A novel hybrid mechanistic-data-driven model identification framework using NSGA-II

Soroosh Sharifi and Arash Massoudieh

ABSTRACT

This paper describes a novel evolutionary data-driven model (DDM) identification framework using the NSGA-II multi-objective genetic algorithm. The central concept of this paper is the employment of evolutionary computation to search for model structures among a catalog of models, while honoring the physical principles and the constitutive theories commonly used to represent the system/processes being modeled. The presented framework provides high computational efficiency through connecting a series of NSGA-II runs which share results. Furthermore, the employment of a multi-objective optimization algorithm enables a unique way of incorporating different aspects of model goodness in the model selection process, and also, at the end of the search procedure, provides a number of potential optimal model structures, making it possible for the modeler to make a choice based on the goal of the modeling. As an illustration, the framework is used for modeling wash-off and build-up of suspended solids (TSS) in highway runoff. The performance of the discovered model confirms the potential of the proposed evolutionary DDM framework for modeling environmental processes.

Key words | data-driven modeling, evolutionary computation, genetic algorithms, NSGA-II, symbolic regression

Soroosh Sharifi
Arash Massoudieh (corresponding author)
The Catholic University of America,
620 Michigan Ave. NE., Washington, DC 20065,
USA
E-mail: *Massoudieh@cua.edu*

NOTATION

A	constant coefficient
C	concentration of mobile constituents in the runoff [M/L^3]
$\mathbf{E}$	vector containing influencing variables
$f_i(X)$	objective function
G	unified model
h	the thickness of the sheer flow [L]
i	rain intensity [L/T]
k	mass exchange rate coefficient
$\mathbf{K}$	the vector representing the controlling processes
L	basin length [L]
m	weighting coefficient
n	number of observed data
n	Manning's roughness coefficient
N	number of archived solutions
P_t	parent population
q	flow-rate per unit length [L^2/T]
Q_t	offspring population
R_t	combination of parent and offspring population
s	slope of the pavement normal to the highway axis [L/L]
S	concentration of attached or adsorbed constituents to the pavement surface [M/L^2]
t	Time
u	flow velocity [L/T]
x	model input variable/distance [L]
$\mathbf{X}$	set of observed data
X	space vector
y	model output variable
Y	set of computed data
$\mathbf{Y}$	vector of outputs
α	exponent in Manning's equation
β	coefficient of attachment and release equations
δ	termination criterion

doi: 10.2166/hydro.2012.026

η constant
μ exponent of attachment and release equations
φ constant
Ω design domain search space
Ω_o objective domain search space
ω constant

SUBSCRIPTS

a attachment
c calibration dataset
f fast release sorption site
r release
s slow release sorption site
t testing dataset

ACRONYMS

COD	Coefficient of determination
DDM	Data-driven modeling
EA	Evolutionary algorithm
EC	Evolutionary computation
EMC	Event mean concentration
EPR	Evolutionary polynomial regression
GA	Genetic algorithm
GP	Genetic programming
MOGA	Multi-objective genetic algorithm
NSGA-II	Non-dominated sorting genetic algorithm II
PDE	Partial differential equation
TSS	Total suspended solids

INTRODUCTION

Data-driven modeling (DDM) has been used for modeling a wide range of systems, including complex environmental phenomena, which are often governed by highly non-linear processes. In contrast to knowledge-driven (also known as physically-based) modeling, which is typically based on the 'physical' principles controlling the system, DDM is based on the analysis of data representing the inputs and outputs of the system. DDM generally involves employing computational intelligence and machine learning techniques to build mathematical models through analysis of coexisting input–output data, while making none or a limited number of assumptions about the 'physical' behavior of the system (Solomantine & Ostfeld 2008). Statistical methods (e.g. Lemke *et al*. 2006), artificial neural networks (e.g. Hsu *et al*. 1995; Sousa *et al*. 2007), fuzzy rule-based systems (e.g. Bardossy *et al*. 1990) and evolutionary algorithms (EAs) (e.g. Giustolisi & Savic 2006; Sharifi *et al*. 2011) are among the many common methods used in DDM of environmental processes.

