology", 3rd ed., John Wiley and Sons, United States, 2005.

[3] Fitts, C. R. "Groundwater Science", 2nd ed., Elsevier, United States, 2012.

[4] Fetter, C.W. "Applied Hydrogeology", 4th ed., Pearson, United States, 2014.

[5] Gomes, G.J.C., Vrugt, J.A., Vargas Jr., E.A. "Toward improved prediction of the bedrock depth underneath hillslopes: Bayesian inference of the bottom-up control hypothesis using high-resolution topographic data." Water Resources Research, 52, 2016. http://dx.doi.org/10.1002/2015WR018147.

[6] Gomes, G.J.C., Vrugt, J.A., Vargas Jr., E.A., Camargo, J.T., Velloso, R.Q., van Genuchten, M.Th. "The role of uncertainty in bedrock depth and hydraulic properties on the stability of a variably-saturated slope", Computers and Geotechnics, 88, 2017. http://dx.doi.org/10.1016/j.compgeo.2017.03.016.

[7] Richards, L.A."Capillary Conduction of Liquid Through Porous Mediums", Physics, 1, 1931. https://doi.org/10.1063/1.1745010

[8] Brooks, R.H., Corey, A.T. "Hydraulic Properties of Porous Medium", Hydrology paper, 3, 1964.

[9] van Genuchten, M.Th. "A closed-form equation for predicting the hydraulic conductivity of unsaturated soils", Soil Science Society of America Journal, 44, pp. 892-898, 1980. https://doi.org/10.2136/sssaj1980.03615995004400050002x.

[10] Scharnagl, B., Vrugt, J.A., Vereecken, H., Herbst, M. "Inverse modelling of in situ soil water dynamics: investigating the effect of different prior distributions of the soil hydraulic parameters. " Hydrol Earth Syst Sci. 15, 2011. http:// dx.doi.org/10.5194/hess-15-3043-2011.

[11] Zachmann, D.W., DuChateau, P.C., Klute, A. " The Calibration of the Richards Flow Equation for a Draining Column by Parameter Identification", Soil Sci. Soc. of Am. J., 45, pp. 1012-1015, 1981. http://doi.org/10.2136/sssaj1981.03615995004500060002xLA

[12] Zachmann, D.W., DuChateau, P.C., Klute, A. " Simultaneous Approximation of Water Capability and Soil Hydraulic Conductivity by Parameter Identification 1", Soil Sci. Soc. of Am. J., 134(3), pp. 157-163, 1982. http://doi.org/10.1097/00010694-198209000-00002

[13] Kool, J.B., Parker, J.C., Van Genuchten, M.T. " Determining Soil Hydraulic Properties from One-Step Outflow Experiments by Parameter estimation: I. Theory and Numerical Studies", Soil Sci. Soc. of Am. J., 49(1), pp. 1348-1354, 1985. http://doi.org/10.2136/sssaj1985.03615995004900060004x

[14] van Dam, J.C., Stricker J. N. M., Drooger, P. " Inverse Method for Determining Soil Hydraulic Functions from One-Step Outflow Experiments", Soil Sci. Soc. of Am. J., 56, pp. 1042-1050, 1992. http://doi.org/10.2136/sssaj1992.03615995005600040007x

[15] Hudson, D.B., Wierenga, P.J., Hills, R.G." Unsaturated Hydraulics Properties from Upward Flow into Soil Cores", Soil Sci. Soc. of Am. J., 60, pp. 388-396, 1996. http://doi.org/10.2136/sssaj1996.03615995006000020009x

[16] Simunek, J.J., Kodesova, R., Gribb, M., van Genuchten, M.T. " Estimating Hysteresis in the Soil Water Retention Function from Cone Permeameter Experiments", Water Resourcers Research, 35, pp. 1329-1345, 1999.http://doi.org/10.1029/1998WR900110

[17] Velloso, R.Q."Estudo Numérico da Estimativa de Parâmetros Hidráulicos em Solos Parcialmente Saturados" (Numerical study on the Estimation of Hydraulic Parameters on Partially Saturated Soils), Master thesis, Pontifical Catholic University of Rio de Janeiro, Department of Civil Engineering, 2000. (in Portuguese).

[18] Morales, M.S.T."Estudo Numérico e Experimental de Problemas de Fluxo Saturado-Não Saturado em Solos" (Numerical and Experimental study on Saturated and Unsaturated Flux Problems), Master thesis, Pontifical Catholic University of Rio de Janeiro, Department of Civil Engineering, 2008. (in Portuguese).

[19] Pinto, J.L.T.M.G. "Determinação das Propriedades Hidráulicas de Solos Residuais do Rio de Janeiro" (Determination of the Hydraulic properties of Residual Soils from Rio de Janeiro), Master thesis, Pontifical CatholicUniversityof Rio de Janeiro, Department of Civil Engineering, 2013. (in Portuguese).

[20] Alves, F.R. "Um Estudo de Procedimentos Numéricos e Experimentais Para Uso no Ensaio de Infiltração Monitorada" (A Numerical and Experimental Method to Apply in the Infiltration Monitored Teste), Master thesis, Pontifical Catholic University of Rio de Janeiro, Department of Civil Engineering, 2017. (in Portuguese).

