# An evolutionary multiobjective strategy for the effective management of groundwater resources

O. Giustolisi,[1] A. Doglioni,[2] D. A. Savic,[3] and F. di Pierro[3]

Received 19 July 2006; revised 26 July 2007; accepted 20 September 2007; published 3 January 2008.

[1] This paper introduces a modeling approach aimed at the management of groundwater resources based on a hybrid multiobjective paradigm, namely Evolutionary Polynomial Regression. Multiobjective modeling in hybrid evolutionary computing enables the user (a) to find a set of feasible symbolic models, (b) to make a robust choice of models and (c) to improve computational efficiency, simultaneously developing a set of models with diverse structural parsimony levels. Moreover, this methodology appears to be well suited to those cases where process input and the boundary conditions are not easily accessible. The multiobjective approach is based on the Pareto dominance criterion and it is fully integrated into the Evolutionary Polynomial Regression paradigm. This approach proves to be effective for modeling groundwater systems, which usually requires (a) accurate analyses of the underlying physical phenomena, (b) reliable forecasts under different hypothetical scenarios and (c) good generalization features of the models identified. For these reasons it is important to construct easily interpretable models which are specialized for well defined purposes. The proposed methodology is tested on a case study aimed at determining the dynamic relationship between rainfall depth and water table depth for a shallow unconfined aquifer located in southeast Italy.

**Citation:** Giustolisi, O., A. Doglioni, D. A. Savic, and F. di Pierro (2008), An evolutionary multiobjective strategy for the effective management of groundwater resources, *Water Resour. Res.*, *44*, W01403, doi:10.1029/2006WR005359.

## 1. Introduction

[2] Water management, the planned development, distribution and use of water resources is an inherently complex problem. *Simonovic* [2000] identifies two broad themes in water management problems: complexity and uncertainty. The first is related to the vast scope of the water resources domain and the intricacy of modeling tools in an environment characterized by continuous, rapid technological development. The latter is associated with often restricted data availability and the temporal/spatial variability of domain parameters that characterize water resources decision making. In semiarid to arid regions, where aquifers are the only freshwater resource, the complexity and uncertainty of problems often mean high resource procurement costs. Therefore the planning and management of groundwater resources presents a particular challenge in those regions deprived of abundant surface water deposits and which depend on pumping from wells [*Siegfried and Kinzelbach*, 2006].

[3] *Custodio* [2002] emphasizes that groundwater resources management, and in particular the degree of exploitation that is sustainable, depends on the detailed and updated characterization of aquifer-development conditions and the measures implemented for their moderation, correction or mitigation. Such management should not be dominated by the unquestioned application of general rules based on indirect data but, rather, sound management ought to be based on a combination of monitoring, aquifer characterization and system modeling.

[4] There already exists a strong tradition of groundwater resources modeling and system optimization within the context of planning and management. For example, *Jones et al.* [1987] used a differential dynamic programming algorithm to solve an unsteady nonlinear optimal control problem in groundwater. In the study of *Bierkens et al.* [2001] the spatiotemporal variation of shallow water table depth is modeled with a regionalized version of an autoregressive exogenous (ARX) time series model. The results therein presented can be used for optimal space-time prediction of water table depth, network optimization, and space-time conditional simulation. *Knotters and Bierkens* [2000] describe the relationship between precipitation excess and water table depth by a physically based ARX model. They show that the physically based ARX model predicts the effect of interventions reasonably well. *McKinney and Lin* [1994] incorporated groundwater simulation models into a genetic algorithm to solve three groundwater management problems: maximum pumping from an aquifer; lowest cost water supply; and minimum cost aquifer remediation. *Wang and Zheng* [1998] applied a Genetic Algorithm (GA) and simulated annealing, coupled with the MODFLOW finite difference groundwater flow model, to optimal groundwater remediation design. The

[1]Civil and Environmental Engineering Department, Technical University of Bari, Engineering Faculty of Taranto, Taranto, Italy.
[2]Department of Environmental Engineering and Sustainable Development, Technical University of Bari, Engineering Faculty of Taranto, Taranto, Italy.
[3]Centre for Water Systems, Department of Engineering, University of Exeter, Exeter, UK.


0043-1397/08/2006WR005359$09.00

model was applied over various management periods and included both fixed and operating costs. The problem of experimental design for parameter estimation was formulated and solved by *McPhee and Yeh* [2006] using a combination of genetic algorithms and gradient-based optimization. *Siegfried and Kinzelbach* [2006] presented a methodology for the determination of optimal, cooperative allocation policies in multiobjective aquifer management problems. They integrated a finite difference aquifer model with an economic model that accounts for water provision costs.

[5] Modeling of most water systems tends to fall under two broad categories: physically based and data-driven. Both have been applied with success and neither is immune from certain challenges. Data-driven modeling techniques are widely used approaches to the modeling of environmental phenomena and may be especially helpful when physically based approaches lead to overly complex problem formulations, poorly characterized boundary conditions and/or intractable solutions. Moreover, the mathematical formulation of the problem generally involves empirically determined coefficients that might depend on statistical analyses or surrogate data. Thus a data-driven approach may produce good results and can serve alongside a physically based counterpart in order to provide a more thorough system characterization. The growing popularity of data-driven approaches can in part be attributed to the increasing availability of monitoring data and information, and to the relatively small number of environmental models based on the first principles articulation of underlying system physics. Despite their popularity and variety, the most commonly adopted model selection approaches, which are based on Single-Objective (SO) optimization, can often be restrictive and lead to an unsatisfactory model choice. For example, under a SO scenario, a small improvement in the fitness indicator (i.e., a quantitative or qualitative measure of the level of agreement between observed and simulated data) can usher in a more complex model to the detriment of a more parsimonious alternative [*Young et al.*, 1996]. Specifically, complexity here refers to the number of independent variables (parameters) comprising the model. In addition to the problem of complexity, the modeler usually has to work with small data sets affected by non-Gaussian errors which can further hamper good model selection.

[6] More robust model selection can be realized employing a Multiobjective (MO) approach, which produces a set of non-dominated models [*Van Veldhuizen and Lamont*, 2000] selected according to multiple and often conflicting objectives. In addition to achieving good agreement (fitness level) with training data, these models must have an acceptable level of parsimony [*Giustolisi and Simeone*, 2006]. This leads to (a) simply structured models, reasonably representative of the system's dynamic behavior, rather than of the specific error realization contained in training data, and (b) accurate selection of those input variables which are physically relevant for the output.

[7] The complexity of the model's structure is evaluated through one or more objective functions. These, along with the objective function that quantifies the degree of agreement between observed and simulated measurements (fitness), constitute the objective space that must be explored for *good* models. The approach proposed in this paper resorts to the Pareto dominance criterion [*Pareto*, 1896] to assess the quality of different models. The criterion is used to select the so-called Pareto solutions (models) where each solution of the set is not dominated by any other solution, i.e., in going from one solution to another, it is not possible to improve on one objective (e.g., reduce the complexity) without making at least one of the other objectives worse (e.g., reduce fitness). This approach permits broad comparative analysis of contiguous non-dominated models with the aim of achieving a proper level of robustness in the selection procedure. In this paper, the use of symbolic models, such as those in the form of clearly understandable formulas based on elementary mathematical functions (polynomials), is proposed. In addition to these models being amenable to mathematical processing (e.g., derivative operations), the MO strategy proposed here facilitates comparison of model structures (e.g., instances of common/uncommon features or the contiguity of formulas).

[8] The approach introduced in this paper is based on a hybrid paradigm, Evolutionary Polynomial Regression (EPR), developed by *Giustolisi and Savic* [2006] and tested on different environmental problems [*Giustolisi et al.*, 2004a, 2007]. It is a two-stage technique, which allows the user to identify models with either an SO or MO optimization procedure. This paper presents the MO features of EPR applied to the modeling of the unknown dynamic relationship between rainfall and groundwater levels for an aquifer located in the region of Apulia, Southeast Italy (near Brindisi). The fundamental premise of the work is that it is possible to identify the best performing model in terms of on-line predictions [*Ljung*, 1987] by evaluating its performance on test data for variable prediction horizons. At the same time, the influence of each input (and combinations thereof) on the predicted outputs (i.e., returned model structures) will be analyzed. The objective functions used are: (a) maximization of the fitness, (b) minimization of the total number of inputs selected by the modeling strategy and (c) minimization of the length of the model expression.

## 2. EPR Background

[9] EPR is a data-driven hybrid technique based on evolutionary computing; its paradigm can be classified as Genetic Programming (GP) [*Koza*, 1992]. While GP has been shown to be effective for modeling [*Babovic and Keijzer*, 2000], it suffers from certain limitations, as demonstrated by *Soule and Heckendorn* [2002]. In the literature, several attempts to overcome these limitations are reported. In particular, *Davidson et al.* [2003] offer an interesting demonstration of the potential of a hybrid evolutionary strategy based on rules for rendering GP more effective. EPR is similar to GP in terms of the class of results it generates (symbolic formulas), but it circumvents some of GP's shortcomings by integrating a GA [*Goldberg*, 1989] with a least squares (LS) approach. Therefore EPR is a two-stage method that 1) searches model structures based on an integer GA and 2) estimates their parameters based on the linear optimization, which represents a simple link between the symbolic and the numerical regressive nature of EPR.