Evolutionary computation (EC) techniques, such as genetic algorithm (GA) (Holland 1975) and genetic programming (GP) (Koza 1992), are robust and adaptive search methods that have been widely used as the running engine of DDM approaches. Having the advantage of revealing the underlying relationships of the system in a white box model, GP is perhaps the most widely used evolutionary method for model induction. GP is able to discover the optimal mathematical functional form that fits the data, as well as the appropriate related numeric coefficients through an evolutionary process. Many applications of GP in environmental modeling have been reported in the literature (Poli *et al*. 2008). Following the work of Davidson *et al*. (1999, 2000), Giustolisi & Savic (2006) and Giustolisi *et al*. (2007) developed an evolutionary DDM method called evolutionary polynomial regression (EPR). In their method, GP was combined with the features of conventional numerical regression to generate suitable polynomial regression equations for describing a phenomenon. Later on, to overcome the shortcomings of the single objective EPR method, including computational efficiency and model structure complexity, EPR was integrated with a multi-objective (MO) optimization strategy based on the Pareto dominance criterion and the EPR- multi-objective genetic algorithm (MOGA) was introduced (Giustolisi & Savic 2009). MO-EPR has the additional advantages of searching for different goals and giving a set of models, ranked on their performance and structural complexity, at the end of each search (El-Baroudy *et al*. 2010). EPR overcomes some of the shortcomings of GP, such as the complexity of the resulting symbolic models and the computational requirements (Davidson

et al. 1999, 2000; Solomantine & Ostfeld 2008) and has been used for environmental modeling (e.g. Giustolisi & Savic 2006; Giustolisi *et al.* 2007).

In many environmental modeling problems, it is important to construct the model based upon the conceptual representation of the processes deemed to control the system while honoring the physical principles as well as the constitutive relationships commonly used. This mathematical representation can stem from the commonly accepted physical, chemical or biological processes, such as advection-dispersion-reaction or mass exchange between various phases, natural physical balance laws, such as conservation of mass, momentum and energy, often represented by one or a few coupled partial differential equations (PDEs). It is theoretically possible that traditional data modeling approaches, such as GP, find model structures that to some degree honor the physical principles governing a system; however, due to the usual high-dimensional search space and limited computational resources and also their intrinsic attributes (e.g. searching for models with higher parsimony) the chances of finding such model structures are very small. To fill this gap, in this paper, a novel hybrid mechanistic-data-driven model identification framework is proposed that uses EC techniques to select among models that honor the physical principles and the common constitutive relationships for a system. This proposed framework works on the basis of: (1) selecting a comprehensive generalized model structure for the processes; (2) employing a MOGA to search in the model structure space; and (3) using a hybrid GA to calibrate the models (i.e. to find optimum values for the constants) and assign measures of goodness of fit to each model. The final set of models obtained from this method become the candidate models that best describe the phenomena.

This paper is organized as follows: The first section provides a brief background on topics which are essential to understanding the proposed DDM framework, i.e. the basics of MO optimization and the MOGA employed in the framework. Then, the evolutionary model identification framework is described in detail. Next, in a case study, the proposed method is used for identifying a model for highway contaminant accumulation, fate and transport. Finally, appropriate conclusions relating to the work are presented.