[21] Marquardt, D.W. "An algorithm for least-squares estimation of nonlinear parameters", J. Soc. Indust Ap. Math, 11, pp. 431-441, 1963. https://doi.org/10.1137/0111030

[22] Vrugt, J. "Markov chain Monte Carlo simulation using the DREAM software package: Theory, concepts, and MATLAB implementation", Envir. Modelling & Software, 75 pp. 273-316, 2016. https://doi.org/10.1016/j.ensoft.2015.08.013

[23] Rezania, M. "Evolutionary Polynomial Regression Based Constitutive Modelling and Incorporation in Finite Element Analysis", Ph. D. thesis, University of Exeter, School of Engineering, Computing and Mathematics, 2008.

[24] Giustolisi, O. "Some techniques to avoid overfitting of artificial neural networks", 5th Conf. Hydroinformatic, IWA publishing London, UK, vol.2, pp. 1465-1477, 2002.

[25] Davidson, J.W., Savic, D., Walters, G.A. "Method for identification of explicit polynomial formulae for the friction in turbulent pipe flow", Journal of Hydroinformatics, 1(10), pp. 115-126. 1999. https://doi.org/10.2166/hydro.1999.0010

[26] Giustolisi, O., Savic, D. "A symbolic data-driven technique based on evolutionary polynomial regression", Journal of Hydroinformatics, 8, pp 207-222, 2006. https://doi.org/10.2166/hydro.2006.020b

[27] Shahin, M.A. "Use of evolutionary computing for modelling some complex problems in geotechnical engineering", Geomechanics and Geoengineering, 10(2), pp. 109-125. 2015. https://doi.org/10.1080/17486025.2014.921333

[28] Ahangar-Asr, A., Faramarzi, A., Mottaghifard, N., Javadi, A. "Modeling of permeability and compaction characteristics of soils using evolutionary polynomial regression", Computers & Geosciences, 37, pp. 1860-1869. 2011. https://doi.org/10.1016/j.cageo.2011.04.015

[29] Ahangar-Asr, A., Javadi, A. "Air losses in compressed air tunneling: A prediction model", Engineering and Computational Mechanics, 169, pp. 1-8. 2016. https://doi.org/10.1680/jencm.15.00023

[30] Ahangar-Asr, A., Javadi, A.A., Johari, A, Chen, Y. "Lateral load bearing capacity modelling of piles in cohesive soils in undrained conditions: An intelligent evolutionary approach", Applied Soft Computing, 24, pp. 822-828. 2014. https://doi.org/10.1016/j.asoc.2014.07.027

[31] Ahangar-Asr, A., Faramarzi, A., Javadi, A."A new approach for prediction of the stability of soil and rock slopes", Engineering Computations, 27(7), pp. 878-893. 2010. https://doi.org/10.1108/02644401011073700

[32] Khoshkroudi, S.S., Sefidkouhi, M.A.G., Ahmadi, M.Z., Ramezani, R. "Prediction of Soil Saturated Water Content Using Evolutionary Polynomial Regression (EPR)", Archives of Agronomy and Soil Science, 12, 2013. https://doi.org/10.1080/03650340.2013.861062

[33] Faramarzi, A., Javadi, AA., Alireza, A. "Numerical implementation of EPR-based material models in finite element analysis", Computers & Structures, 118, pp. 100-108. 2013. https://doi.org/10.1016/j.compstruc.2012.10.002

[34] Faramarzi, A., Javadi, A., Alani, M.A."EPR-based material modelling of soils considering volume changes", Computers & Geosciences, 48, pp. 73-85. 2012. https://doi.org/10.1016/j.cageo.2012.05.015

[35] Javadi, A., Ahangar-Asr, A., Johari, A., Faramarzi, A., Toll, D. "Modelling stress-strain and volume change behavior of unsaturated soils using an evolutionary based data mining technique, an incremental approach", Engineering Application of Artificial Intelligence, 25(5), pp. 926-933. 2012. https://doi.org/10.1016/j.engappai.2012.03.006

[36] Hussain, M.S., Javadi, A.A., Ahangar-Asr, A., Farmani, R. "A surrogate model for simulation-optimization of aquifer systems subjected to seawater intrusion", Journal of Hydrology, 523, pp 542-554, 2015. https://doi.org/10.1016/j.jhydrol.2015.01.079

[37] Ahangar-Asr, A., Johari, A., Javadi, A.A. "An evolutionary approach to modelling the soil-water curve in unsaturated soils", Computers & Geosciences, 43, pp 25-22, 2012. https://doi.org/10.1016/j.cageo.2012.02.021

[38] Storn, R., Price, K. "Differential evolution: a simple and efficient heuristic for global optimization over continuous spaces." Journal of Global Optimization, 11, 1997. http://dx.doi.org/10.1023/A:1008202821328.

[39] Schaap, M. "Rosetta: a computer program for estimating soil hydraulic parameters with hierarchical pedotransfer functions", Journal of Hydrology, 251(3), pp 163-176, 2001. https://doi.org/10.1016/S0022-1694(01)00466-8

[40] Gomes, G.J.C., Vargas Jr., E.A., "A coupled system based on Differential Evolution for the determination of Rainfall intensity equations", Brazilian Journal of Water Resources, 23, 2018. http://dx.doi.org/10.1590/2318-0331.231820170165

[41] Laucelli, D., Romano, M., Savic, D., Giustolisi, O. "Detecting anomalies in water distribution networks using EPR modelling paradigm", Journal of Hydroinformatics, 18(3), pp 409-427, 2016. https://doi.org/10.2166/hydro.2015.113