[10] EPR focuses its search on pseudo-polynomial structures, which are summarized as [*Giustolisi and Savic*, 2006]

$$\mathbf{Y} = a_0 + \sum_{j=1}^{m} a_j \cdot (\mathbf{X}_1)^{\mathbf{ES}(j,1)} \cdot \ldots \cdot (\mathbf{X}_k)^{\mathbf{ES}(j,k)} \cdot f\left((\mathbf{X}_1)^{\mathbf{ES}(j,k+)}\right) \cdot \ldots \cdot f\left((\mathbf{X}_k)^{\mathbf{ES}(j,2k)}\right) \quad \text{case } 0$$

$$\mathbf{Y} = a_0 + \sum_{j=1}^{m} a_j \cdot f\left((\mathbf{X}_1)^{\mathbf{ES}(j,1)} \cdot \ldots \cdot (\mathbf{X}_k)^{\mathbf{ES}(j,k)}\right) \quad \text{case } 1$$

$$\mathbf{Y} = a_0 + \sum_{j=1}^{m} a_j \cdot (\mathbf{X}_1)^{\mathbf{ES}(j,1)} \cdot \ldots \cdot (\mathbf{X}_k)^{\mathbf{ES}(j,k)} \cdot f\left((\mathbf{X}_1)^{\mathbf{ES}(j,k+1)} \cdot \ldots \cdot (\mathbf{X}_k)^{\mathbf{ES}(j,2k)}\right) \quad \text{case } 2$$

$$\mathbf{Y} = g\left(a_0 + \sum_{j=1}^{m} a_j \cdot (\mathbf{X}_1)^{\mathbf{ES}(j,1)} \cdot \ldots \cdot (\mathbf{X}_k)^{\mathbf{ES}(j,k)}\right) \quad \text{case } 3 \tag{1}$$

where $\mathbf{X}_i$ are the vectors of candidate inputs; **ES** is the matrix of exponents (coded as integers in the GA); $f$ and $g$ are user-specified functions; $a_j$ are constant values and m is the length of the expressions. Note that the last structure of equation (1) requires the assumption of an invertible $g$-function, because of subsequent parameter estimation. Moreover, when **ES** is zero the input variable assumes a constant value of one and is then deselected.

[11] Let $j$ represent the subscript associated to constant values, the parameters $a_j$ are estimated by an LS method integrated in the EPR procedure. The LS guarantees a biunique correspondence between the structure and its constant values. In addition to the usual LS search, the user can force the LS to search for structures that contain only positive constant values ($a_j > 0$) according to the approach introduced by *Lawson and Hanson* [1974]. In environmental modeling, there is a high probability that the negative constant values ($a_j < 0$) are selected in order to offset the particular realization of errors related to the finite data set. The models thus identified were shown to be more generic and physically sound than those obtained with the incorporation of negative values. Moreover, the formulae generated by EPR often lend themselves to a physical interpretation of their monomial components.

[12] Among the equations in (1), the structure denoted as case 0 was chosen. A simpler example of an equation belonging to case 0 is

$$y = a_1 \cdot x_1^{\alpha} \cdot x_2^{\beta} + a_2 \cdot x_2^{\sigma} \cdot x_3^{\gamma} \cdot x_4^{\delta} + a_3 \cdot x_1^{\beta} \cdot x_4^{\alpha} + a_0 \tag{2}$$

where $y$ is the output of the system/process; $a_1$,$a_2$, $a_3$, $a_0$, are constant values; $x_1$, $x_2$, $x_3$, $x_4$, are inputs selected by the process among the user-specified range of candidates and $\alpha$, $\beta$, $\gamma$, $\delta$ are process-selected exponents from a set prespecified by the user. Although the general configuration of the structures is defined by the user (i.e., inputs, exponents and maximum expression length), EPR can return simplified structures according to the strategy it pursues.

[13] The space of candidate formulas can be explored by EPR according to two main strategies: (1) an SO search and (2) an MO approach. Although the effectiveness of the SO approach in environmental modeling has been demonstrated [*Giustolisi et al.*, 2007], it presents some drawbacks. The MO approach outperforms that of SO, since it explores the space of candidate formulas by assuming just the maximum number of constants ($a_j$). This article focuses on the MO strategy.

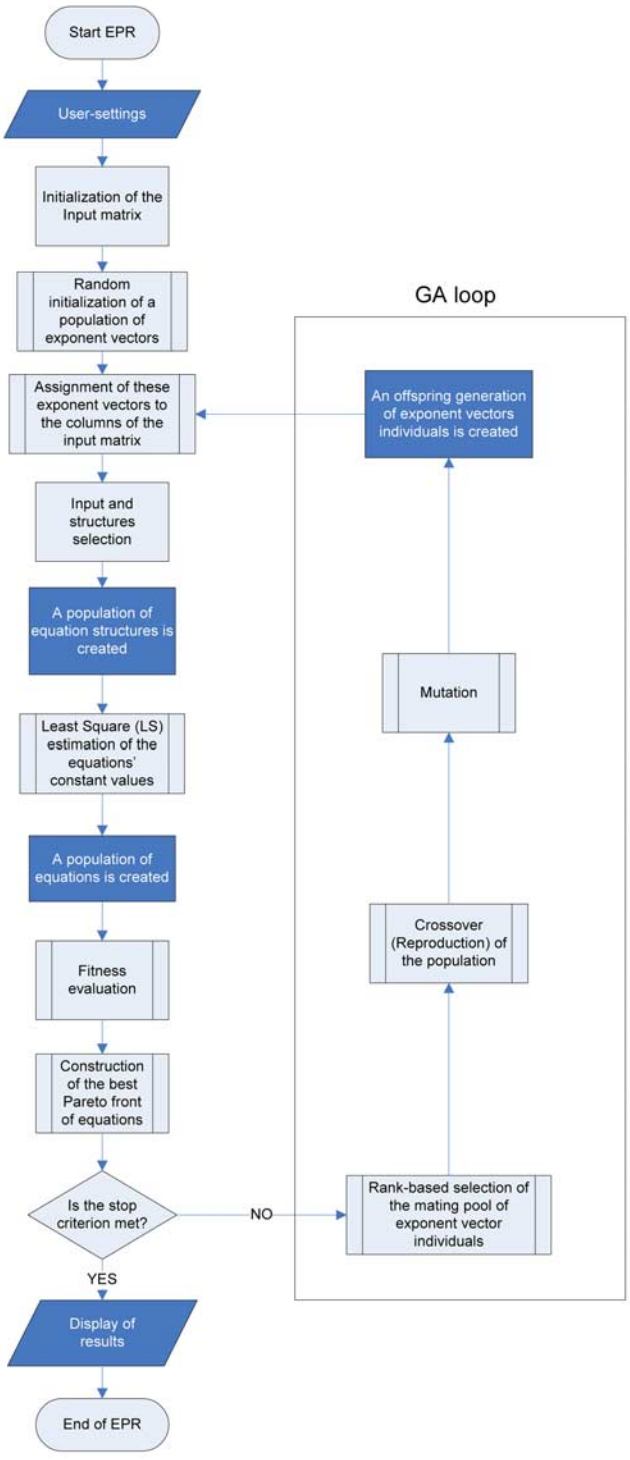


**Figure 1.** EPR flowchart.

[14] The objectives assumed for this search are three: (1) the number of constants ($a_j$), (2) the total number of inputs ($\mathbf{X}_k$) represented in each formula and (3) the models' fitness to data. Furthermore, this approach ranks the generated formulas according to their fitness, the number of constant values and the total number of inputs incorporated in each formula. In Figure 1, the EPR flowchart is given, outlining the steps of the procedure.

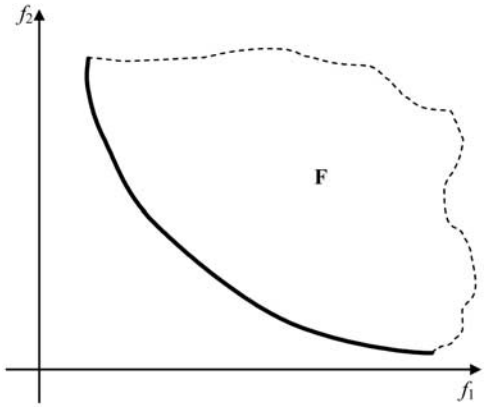


**Figure 2.** Representation of the Pareto front for a design space F of a dual-objective optimization problem. The bold line in the left part of the plot is the Pareto front.

[15] Finally, the GA used for the evolutionary stage of EPR is OPTIMOGA [*Giustolisi et al.*, 2004b], which is employed to select the set of independent variables ($\mathbf{X}_k$) that must form the model structure. Further details on OPTIMOGA are discussed by *Giustolisi et al.* [2004b]; *Giustolisi and Savic* [2006] accurately describe how OPTIMOGA is applied in EPR in order to conduct the structural identification of models.

### 2.1. Multiobjective Modeling by EPR

[16] Although the SO approach of EPR has proven effective in several applications [*Giustolisi and Savic*, 2006; *Giustolisi et al.*, 2004a; 2007], optimization results are often difficult to interpret. In fact, the set of candidate models could be either ranked according to their fitness or to their structural complexity. However, sorting models according to their intricacy requires subjective judgment and, consequently, the process risks being biased by the analyst's experience rather than being purely based on mathematical criteria [*Young et al.*, 1996]. To avoid this pitfall, a MO strategy is implemented and integrated into EPR to improve both the post-processing phase and the general modeling framework. Such a strategy allows model ranking according to both the Coefficient of Determination (CoD) and structural complexity (i.e., the number of parameters and total number of inputs in the symbolic expression).

[17] To date, MO Evolutionary Computing has inspired several algorithms and a comprehensive discussion can be found in the work of *Coello Coello et al.* [2002]. These algorithms often exhibit pronounced qualitative differences, but all possess a common element; that is, a procedure to assess the quality of solutions given a set of objective functions. A number of alternative procedures have been presented over the years, but those based on the concept of Pareto efficiency have attracted a greater consensus [*Deb*, 2001]. Consequently, the ranking procedure featured in the MO approach of EPR, for assigning fitness values to the solutions generated, is based upon it.