BACKGROUND

Evolutionary computation

Inspired by Darwin's theory of natural evolution and motivated by the development of computer technologies, EC was introduced in the 1960s as a robust and adaptive search method (Back *et al.* 1997). Simulating the natural evolutionary process and mimicking the main principle of 'survival of the fittest'. these techniques are able to search for the best ('fittest') solution(s) among a large number of possible candidates. Conceptually, all EC methods are based on initializing a random population of potential candidates (chromosomes) using a coding scheme, evaluating each individual within the population and giving the fitter solutions higher chances of 'survival' or 'reproduction'. In the search for the best solution, evolution tries to gradually improve the quality of individuals by selecting, recombining (crossover) and altering (mutation) the fittest individuals. Any algorithm that adopts this general procedure is called an EA. More than 30 years of practical application of EC in various fields has demonstrated that it is capable of dealing with a large variety of problems (Back & Schwefel 1993). For an in-depth description of EC and its extended branches of derivatives, the reader is referred to Holland (1975), Goldberg (1989), Coley (1999), Koza (1992), Michaelwicz (1992), Back *et al.* (1997), Osyczka (2002) and De Jong (2006).

In this research, two EC techniques, a MOGA called non-dominated sorting genetic algorithm II (NSGA-II) (Deb *et al.* 2000, 2002) and a hybrid GA (Fan *et al.* 2006; Massoudieh *et al.* 2008), are combined to structure a model identification framework. For the sake of brevity, only a brief description of the basics of MO optimization and the NSGA-II method are provided in the following sub-sections.

Multi-objective optimization

The concept of optimizing multiple, but equally important, objective functions was originally introduced by two economists, Edgeworth and Pareto (Chipman 2006). In contrast to single objective optimization problems, MO optimization problems may not have a single solution that simultaneously optimizes all objectives (Hirschen & Schafer 2006). In fact,

there exists a set of equally good optimum solutions (trade-offs), none of which, without any further information or subjectivity, can be said to be better than the others in terms of one or more aspects.

In order to distinguish between better and worse solutions of a MO problem, it is necessary to rank them according to an order criterion. The Pareto dominance concept (Goldberg 1989) is commonly used to compare two solutions and to classify them as dominated or non-dominated solutions. Based on this concept, for any two solutions (e.g. X_1 and X_2) in the feasible design domain search space, Ω, assuming a minimization problem, X_1 is said to dominate X_2 if:

$$\begin{aligned} &f_i(X_1) \le f_i(X_2) \text{ for all } i \in [1,2,\ldots M] \\ &\text{and} \\ &f_j(X_1) < f_j(X_2) \text{ for at least one } j \in [1,2,\ldots M] \end{aligned} \quad (1)$$

where f_i is the ith objective function and M is the total number of objectives. The union of all non-dominated solutions is called the 'Pareto optimal set' of solutions and its corresponding image in the objective space, Ω_0 (i.e. the space where candidate solutions are projected through the objective functions) is known as the 'Pareto optimal front'.

To promote the required diversity and to more effectively search the solution space, the concept of ε-dominance was introduced by Laumanns *et al.* (2002). Based on this concept, for any two solutions (e.g. X_1 and X_2) in the feasible design domain search space, Ω, assuming a minimization problem, X_1 is said to ε-dominate X_2 if for $\varepsilon > 0$:

$$(1-\varepsilon) f_i(X_1) \le f_i(X_2) \text{ for all } i \in [1,2,\ldots M] \quad (2)$$

where ε is a user-specified precision vector. Figure 1 illustrates the concepts of dominance and ε-dominance for a minimization problem.

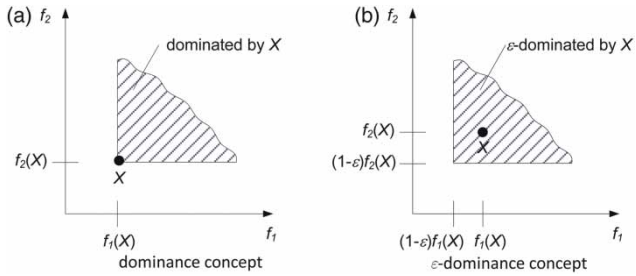

Figure 1 | Visualizing the concepts of (a) dominance (b) ε-dominance (adopted from Laumanns *et al.* 2002).