[18] A brief definition of the MO approach applied to the optimization problems is given. Generally speaking, the MO optimization problem consists in finding a vector of decision variables, which satisfies constraints and optimizes a vector function whose elements represent the objective functions [*Coello Coello*, 1999]. These objective functions are representative of performance criteria which are usually non-commensurate. Therefore the goal of such optimization is to unearth a set of solutions which are acceptable for the designer/analyst. This formally corresponds to finding the vector of decision variables

$$\bar{x}^* = \left[x_1^*, x_2^*, \ldots, x_n^*\right]^{\mathrm{T}}$$

satisfying the $m$ inequality constraints

$$g_i(\bar{x}) \geq 0 \quad i = 1, 2, \ldots, m \tag{3}$$

and the $p$ equality constraints

$$h_i(\bar{x}) = 0 \; i = 1, 2, \ldots, p$$

which optimize the following

$$\bar{f}(\bar{x}) = [f_1(\bar{x}), f_2(\bar{x}), \ldots, f_k(\bar{x})]^{\mathrm{T}} \tag{4}$$

[19] It is now given the definition of Pareto optimal set. A vector $\bar{x}^*$ is Pareto optimal if there exists no feasible vector $\bar{x}$ which would decrease some criterion without causing a simultaneous increase in at least one other criterion [*Coello Coello*, 1999; *Van Veldhuizen and Lamont*, 2000]. However, Pareto optimality almost never implies a single solution, but rather a set of solutions that constitute the non-inferior or non-dominated solution set.

[20] For instance, if minimization of a dual objective problem is sought, the minima in the Pareto sense are on the boundary of the design space **F**. Looking at Figure 2, the bold line in the left zone of the picture is the Pareto front, while the dotted line represents the boundary of the design space **F**. The front is usually constructed by computing each point since its analytical expression is not easy to be found [*Coello Coello*, 1999].

[21] The objective functions subject to minimization in MO EPR are: (a) (1-CoD), which addresses the performance of models in terms of fitness to data evaluated on the 1-step-ahead prediction; (b) the number of constant values $a_j$ and (c) the total number of inputs involved in the symbolic expression. Those objective functions reported as (b) and (c) relate to the structural complexity of the models. Note that the total number of inputs corresponds to the number of times each input is involved in the symbolic expression. The user must set the maximum number of constant values, which poses an upper limit on the length of the candidate expressions. Therefore EPR seeks the best non-dominated models with respect to both structural complexity and fitness. Clearly, this approach enables an improved post-processing phase compared to SO EPR as the identified symbolic expressions are automatically ranked according to both their fitness and complexity.

### 2.2. Motivation for Using EPR in Groundwater Modeling

[22] The original approach in EPR can make it a reliable strategy in modeling environmental problems. These problems are frequently addressed by physically based strategies in which the analyst constructs a representation of the

system based on a mathematical representation of the known physics and then tests multiple hypothetical scenarios to simulate the effects that a particular input can produce on the output. While such models can perform remarkably well due to their solid theoretical grounding, they are often not easily applicable since they can be highly complex and demand accurate calibration based on data which is usually unavailable, costly or inadequate for such purpose [*Coppola et al.*, 2002]. Consequently, the analyst must often resort to simplifying assumptions, possibly undermining the theoretical purity (and performance) of the model. Often, when faced with intractability problems and/or an imperfectly understood system, a simplified physical representation may be excessively rudimentary with only narrow applicability. An alternative strategy intended to mitigate these potential drawbacks is data-driven modeling which can equip analysts with easily approachable models (i.e., formulas), that are simply structured and, perhaps, more meaningful with respect to system dynamics. EPR is designed to possess these features. The combination of GA, for finding the best functional structures, and LS, for evaluation of the constants $a_j$, offers certain advantages. On the one hand, a two-way unique relationship between the model structure and the constants is guaranteed by LS and, on the other, GA performs a fairly global exploration of the model space (symbolic expressions) given a set of objective functions. These should be carefully chosen during the model construction stage and ought to embody any a priori knowledge of the natural phenomenon being studied.

## 3. The Case Study

[23] A case study designed to identify the unknown dynamic relationship between average daily rainfall and groundwater levels for an aquifer located in the vicinity of Brindisi in Apulia (Southeast Italy) is presented. It represents an interesting case study because it is a relatively straightforward hydrogeological system occupying a modest area (about 200–300 km$^2$) and consisting of a shallow aquifer recharged only by direct rainfall and thus serves as an ideal object for study of the relationship between groundwater level and precipitation. Moreover, the management of groundwater resources in a Mediterranean climate zone, like that which prevails in Southern Italy, is an important issue since it is typical of regions susceptible to periodic acute water shortages. The shallow aquifer considered in this study is a critical source of irrigation water that plays an important role in food security and the regional economy. Furthermore, substantial pumping from wells in late spring, summer and early autumn is depleting reserves and aggravating water quality, a situation which is common in numerous places globally. In this scenario, the prediction of groundwater levels based on simple models that exploit monitoring data such as measured groundwater levels and rainfall depths (often easily collected and readily available from the national/local hydrographic services) can support the regulation, planning and conservation of local water resources.

[24] Due to its inherent non-linearity, and the presence of unknown extra inputs and boundary conditions, this problem poses significant computational difficulty and previous attempts to confront it using physically based approaches were found wanting. For instance, *Yi and Lee* [2004] presented a methodology based on Transfer Function-Noise (TFN) models. They emphasize that groundwater heads are usually collected as incomplete time series and that the time intervals among the samples is not uniformly distributed throughout the whole data record. To improve tractability, the TFN approach was integrated with the Kalman filtering and the maximum likelihood criterion, as suggested by *Jones* [1980].

### 3.1. Background to Data

[25] The groundwater system consists of a shallow unconfined aquifer of about 300 km$^2$ near Brindisi, in the northern part of the Salento Peninsula of Apulia (see Figure 3).

[26] This groundwater reservoir is an open system supplied only by direct rainfall but empties to a deeper regional aquifer that is in turn drained by a well developed surface channel network. A detailed description of the aquifer, based on stratigraphic data from boreholes, outcrop observations, satellite radar images, permeability tests and piezometric measurements is made by *Ricchetti and Polemio* [1996].

[27] The aquifer lies in sandy soils that outcrop extensively in the wide structural tectonic depression spanning two large calcareous blocks of the Apulian calcareous platform: Murge and Salento. Permeability of the surface aquifer's constitutive soils is low, ranging from $8 \times 10^{-6}$ m/s to $1.4 \times 10^{-4}$ m/s, while in the impervious clayish layer, the permeability ranges between $2 \times 10^{-6}$ m/s and $1 \times 10^{-7}$ m/s.

[28] This particular stratigraphic sequence establishes in effect a complex double-aquifer system, the former being the deep regional costal aquifer that is housed in the limestone fissured platform; the latter being a shallow unconfined aquifer situated in the quaternary deposits. Subappenine clays act as a barrier to infiltration and thus render sub-horizontal circulation of water possible in the overlying soil, creating an aquifer supplied only by direct rainfall. Significant water losses in favor of the deeper regional aquifer may intervene, especially since the latter is also unconfined and the hydraulic gradient of the flow through the clay layer may easily reach values higher than one, especially where the thickness is reduced. On the basis of this consideration, it can be argued that important interactions may take place locally between the two aquifers. The entity of this interaction can only be roughly conjectured, since geological data derived from only 45 wells irregularly distributed over the entire area.

[29] In order to study the relationship between groundwater level variations and rainfall, the data that have been used consist of measured phreatic levels from a well located close to Brindisi and rainfall data from the city's rain gauge station, both belonging to the Italian National Hydrographic Service. Phreatic records are available for a long period (1952–1996) and some qualitative considerations can be expressed on the basis of simple preliminary analyses of these data. The study area is characterized by a typical Mediterranean climate having every year a single dry period and a single wet period. Aquifer recharge takes place mainly in the first three months of the year while the rainy autumn months do not contribute to replenishment since infiltrated water is first sequestered by the soil in order to restore its field capacity. The largest variations in pluviometric regime occur in March and April, experiencing a minimum during

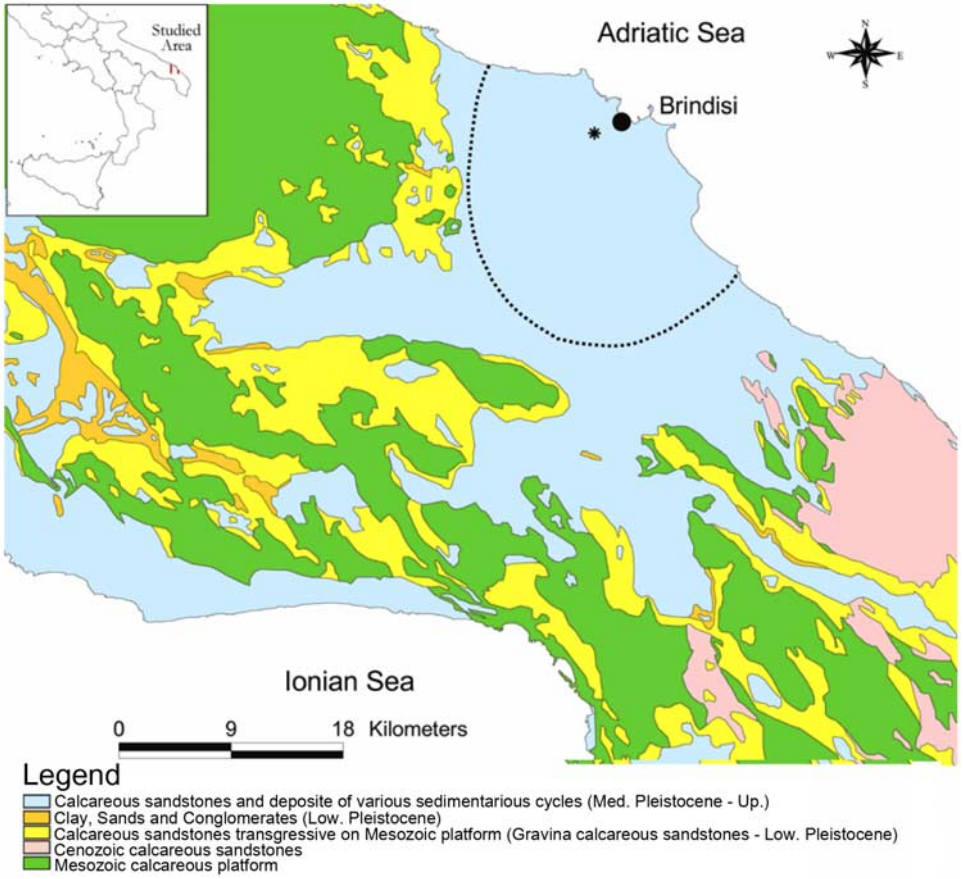


**Figure 3.** Location of the sampling well. The star represents the sampling well, solid circle is the rain gauge station and the dotted line represents the approximate bound of the aquifer.

summer when evapotranspiration is most intense. Recharge is more evident on groundwater level fluctuations when passing from autumn to spring if no acute or anomalous events (periods of atypical precipitation pattern) take place. In such cases, surface runoff is typically more pronounced at the expense of infiltration.

[30] Data sets consists of 528 observations: rainfall data series consist of daily values averaged on a monthly basis, measured in mm, and groundwater data are the average monthly values of the level of the water's free surface in the well. The mouth of the well is located at 35.92 m a.s.l. (above sea level). Both the rainfall and groundwater data series cover a 44-a period (January 1953 to December 1996). Figures 4 and 5 show the time plots of rainfall and groundwater levels, respectively. The data used in order to construct the models, referred to as the training set, range over a period of 300 months (January 1953 to December 1977).

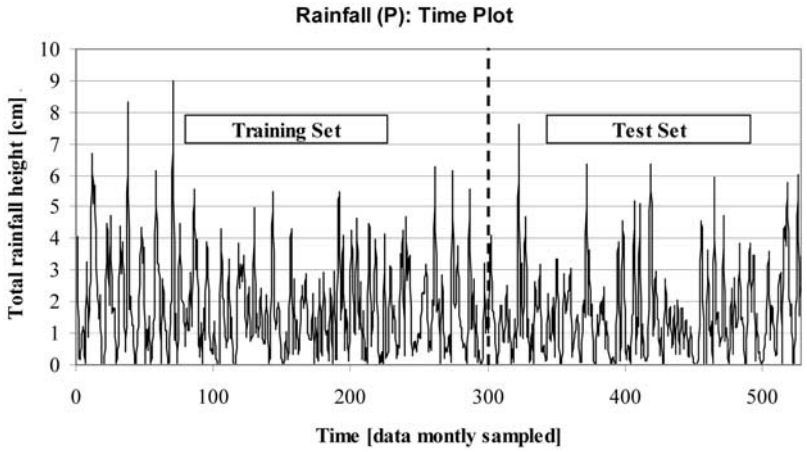


**Figure 4.** Time series of rainfall, data are divided for modeling purpose into a training set and a test set.

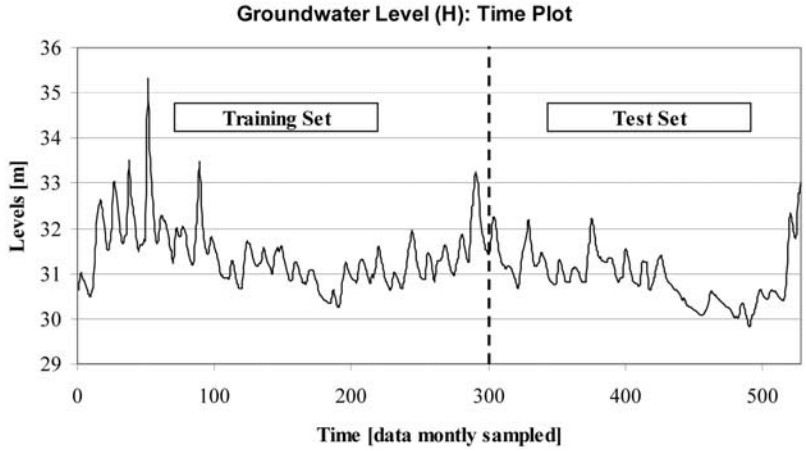


**Figure 5.** Time series of groundwater levels, data are divided for modeling purpose into a training set and a test set.

[31] The remaining 228 months of data were used for validation purposes and are referred to as unseen data. Specifically, these were applied to assess the quality of the models' predictions at 1, 2, 4, 6, and 12 months into the future. In the 44 years of observations considered, a decay in phreatic levels is apparent, as depicted in Figure 6 where the annual mean levels of the aquifer are indicated. In order to highlight this decay a linear fitting of the average values is reported in Figure 6; the linear fitting emphasizes a decreasing trend, which can be quantified as an average decrease of 3.3 %, corresponding to 1.04 m over 44 years.

[32] Once the Pareto front of models is obtained and the modeling phase is concluded (i.e., candidate models have been ascertained), the unseen data are employed to test the capability of these models with similar data, but drawn from a different time window; that is, the data represent the same variables and underlying phenomena.

### 3.2. Preliminary Modeling Aspects

[33] The modeling phase was carried out according to the following assumptions:

[34] ● The set of variables considered as candidate input to the models are: $H_{t-1}$ and $H_{t-2}$ as past measured values of the groundwater head, and $P_t$, $P_{t-1}$, $P_{t-2}$, $P_{t-3}$, $P_{t-4}$, $P_{t-5}$, $P_{t-6}$, $P_{t-7}$, $P_{t-8}$, $P_{t-9}$, $P_{t-10}$, $P_{t-11}$, $P_{t-12}$ as measured values of the rainfall depths. Subscripts denote the measurement time: for instance $t - 2$ indicates the groundwater level observed two months before the present ($t$). These candidate inputs to the models have been selected according to aquifer response (3–4 months delayed) to the rainfall perturbations reported by *Ricchetti and Polemio* [1996]. A longer delay (up to one year) is also considered in order to investigate the effects of aquifer recharge due to non-local rainfall. Past output measures $H_{t-1}$ and $H_{t-2}$

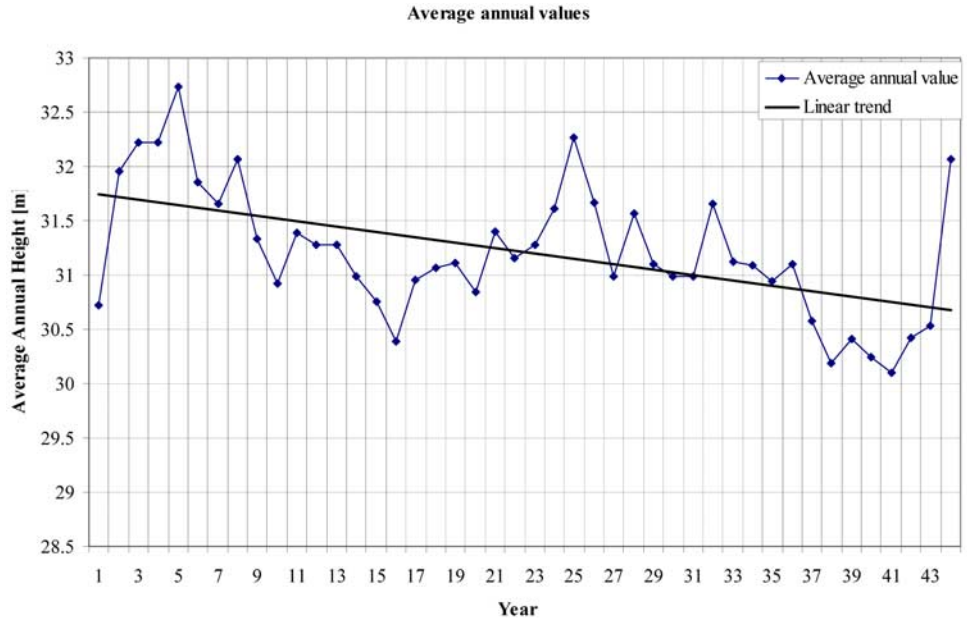


**Figure 6.** Average annual values of piezometric heads and linear trend of data over the 44 years investigation period.

have been incorporated to reflect the persistence of piezometric head variation.

[35] • The possible model structures, see equation (2), are assumed to be polynomial only.

[36] • The polynomial expressions consist of four terms at most, excluding the bias term (if selected by the procedure).

[37] • Each monomial term is the product of the methodology-selected inputs to the power of the exponents selected by EPR in the pre-specified set {0; 0.5; 1; 2}. The exponent 0 allows the procedure to deselect the unnecessary inputs, the exponent 0.5 smoothes the effect of the input, the exponent 1 introduces a linear effect to the input and, finally, the exponent 2 amplifies the effect of the input.

[38] • The LS estimate of the constant $a_j$ is constrained to positive values according to the approach by *Lawson and Hanson* [1974].

[39] • Data are never scaled.

[40] • The optimization parameters are: 24000 generations, initial population size 100 elements, probability of crossover 0.4 and probability of mutation 0.1.

[41] • The number of potential candidate solutions among which EPR searches is $1.33 \times 10^{36}$.