NSGA-II

1. Due to their intrinsic nature of evolving a set of possible solutions (population), which enables them to find multiple members of the Pareto optimal set in a single run of the algorithm, EC techniques including GAs are ideal candidates for solving MO problems (Fonseca & Fleming 1995). Among the many MOGA proposed in the literature, a second generation MOGA called the NSGA-II (Deb *et al.* 2000, 2002) was selected to be used in the model identification framework. Initially introduced by Deb *et al.* (2002), NSGA-II is an efficient MOGA that follows an 'elitism strategy', and employs a fast, non-dominated sorting algorithm, making it capable of finding multiple Pareto solutions in a single run. The main features of this method are expressed as follows (Deb *et al.* 2002): An elite-preserving operator is used to preserve the best found solutions of each generation and to include them in the next generation.
2. The non-dominated Pareto optimal fronts are sorted using a fast non-dominated sorting algorithm (Deb *et al.* 2002). This approach increases the convergence speed significantly by reducing the computational burden.
3. A two-level ranking method is used to assign the effective fitness of solutions during the selection process. At first, solutions are sorted according to their dominance level and are organized into fronts (i.e. groups of similar Pareto-optimality). Each group is then assigned a rank based on their optimality level. Subsequently, within each front, individual solutions are ranked according to a density measure using the crowding operator. This operator measures the distance of each solution with respect to the solutions that surround it on its front. Consequently, the solutions residing in less crowded regions of the objective space are given higher ranks, which ultimately results in a better diversity of individuals.

The procedures involved in the NSGA-II algorithm are depicted in Figure 2. In this figure, P_t is the parent population, Q_t is the offspring population, R_t is the combined

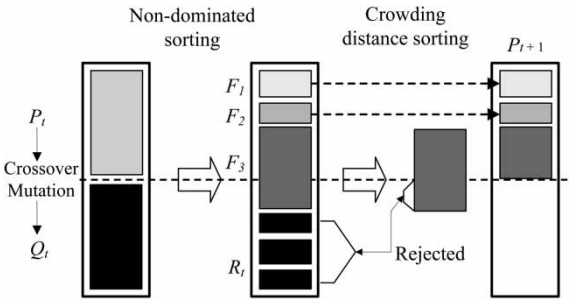

Figure 2 | Procedure of NSGA-II (adopted from Deb *et al.* 2002).

population ($R_t = P_t \cup Q_1$) and F_i are the sorted Pareto fronts of R_t. For an in-depth explanation of this method, the reader is referred to Deb *et al.* (2002).

Studying a variety of test cases (e.g. Deb *et al.* 2002; Khare *et al.* 2003), it has been shown that compared to other elitist multi-objective EAs, NSGA-II has a better diversity preservation and therefore is able to compete with them in terms of convergence to the true Pareto optimal front in both constraint and non-constraint problems (Nazemi *et al.* 2006). This superiority has led to the successful application of NSGA-II in several real-world problems, such as long-term groundwater monitoring design (Reed *et al.* 2007), water distribution network design (Babayan *et al.* 2005), and calibrating hydrological models (Bekele & Nicklow 2007).

The following section addresses a novel application of NSGA-II to DDM: this algorithm is used as the running engine of a hybrid model identification framework.

HYBRID MECHANISTIC-DATA-DRIVEN MODEL IDENTIFICATION FRAMEWORK

The main objective of this research is to develop an efficient mechanistic-data-driven model identification framework. The novelty of the proposed framework is two-fold: first, having the ability to search for models that incorporate any considered physically based or constitutive relationships of the system; and second, employing a MO search procedure which brings into account several measures of the goodness of the model, including its capability in reproducing the observed data and its degree of *transferability* (i.e. the degree to which the model can be generalized or transferred to similar situations and settings). In summary, the framework is a hybrid that combines two approaches: (1) using NSGA-II to search the feasible model structure space; and (2) using a GA to calibrate the models. In the following subsections, the elements and procedures employed in the proposed DDM framework are presented in detail.

Model structure identification

The first step in the framework is to identify the most general form of the model intended to be included in the model identification. This general form, which encompasses numerous sub-models comprising the category of models to be used in model identification, is henceforth referred to as the 'unified' model structure. This task consists of (Evsukoff *et al.* 2006): (1) selecting the model *type*, i.e. the main relationships of input and output variables; (2) selecting the model *size* in terms of the number of input and output variables; and (3) selecting the model's *parameterization* according to model type and size, i.e. determining the final structure of the unified model. Equation (3) shows the formulation of a unified model in its most general form:

$$G(\mathbf{Y}, t, \mathbf{X}, \mathbf{E}, \mathbf{K}) = 0 \tag{3}$$

where G can be a single or a set of functions, operators (i.e. partial derivative with respect to the independent variables $\mathbf{X}$ or t, etc.) or a combination of them derived based on the underlying physical principles or constitutive theories governing the system (e.g. mass balance or advection-dispersion). $\mathbf{Y}$ is the vector of outputs, t is time, $\mathbf{X}$ is the space vector, $\mathbf{E}$ is the vector containing other influencing variables (e.g. rain intensity, shear stress, etc.) and $\mathbf{K}$ is the vector representing the controlling processes, (e.g. mass exchange rate, etc.) referred to as process functions. In general, the components of the $\mathbf{K}$ vector can be functions of outputs, external parameters, time and space:

$$\mathbf{K} = \begin{bmatrix} \cdot \\ \cdot \\ K_i = K_i(\mathbf{E}, \mathbf{Y}, \mathbf{X}, \mathbf{t}; A_{0i}, A_{1i}, \ldots) \\ \cdot \\ \cdot \end{bmatrix} \tag{4}$$

where K_i is a function controlling process i which contains unknown constants A_{ij}. The form of the K_i functions are

determined so that elements of the function can be removed by dropping some of the A_{ij} constants (e.g. a combination of additive, multiplicative or power forms) resulting in sub-models from the unified model structure. For example:

$$K_i = A_{0i} + A_{1i}E_{1i}^{A_{2i}} + A_{3i}A_{2i}^{A_{4i}} + A_{5i}x^{A_{6i}} \quad (5a)$$

$$K_j = A_{1j} + E_{1j}^{A_{2j}} \quad (5b)$$

The feasible model search space, Ω, of a unified model is the set of all of its meaningful sub-models. The goal of the model identification is to find the optimal set of non-zero A_{ij} constants that leads to the best model in terms of reproducing the measured data available for $\mathbf{Y}$, and transferability as defined by the relative decline in model fit when it is applied to unseen data.

Model search process

Representation (encoding)

Binary integer coding is used to identify sub-models from the unified model structure. Each individual (sub-model) is encoded by a binary string chromosome where each gene represents the function controlling a process (K_i) and each bit in the gene acts as a switch, determining which of the function's constants (A_{ij}) will be active (bit value = 1) or inactive (bit value = 0). For example, considering Equations (5a) and (5b) as the elements of the $\mathbf{K}$ vector, the individuals would be encoded using a 9-bit binary string. Examples of encoded individuals are illustrated in Figure 3. To initialize the search process, a set of models is produced by generating a number of random binary strings.

Model calibration (parameter estimation)

The parameter estimation is performed on a sub-set of observed data called the 'calibration set'. Once a set of random models are generated, for each generated model structure, the constants (A_{ij} in Equation (4)) are estimated via maximization of a linear combination of the square of Pearson's correlation coefficient, r_{XY}^2, and the coefficient of determination (COD):

$$f(X,Y) = m_1(r_{XY}^2) + m_2(\mathrm{COD}) = m_1\left(\frac{(\mathrm{Cov}(X,Y))^2}{\mathrm{Var}(X)\mathrm{Var}(Y)}\right) + m_2\left(1 - \frac{\sum_n (Y-X)^2}{\sum_n (X - \mathrm{avr}(X))^2}\right) \quad (6)$$

where n is the number of observed data, m_1 and m_2 are weight coefficients that can be adjusted according to the modelers choice and X and Y are the sets of observed and computed data, respectively. It is acknowledged that other measures of goodness of fit could have been used. However, it is felt that those listed above are a good choice leading to an optimum model. This is because r_{XY}^2 is an indicator of how good the model captures the trends in the observed data, while COD measures how good the model captures the magnitude of the difference between the observed data and model outputs.