[42] The main fitness indicator considered in this paper is the Coefficient of Determination (CoD),

$$\mathrm{CoD} = 1 - \frac{\mathrm{N}-1}{\mathrm{N}} \frac{\sum_{\mathrm{N}} \left(\hat{H} - H_{\mathrm{exp}}\right)^2}{\sum_{\mathrm{N}} \left(H_{\mathrm{exp}} - \mathrm{avg}\left(H_{\mathrm{exp}}\right)\right)^2} = 1 - \mathrm{k} \cdot \mathrm{SSF}$$
$$\mathrm{k} = \frac{2(\mathrm{N}-1)}{\sum_{\mathrm{N}} \left(H_{\mathrm{exp}} - \mathrm{avg}\left(H_{\mathrm{exp}}\right)\right)^2} \tag{5}$$

where N is the number of samples, $H$ and $H_{\mathrm{exp}}$ are the values of groundwater head simulated by the model and measured, respectively, and avg($H_{\mathrm{exp}}$) represents the average value of measured groundwater heads evaluated for the N samples. It can be seen from equation (5) that CoD and SSE (Sum of Squared Errors) are clearly related (note that the value $k$ does not depend on the particular model).

[43] The set of non-dominated models identified by EPR defines a global scenario of possible model structures which is presented to the analyst who must then select the best candidate for the problem at hand. This final selection is guided by an analysis of the similarities and differences among formulae and through consideration of the trade-off between structural complexity and fitness level attained. Therefore the user can identify those terms/inputs that are common among the models and assess which terms/inputs are discarded by the methodology when the structural complexity decreases. Moreover, this analysis permits identification of terms that appear in one model only and such terms are likely to be weakly related to the physical phenomenon, but rather to the specific error realization contained in data.

### 3.3. Modeling Results

[44] In this section, the entire set of non-dominated EPR models is presented keeping in mind that the goal is to furnish a decision support strategy and not strictly a model suitable for a unique case study. EPR identified 24 non-dominated models, described by the equations (6) to (29).

$$H_t = 0.036285 \cdot P_{t-1} \cdot P_{t-2}^{0.5} + 0.0016503 \cdot P_t^2 \cdot P_{t-1}^{0.5} \cdot P_{t-3}^{0.5} \cdot P_{t-5}^{0.5} \cdot P_{t-6}^{0.5} \cdot P_{t-10}^{0.5} + 0.872 \cdot H_{t-1} + 3.0356 \cdot 10^{-5} \cdot H_{t-1} \cdot P_{t-2} \cdot P_{t-3} \cdot P_{t-4} \cdot P_{t-12}^2 + 3.888 \tag{6}$$

$$H_t = 0.000967 \cdot P_{t-2} \cdot P_{t-3} \cdot P_{t-4} \cdot P_{t-12}^2 + 0.0016499 \cdot P_t^2 \cdot P_{t-1}^{0.5} \cdot P_{t-3}^{0.5} \cdot P_{t-5}^{0.5} \cdot P_{t-6}^{0.5} \cdot P_{t-10}^{0.5} + 0.036299 \cdot P_{t-1} \cdot P_{t-2}^{0.5} + 0.87292 \cdot H_{t-1} + 3.8593 \tag{7}$$

$$H_t = 0.038174 \cdot P_{t-1} \cdot P_{t-2}^{0.5} + 0.0038309 \cdot P_t^2 \cdot P_{t-3}^{0.5} \cdot P_{t-5}^{0.5} \cdot P_{t-10}^{0.5} + 0.87837 \cdot H_{t-1} + 2.9384 \cdot 10^{-5} \cdot H_{t-1} \cdot P_{t-2} \cdot P_{t3} \cdot P_{t-4} \cdot P_{t-12}^2 + 3.6751 \tag{8}$$

$$H_t = 0.000936 \cdot P_{t-2} \cdot P_{t-3} \cdot P_{t-4} \cdot P_{t-12}^2 + 0.0038306 \cdot P_t^2 \cdot P_{t-3}^{0.5} \cdot P_{t-5}^{0.5} \cdot P_{t-10}^{0.5} + 0.038187 \cdot P_{t-1} \cdot P_{t-2}^{0.5} + 0.87926 \cdot H_{t-1} + 3.6473 \tag{9}$$

$$H_t = 0.00096063 \cdot P_{t-2} \cdot P_{t-3} \cdot P_{t-4} \cdot P_{t-12}^2 + 0.0043271 \cdot P_t^2 \cdot P_{t-3}^{0.5} \cdot P_{t-6}^{0.5} + +0.037119 \cdot P_{t-1} \cdot P_{t-2}^{0.5} + 0.87897 \cdot H_{t-1} + 3.6592 \tag{10}$$

$$H_t = 0.024502 \cdot P_t^{0.5} \cdot P_{t-1} \cdot P_{t-2}^{0.5} + 3.1445 \cdot 10^{-5} \cdot H_{t-1} \cdot P_{t-2} \cdot P_{t-3} \cdot P_{t-4} \cdot P_{t-12}^2 + 0.88525 \cdot H_{t-1} + 3.4974 \tag{11}$$

$$H_t = 0.00093406 \cdot P_{t-2} \cdot P_{t-3} \cdot P_{t-4} \cdot P_{t-12}^2 + 0.037746 \cdot P_{t-1} \cdot P_{t-2}^{0.5} + 0.0059524 \cdot P_t^2 \cdot P_{t-4}^{0.5} + 0.8864 \cdot H_{t-1} + 3.4189 \tag{12}$$

$$H_t = 0.0010018 \cdot P_{t-2} \cdot P_{t-3} \cdot P_{t-4} \cdot P_{t-12}^2 + 0.024506 \cdot P_t^{0.5} \cdot P_{t-1} \cdot P_{t-2}^{0.5} + 0.88619 \cdot H_{t-1} + 3.4678 \tag{13}$$

$$H_t = 0.00095454 \cdot P_{t-2} \cdot P_{t-3} \cdot P_{t-4} \cdot P_{t-12}^2 + 0.037688 \cdot P_{t-1} \cdot P_{t-2}^{0.5} + 0.0064378 \cdot P_t^2 + 0.89239 \cdot H_{t-1} + 3.2279 \tag{14}$$

$$H_t = 0.0009458 \cdot P_{t-2} \cdot P_{t-3} \cdot P_{t-4} \cdot P_{t-12}^2 + 0.040328 \cdot P_{t-1} \cdot P_{t-2}^{0.5} + 0.88272 \cdot H_{t-1} + 3.5611 \tag{15}$$

$$H_t = 0.00052034 \cdot P_{t-3}^2 \cdot P_{t-4} \cdot P_{t-12}^2 + 0.041465 \cdot P0_{t-1} \cdot P_{t-2}^{0.5} + 0.0063853 \cdot P_t^2 + 0.90026 \cdot H_{t-1} + 2.9772 \tag{16}$$

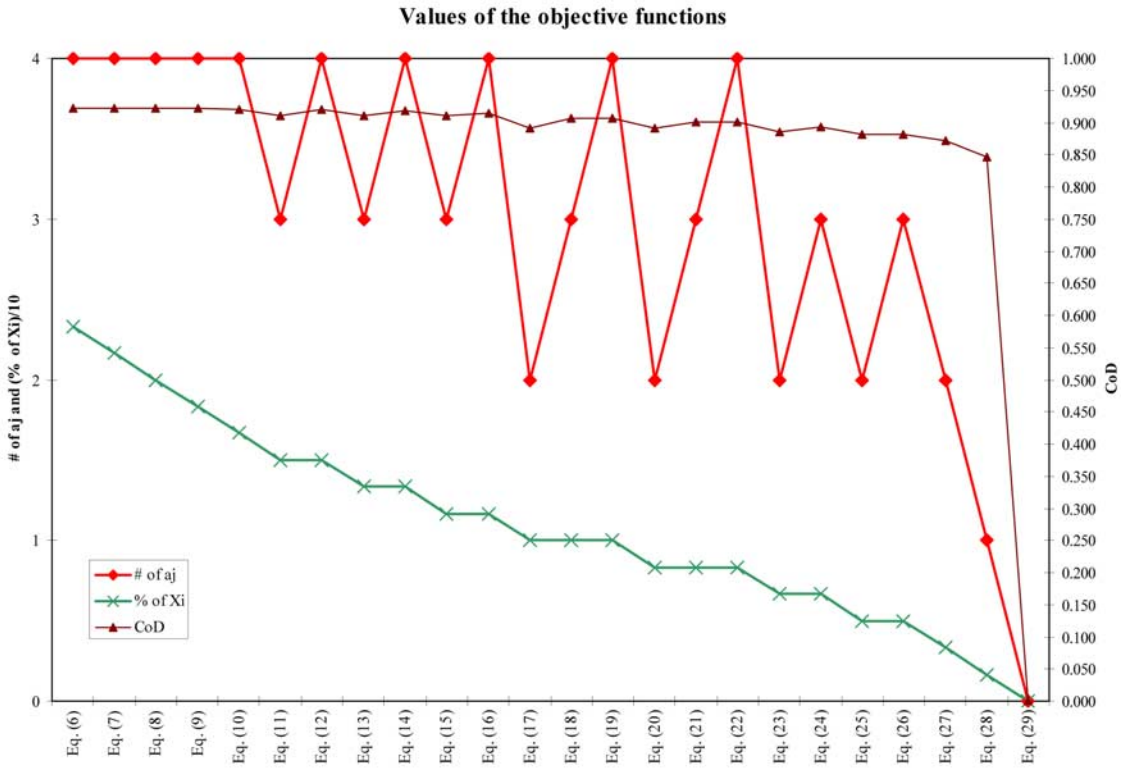


**Figure 7.** Representation of the values of the objective functions, % of Xi, # of aj and CoD, for the whole set of models found by EPR.

$$H_t = 0.8913 \cdot H_{t-1} + 0.000078477 \cdot H_{t-1} \cdot P_{t-1} \cdot P_{t-2}^{0.5} \cdot P_{t-3}^{0.5} \cdot P_{t-12}^{0.5} + 3.3297 \quad (17)$$

$$H_t = 0.00051611 \cdot P_{t-3}^2 \cdot P_{t-4} \cdot P_{t-12}^2 + 0.044044 \cdot P_{t-1} \cdot P_{t-2}^{0.5} + 0.89059 \cdot H_{t-1} + 3.3107 \quad (18)$$

$$H_t = 0.00083735 \cdot P_{t-3}^2 \cdot P_{t-12}^2 + 0.042462 \cdot P_{t-1} \cdot P_{t-2}^{0.5} + 0.0060721 \cdot P_t^2 + 0.91195 \cdot H_{t-1} + 2.612 \quad (19)$$

$$H_t = 0.024842 \cdot P_{t-1} \cdot P_{t-2}^{0.5} \cdot P_{t-3}^{0.5} \cdot P_{t12}^{0.5} + 0.89375 \cdot H_{t-1} + 3.2524 \quad (20)$$

$$H_t = 0.00084387 \cdot P_{t-3}^2 \cdot P_{t-12}^2 + 0.04483 \cdot P_{t-1} \cdot P_{t-2}^{0.5} + 0.90249 \cdot H_{t-1} + 2.9371 \quad (21)$$

$$H_t = 0.00084049 \cdot P_{t-3}^2 \cdot P_{t-12}^2 + 0.039485 \cdot P_{t-2} + 0.055426 \cdot P_{t-1} + 0.90273 \cdot H_{t-1} + 2.865 \quad (22)$$

$$H_t = 0.037769 \cdot P_{t-1} \cdot P_{t-3}^{0.5} \cdot P_{t-12}^{0.5} + 0.90478 \cdot H_{t-1} + 2.8975 \quad (23)$$

$$H_t = 0.00090171 \cdot P_{t-3}^2 \cdot P_{t-12}^2 + 0.064128 \cdot P_{t-1} + 0.91296 \cdot H_{t-1} + 2.5959 \quad (24)$$

$$H_t = 0.049876 \cdot P_{t-1} \cdot P_{t-2}^{0.5} + 0.91332 \cdot H_{t-1} + 2.6136 \quad (25)$$

$$H_t = 0.11944 \cdot P_{t-2}^{0.5} + 0.059274 \cdot P_{t-1} + 0.91334 \cdot H_{t-1} + 2.4827 \quad (26)$$

$$H_t = 0.070405 \cdot P_{t-1} + 0.92572 \cdot H_{t-1} + 2.2141 \quad (27)$$

$$H_t = 0.9178 \cdot H_{t-1} + 2.5874 \quad (28)$$

$$H_t = 31.449 \quad (29)$$

[45] Note that the set of 24 models found by EPR range from the simple model representation of the average value (equation (29)), through to the linear models of equations (27) and (28), and then on to more ornate configurations. A graphical depiction of the percentage of $X_i$ (the ratio between the total number of inputs automatically selected in the equation and the product of the total number of possible inputs specified by the user with the number of terms in the equation excluding bias) and of the number of $a_j$, together with the variation of the fitness indicator CoD are given in Figure 7.

**Table 1.** Values of the CoD Computed on the Test Set (i.e., Data Never Used During the Model Construction Stage)[a]

| | CoD 1-Month | CoD 2-Months | CoD 4-Months | CoD 6-Months | CoD 12-Months |
|---|---|---|---|---|---|
| equation (6) | 0.944054 | 0.867987 | 0.724002 | 0.636467 | 0.473369 |
| equation (7) | 0.94484 | 0.867673 | 0.726053 | 0.639654 | 0.478219 |
| equation (8) | 0.9438 | 0.866564 | 0.724675 | 0.646738 | 0.478106 |
| equation (9) | 0.944527 | 0.86688 | 0.725609 | 0.649092 | 0.479572 |
| equation (10) | 0.949937 | 0.877083 | 0.749924 | 0.672235 | 0.532415 |
| equation (11) | 0.951897 | 0.880704 | 0.74627 | 0.666323 | 0.517469 |
| equation (12) | 0.950605 | 0.885818 | 0.773038 | 0.707877 | 0.559316 |
| equation (13) | 0.952407 | 0.881456 | 0.746214 | 0.66735 | 0.520753 |
| equation (14) | 0.954039 | 0.886569 | 0.764732 | 0.698843 | 0.582876 |
| equation (15) | 0.945791 | 0.870626 | 0.729355 | 0.647163 | 0.466491 |
| **equation (16)** | **0.955403** | **0.891305** | **0.773413** | **0.716306** | **0.618291** |
| equation (17) | 0.942452 | 0.865273 | 0.713022 | 0.620872 | 0.412995 |
| equation (18) | 0.94662 | 0.876716 | 0.749949 | 0.679858 | 0.516477 |
| equation (19) | 0.954074 | 0.885667 | 0.763183 | 0.701854 | 0.605834 |
| equation (20) | 0.942818 | 0.864418 | 0.715113 | 0.623731 | 0.417457 |
| equation (21) | 0.944611 | 0.873186 | 0.741903 | 0.668758 | 0.505257 |
| equation (22) | 0.945849 | 0.868564 | 0.736682 | 0.659693 | 0.466125 |
| equation (23) | 0.938039 | 0.864983 | 0.713067 | 0.622859 | 0.395826 |
| equation (24) | 0.939169 | 0.865237 | 0.706833 | 0.618578 | 0.421792 |
| equation (25) | 0.950381 | 0.885491 | 0.758338 | 0.687343 | 0.522938 |
| equation (26) | 0.948916 | 0.874795 | 0.752142 | 0.691323 | 0.481588 |
| equation (27) | 0.94767 | 0.876503 | 0.723026 | 0.637821 | 0.445139 |
| equation (28) | 0.920425 | 0.764546 | 0.435187 | 0.232252 | −0.182268 |
| equation (29) | −0.934387 | −0.9276 | −0.938165 | −0.937898 | −0.936759 |

[a]These values give an idea of the generalization capabilities of EPR models for the case study. The bold character identifies the row corresponding to the selected equation (16).

[46] The performance in terms of CoD is evaluated on the data test set and summarized in Table 1. A bootstrap procedure [*Efron*, 1979] was applied for the CoD of the test set in order to improve the robustness of its estimation. For this purpose, the data were re-sampled 1000 times; Table 1 provides the CoD values averaged on the 1000 samples obtained for each class of prediction (1-, 2-, 4-, 6-, and 12-month).

[47] The collection of groundwater heads was affected by some missing samples in the data. This did not represent a problem for the procedure, since a model-based reconstruction was undertaken by EPR during the modeling phase [*Giustolisi et al.*, 2004a]. In particular, the missing data were interpolated using models found during the evolutionary search. However, this reconstruction did not bias the value of the fitness objective functions since the EPR-generated data were excluded prior to their evaluation.

[48] On-line predictions of the groundwater head at different time horizons are presented in Figures 8 and 9; presentation with two figures is designed to increase clarity. Note that, although the choice of the prediction horizons is motivated by management needs, a planning horizon of 12 months is reasonable for simulating the behavior of the aquifer which in turn influences the management policies that can be adopted. On the other hand, a prediction horizon of 1 month can be useful for the adoption of emergency policies, for instance related to an anomalous dry period or excessive pumping.

[49] During model construction, the objective function related to model fitness was estimated on a very short term prediction (i.e., one month ahead). This can be seen in a twofold scenario: on the one hand, it allowed the EPR procedure to estimate the CoD very rapidly; on the other, this can represent a disadvantage when models based on such a fitness criterion are employed for longer prediction horizons. Anyway, concerning the relatively long term predictions herein considered, the uncertainty related to those scenarios is such that the accuracy related to the EPR-returned models can be considered acceptable. Looking at the equation (16), the CoD related to the 12-month forecast (i.e., 12-step-ahead) is estimated as 0.618291 (note a CoD of 1 is a perfectly fitting model) on the test set (see Table 1). Such accuracy can be considered reasonable for stakeholders who want to plan the use of the aquifer for the next year without risking its overuse. For longer timescales, such as those spanning several years, it would be more fitting the use of physically based models which incorporate more variables and actions related to the evolution of the aquifer.

[50] The predictions related to the test set were never used for model construction but are related to the model in equation (16). Focus was placed on this particular choice because it exhibits the best CoD values for the test set for all the prediction horizons considered, see Table 1. In addition, its symbolic structure is similar to those of the other non-dominated models that were identified. This suggests that the terms in the selected equation could reasonably accurately reflect the physics of the phenomenon instead of being led astray by the particular realization of errors in the data.

## 4. Discussion of Results

[51] EPR identified 24 non-dominated models with differing structural complexity and performance. In a decision support scenario, it is important not only to observe the performance of each model, but their structural variations and the contiguities of inputs and monomial terms.