The evolutionary DDM framework employs a hybrid GA, based on Fan *et al.* (2006), for parameter estimation. The hybrid GA utilizes a combination of three optimization methods, including simple binary GA, adaptive random numerical shaking (Massoudieh *et al.* 2008)

Chromosome	Sub-model
1 1 1 1 1 1 1 1 1	$K_i = A_{0i} + A_{1i}E_{1i}^{A_{2i}} + A_{3i}E_{2i}^{A_{4i}} + A_{5i}x^{A_{6i}}, \quad K_j = A_{1j}E_{1j}^{A_{2j}}$
0 1 1 1 1 1 1 1 1	$K_i = A_{1i}E_{1i}^{A_{2i}} + A_{3i}E_{2i}^{A_{4i}} + A_{5i}x^{A_{6i}}, \quad K_j = A_{1j}E_{1j}^{A_{2j}}$
1 1 0 0 0 1 1 1 0	$K_i = A_{0i} + A_{1i}E_{1i} + A_{5i}x^{A_{6i}}, \quad K_j = A_{1j}E_{1j}$
0 0 1 1 0 1 0 0 1	$K_i = E_{1i}^{A_{2i}} + A_{3i}E_{2i} + A_{5i}x, \quad K_j = E_{1j}^{A_{2j}}$
1 0 0 1 1 0 0 0 0	$K_i = A_{0i} + A_{3i}E_{2i}^{A_{4i}}$

Figure 3 | Examples of the integer binary coding.

and an elite member enhancement using the SIMPLEX algorithm (Nelder & Mead 1965). The method starts the calibration process with the binary GA. For this purpose, the real numbers representing the model parameters with four decimal points are transformed into binary number. The sequences of binary numbers representing the parameters of one parameter set compose the chromosomes. A single-point cross-over is used to generate the offsprings. Mutation is applied by randomly changing some of the genes. A mutation probability of 0.008 was found to perform well and was used. When the improvement in the maximum fitness of the population during several generations is stalled, the best parameter set is modified using a series of random shaking steps, i.e. a Gaussian random number is added to all the values of parameter (i.e. phenotypes), and the new parameter set is accepted only if it results in a better solution. The standard deviation of the Gaussian distribution used in generating the random numbers is reduced in a stepwise manner. This helps the GA to explore the local neighborhood of the best parameter set in the population for possible better solutions. When improvement in the solution is gained through the random shaking step, the algorithm goes back to simple binary GA. The SIMPLEX method is applied to the best individual in any instance when neither the binary GA nor the random shaking is capable of improving the maximum fitness.

Model evaluation and selection

In a practical sense, a good model should not only be able to reproduce the observed data well, but should also be *transferable*. Employing a MO framework enables simultaneous assessment of these two qualities of the model by evaluating more than one goodness of fit criterion at a time. To evaluate the transferability of models, each of the calibrated models are evaluated based on another set of data from the system, namely the 'testing set' and the corresponding r^2_{XY} and COD values are obtained. The following two measures of fitness, aimed at assessing the predictive quality of the models for both calibration and testing datasets, were chosen as quantitative measures of the goodness of a model, and therefore, used as the objective functions to be maximized:

$$f_1(X) = m_3(r^2_{XY}|_c + \mathrm{COD}_c) + m_4(r^2_{XY}|_t + \mathrm{COD}_t) \tag{7}$$