[52] Even if the symbolic structures returned by EPR can tempt the user to advance physical explanations, such

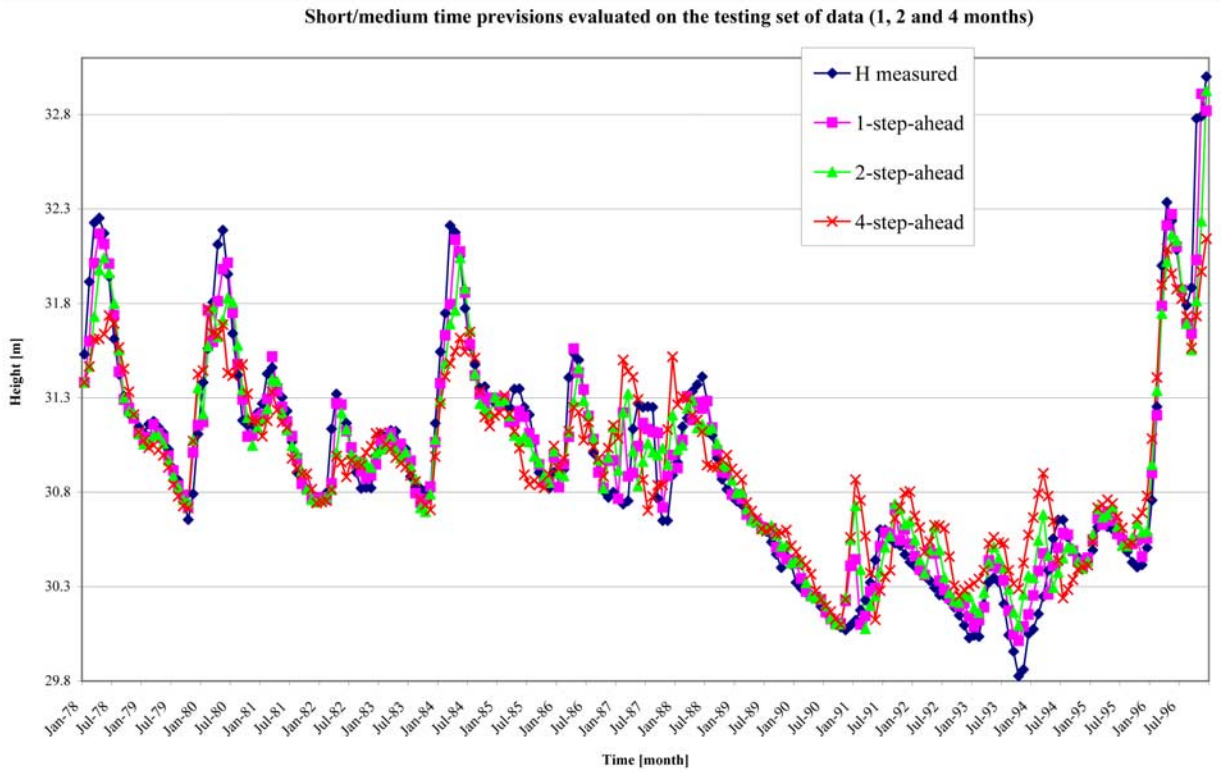


**Figure 8.** Groundwater head-on-line prediction at 1, 2, 4 months ahead computed on the test set (unseen data).

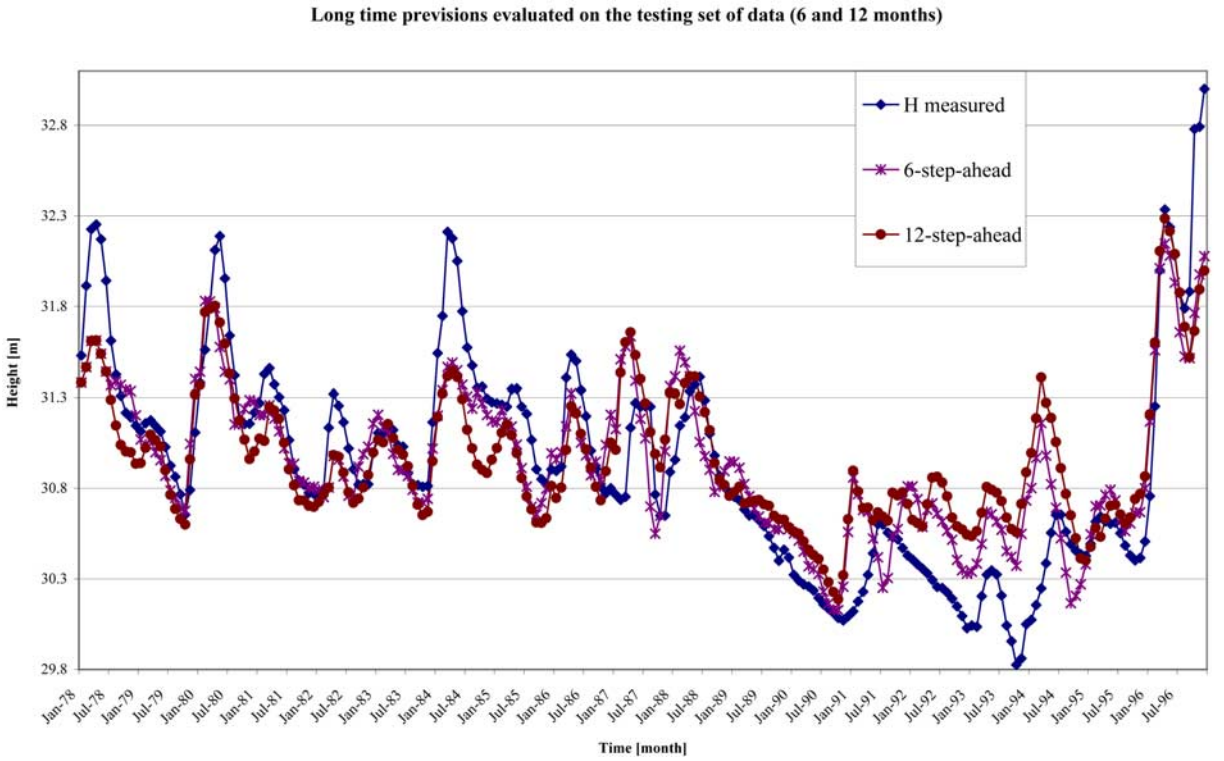


**Figure 9.** Groundwater head-on-line prediction at 6 and 12 months ahead computed on the test set (unseen data).

interpretations may be difficult to prove and are thus somewhat speculative. For this reason, only the similarities among EPR and past physically based studies dealing with the same aquifer are reported.

[53] All models encompass the term $H_{t-1}$, which relates to the persistency of groundwater head variations. The other additive terms depend solely on the rainfall input. Among these, the most frequent is $P_{t-1} \cdot P_{t-2}^{0.5}$, which is common to 13 equations and uniformly distributed on the set of equations. A physically sound interpretation of this term can be associated with the influence of the previous month's rainfall, and that of two months prior, on the head in the aquifer. Even if a physical interpretation of data-driven models is tenuous, this influence is consistent with the empirical response analysis introduced by *Ricchetti and Polemio* [1996]. It is interesting to note that there are 7 terms (common to many equations) that contain $P_{t-12}$, which is the key component related to the rainfall observed one year previously.

[54] The term $H_{t-1}$ is always linear with the exception for model (29), which is the simplest non-dominated model, and it represents the average value of the samples in the training set. Although such a model has no apparent meaning, it actually represents an extreme solution on the Pareto front. The remaining models present increasing levels of structural parsimony up to equation (6), which is the most complex according to its ranking by EPR and represents the other extreme solution on the front; though it scores the highest CoD on the training data, it does not exhibit the best fitness for the test set.

[55] The analysis of the monomial terms belonging to the models along the Pareto front suggests an evolution of these components along the front deriving from a gradual pruning of those inputs considered superfluous by the parsimony criteria. Following this line of reasoning, it is interesting to note that $P_{t-12}$ is never considered superfluous.

[56] As previously mentioned, the exponents were chosen from the set {0; 0.5; 1; 2}, with their possible meanings introduced in Paragraph 3.2. Some inputs were never chosen, but the majority (9 inputs) were selected in all models. Inputs completely discarded by EPR are 5: $H_{t-2}$, $P_{t-7}$, $P_{t-8}$, $P_{t-9}$ and $P_{t-11}$. It is also interesting to observe that inputs $P_{t-1}$ to $P_{t-4}$ were the most frequently selected, in agreement with results already reported in the literature [*Ricchetti and Polemio*, 1996].

[57] With respect to the parameters pertaining to the optimization process, such as number of generations, crossover and mutation rate and population size, it was observed that their relatively wide variation did not strongly influence results. In particular, increasing the number of generations did not produce better results, thus the returned Pareto front presumably represented a good approximation of the best set of non-dominated models. Similarly, variations of the crossover and mutation rates were observed to bear little influence on both the results and on the searching space. However, the variation of these parameters are bounded to a maximum of +0.3 or −0.2 for the crossover rate and +0.3 for the mutation. The population size was chosen as good compromise between search efficiency and the need to generate a good random starting point. Nonetheless, an increasing of the population size was observed to delay the process without proffering any significant advantage. Conversely, reducing the population size by an appreciable amount compromised the initial set of models, leading to poorer results.

[58] Finally, some brief comments on equation (16) are in order. As was previously discussed, this equation was chosen for two reasons: (1) it scores the highest fitness for the test set, and (2) despite its apparently simple structure, it encompasses those terms and inputs that are common to the other non-dominated equations that were identified. Equation (16) contains the product $P_{t-1} \cdot P_{t-2}^{0.5}$ as a component, the second most diffused term among the equations. Moreover, it includes the term $H_{t-1}$, which is common to all the equations save (29), and two additional terms. The remaining terms are $P_{t-3}^2 \cdot P_{t-4} \cdot P_{t-12}^2$ and $P_t^2$: the former is essentially common to other equations; however, a physically sound interpretation of these terms is not realistic. Equation (16) also contains the term $P_t^2$, which introduces a hypothetical correlation between the present groundwater level and the current rainfall. Overall, this term is uncommon since it is accounted for, alone or combined with other terms, in two and seven additional equations, respectively. This input does not have a direct physical interpretation; in fact, those blocks containing $P_t^2$, alone or together with other inputs, are progressively expunged moving toward the region of the Pareto front where the parsimony-related objective prevails. In fact, commencing with equation (19), those terms progressively disappear. Thus in summary, equation (16) is quite fit while being very similar to the other equations identified in the process. Furthermore, it is characterized by low structural complexity, both in terms of the number of selected inputs as well as the number of inputs involved in each term.

## 5. Conclusions

[59] The use of EPR in support of groundwater resource planning is described. In particular, the procedure was applied to capture the dynamic relationship between groundwater heads (output) and rainfall depths (input). This showed to be potentially useful for planning purposes in situations where the availability of data is limited and boundary conditions are complex are unknown with accuracy. The case study related to an existing aquifer illustrates its effectiveness in a typical situation; that is, a heavily relied upon groundwater resource and the general absence of a clear aquifer management strategy. Moreover, the scenario is particularly interesting because it is not readily approachable due to lacunae in groundwater data and non-uniform data collection [*Yi and Lee*, 2004], in addition to the nonlinearity imposed by vadose zone effects.

[60] In a wider sense, the application of EPR to the study of this phenomenon suggested the possible advantages of evolutionary multiobjective modeling in general. These are: (1) the expressions cover a wide range of solutions which represent the best models for different structural complexities, (2) several important aspects in the analysis of an environmental system are considered as evident in the analysis of the Pareto front of equations and (3) the algorithm is more computationally efficient compared with

the multiple single-objective runs for separately analyzing fitness and complexity. These features allow the user to select from among a robust group of models, since a comprehensive set of possible structures can be developed for each purpose. Even if a single model is ultimately settled upon, a wide range of models can be helpful for understanding which terms/inputs are physically meaningful and which can comfortably be eschewed for the sake of model parsimony, while simultaneously striving for a degree of generality.

## Notation and Acronyms (as encountered in the text)

| | |
|---|---|
| ARX | AutoRegressive eXogenous model |
| GA | Genetic Algorithm |
| SO | Single Objective |
| MO | Multiobjective |
| EPR | Evolutionary Polynomial Regression |
| GP | Genetic Programming |
| LS | Least Square |
| $a_j$ | Model parameters for EPR expressions |
| $X_i$ | vectors of candidate inputs for EPR |
| **ES** | matrix of candidate exponents for EPR |
| $f$ | used-specified function |
| $g$ | user-specified function |
| m | length of EPR expressions |
| $y$ | estimated output of the generic system/process |
| $j$ | subscript associated to constant values |
| $x_i$, | inputs selected by the process among the user-specified range of candidates $\mathbf{X}_i$ |
| $\alpha$, $\beta$, $\gamma$, $\delta$ | process-selected exponents from a set pre-specified by the user |
| CoD | Coefficient of Determination |
| $\overline{x}$* | vector of decision variables |
| $g_i$ | set of inequality constraints |
| $h_i$ | set of equality constraints |
| $\overline{f}(\overline{x})$ | objective functions |
| **F** | design space |
| TFN | Transfer Function-Noise |
| $H_{t-i}$ | values of groundwater head |
| $P_{t-i}$ | values of rainfall heights |
| $t$ | sampling time |
| SSE | Sum of Square Errors |
| N | number of samples |
| $\widehat{H}$ | model returned groundwater height |
| $H_{\text{exp}}$ | measured groundwater height |
| avg | average value |

[61] **Acknowledgments.** The authors would like to thank the Technical University of Bari and in particular the Engineering Faculty of Taranto for granting this paper. Moreover, the authors are grateful to Davide Mancarella for his contribution to the section named "Background to data" and Andrew Colombo for his assistance in editing the text.

## References

Babovic, V., and M. Keijzer (2000), Genetic programming as a model induction engine, *J. Hydroinformatics*, *2*(1), 35–61.

Bierkens, M. F. P., M. Knotters, and T. Hoogland (2001), Space-time modeling of water table depth using a regionalized time series model and the Kalman filter, *Water Resour. Res.*, *37*(5), 1277–1290, doi:10.1029/2000WR900353.

Coello Coello, C. A. (1999), A comprehensive survey of evolutionary-based multiobjective optimization techniques, *Knowledge and Information Systems*, *1*(3), 269–308.

Coello Coello, C. A., D. A. Van Veldhuizen, and G. B. Lamont (2002), *Evolutionary algorithms for Solving Multi-Objective Problems*, Kluwer Academic, New York.

Coppola, E. A., L. Duckstein, and D. Davis (2002), Fuzzy rule-based methodology for estimating monthly groundwater recharge in a temperate watershed, *J. Hydrol. Eng.*, *7*(4), 326–335, doi:10.1061/(ASCE)1084-0699(2002)7:4(326).

Custodio, E. (2002), Aquifer overexploitation: What does it mean?, *Hydrogeology J.*, *10*(2), 254–277.

Davidson, J. W., D. A. Savic, and G. A. Walters (2003), Symbolic and numerical regression: Experiments and applications, *Information Sciences*, *150*(1/2), 95–117.

Deb, K. (2001), *Multi-Objective Optimization using Evolutionary Algorithms*, John Wiley & Sons, Chichester, UK.

Efron, B. (1979), Bootstrap methods. Another look at the jackknife, *The Ann. of Statist.*, *7*, 1–26.

Giustolisi, O., and D. A. Savic (2006), A symbolic data-driven technique based on evolutionary polynomial regression, *J. Hydroinformatics*, *8*(3), 207–222.

Giustolisi, O., and V. Simeone (2006), Multi-objective strategy in artificial neural network construction, *Hydrol. Sci. J.*, *3*(51), 502–523.

Giustolisi, O., D. A. Savic, and A. Doglioni (2004a), Data Reconstruction and Forecasting by Evolutionary Polynomial Regression. In Proceedings of the 6th International Conference on Hydroinformatics 2004, Liong, Phoon, and Babovic, eds., World Sci. Publishing Company, Singapore.

Giustolisi, O., A. Doglioni, D. Laucelli, and D. A. Savic (2004b), A proposal for an effective multiobjective non-dominated genetic algorithm: the OPTimised Multi-Objective Genetic Algorithm, OPTIMOGA, *Report 2004/07*, School of Engineering Computer Science and Mathematics, Centre for Water Systems, Univ. of Exeter, UK.

Giustolisi, O., A. Doglioni, D. A. Savic, and B. W. Webb (2007), A multi-model approach to analysis of environmental phenomena, *Environmental Modelling & Software*, *22*(5), 674–682.

Goldberg, D. E. (1989), *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison Wesley.

Jones, R. H. (1980), Maximum likelihood fitting of ARMA models to time series with missing observations, *Technometrics*, *22*, 389–395.

Jones, L., R. Willis, and W. W. G. Yeh (1987), Optimal control of nonlinear groundwater hydraulics using differential dynamic programming, *Water Resour. Res.*, *23*, 2097–2107.

Knotters, M., and M. F. P. Bierkens (2000), Physical basis of time series models for water table depths, *Water Resour. Res.*, *36*(1), 181–188, doi:10.1029/1999WR900288.

Koza, J. R. (1992), *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. MIT Press, Cambridge, MA, USA.

Lawson, C. L., and R. J. Hanson (1974), *Solving Least Squares Problems*, 161 pp., Prentice-Hall.

Ljung, L. (1987), *System Identification: Theory for the User*. Prentice-Hall, Englewood Cliffs, New Jersey.

McKinney, D. C., and M. D. Lin (1994), Genetic algorithm solution of groundwater management models, *Water Resour. Res.*, *30*(6), 1897–1906.

McPhee, J., and W. W. G. Yeh (2006), Experimental design for groundwater modeling and management, *Water Resour. Res.*, *42*(2), W02408, doi:10.1029/2005WR003997.

Pareto, V. (1896), *Cours D'Economie Politique*., Rouge and Cic, Vol. I and II, Lausanne, Switzerland.

Ricchetti, E., and M. Polemio (1996), L'acquifero superficiale del territorio di Brindisi: dati geoidrologici diretti e immagini radar da satellite (The shallow aquifer of Brindisi: geohydrological direct data and satellite radar images), *Memorie della Società Geologica Italiana*, *51*, 1059–1074, 11 ff.

Siegfried, T., and W. Kinzelbach (2006), A multiobjective discrete stochastic optimisation approach to shared aquifer management: Methodology and application, *Water Resour. Res.*, *42*(2), W02402 doi:10.1029/2005WR004321.

Simonovic, S. P. (2000), Tools for water management: One view of the future, *Water International*, *25*(1), 76–88.

Soule, T., and R. B. Heckendorn (2002), An analysis of the cause of the code growth in genetic programming, *Genetic Programming and Evolvable Machines*, *3*, 283–309.

Van Veldhuizen, D. A., and G. B. Lamont (2000), Multiobjective Evolutionary Algorithms Analyzing the State-of-the-Art, *Evolutionary Computation*, *8*(2), 125–144.

Wang, M., and C. Zheng (1998), Application of genetic algorithms and simulated annealing in groundwater management: formulation and comparison, *J. Am. Water Resour. Assoc.*, *34*(3), 519–530.

Yi, M.-J., and K. K. Lee (2004), Transfer function-noise modelling of irregularly observed groundwater heads using precipitation data, *J. Hydrol.*, *288*, 272–287.

Young, P., S. Parkinson, and M. Lees (1996), Simplicity out of complexity in environmental modelling: Occam's razor revisited, *J. Applied Statistics*, *23*(2–3), 165–210.

F. di Pierro and D. A. Savic, Centre for Water Systems, Department of Engineering, University of Exeter, Harrison Building, North Park Road, Exeter EX4 4QE, UK. (f.di-pierro@ex.ac.uk; d.savic@ex.ac.uk)

A. Doglioni, Department of Environmental Engineering and Sustainable Development, Technical University of Bari, Engineering Faculty of Taranto, V.le del Turismo n. 8, 74100 Taranto, Italy. (a.doglioni@ poliba.it)

O. Giustolisi, Civil and Environmental Engineering Department, Technical University of Bari, Engineering Faculty of Taranto, V.le del Turismo n. 8, 74100 Taranto, Italy. (o.giustolisi@poliba.it)