$$f_2(X) = \mathrm{Min}\left\{ \begin{array}{c} (r^2_{XY}|_t + \mathrm{COD}_t)/(r^2_{XY}|_t + \mathrm{COD}_t \\ +r^2_{XY}|_c + \mathrm{COD}_c) \\ 0.5 \end{array} \right\} \tag{8}$$

where m_3 and m_4 are weighting coefficients and the subscripts c and t denote the calibration and testing datasets, respectively. Equation (7) is the weighted sum of model's r^2_{XY} and COD for calibration and testing sets. Furthermore, Equation (8) gives the normalized performance of the model for the test data relative to the overall model performance on the calibration and the test data. Since there is no additional benefit when the model does a better job on the testing data compared to the calibration data, the maximum value this objective function can take is set to 0.5.

Once the objective functions are evaluated for all models (individuals), they are ranked based on their non-domination level and the crowding measure. Then, a binary tournament selection method (Goldberg 1989) is used to select the parents for reproduction. Single point uniform crossover and mutation operators (Coley 1999) are used in this implementation of NSGA-II to generate offspring. Then, as explained in the background section, the set of initial populations are combined with the offspring. The accumulated set is then sorted again, and the population of the next generation is selected.

Improving the search efficiency and convergence speed

Since the ability of EAs, including NSGA-II, in converging towards the global optimum(s) can be affected by the values of the internal parameters as well as the termination criteria, to ensure a comprehensive and robust search of the objective space, multiple runs of the search with different seeds for the initial population are normally performed (Lee & Doong 2008; Sharifi 2009). Performing replicate searches is often computationally expensive and time consuming, especially if the system is complex and the function evaluations are expensive. Furthermore, EAs

require the user to define its internal parameters, such as population size, number of generations, probability of crossover, probability of mutation, etc. The traditional method of obtaining optimum internal parameters requires intensive trial-and-error analysis, which might be time consuming. To minimize the computational time, and to improve the efficiency of the model search algorithm, a slightly modified version of the ε-dominance archiving and automatic parameterization techniques, introduced by Reed *et al.* (2003, 2007), was adopted. The employed technique attempts to approximate the Pareto optimal set by connecting a series of NSGA-II runs as explained in the following steps:

1. The evolutionary search is initiated by running an NSGA-II with a small population (i.e. four individuals). The probabilities of crossover and mutation are set as 0.5 and 1/population size, respectively.
2. To maintain a diverse representation of the Pareto optimal set, an archive is formed to store the non-dominated solutions generated in every generation of the NSGA-II runs. The concept of ε-dominance (Laumanns *et al.* 2002; Deb *et al.* 2003) is used to update the archive. Accordingly, depending on the desired level of precision for quantifying each objective, appropriate ε-values are defined, and the set of ε-non-dominated solutions are stored in the archive.
3. The algorithm is continued until the pre-specified termination criteria are met (i.e. number of generations). Then, the percentage of change in the number of ε-non-dominated archive members in two consecutive NSGA-II runs is calculated. If the progress is less than a user defined value, δ, then a counter is incremented. Otherwise it is set to zero:

$$\text{if } \left(\frac{|N_i - N_{i-1}|}{N_{i-1}}\right) \times 100 < \delta \quad \text{then} \quad \text{counter} = \text{counter} + 1 \quad \text{else} \quad \text{counter} = 0 \tag{9}$$

where N_i is the number of archived solutions at run i. When the counter value reaches a user pre-defined threshold, w, indicating stalled improvement in w consecutive runs, the search is stopped. Otherwise, the search is continued by running a new NSGA-II with a doubled population size.

4. To speed up the convergence of the new NSGA-II run, the best results of the previous runs are included in the initial population by injecting solutions from the ε-domination archive formed at the end of the previous run. The diagram in Figure 4 schematically illustrates the structure of the proposed DDM framework.

CASE STUDY: HIGHWAY POLLUTANT BUILD-UP AND REMOVAL MODEL

Highways comprise a large fraction of the surface area in urban and non-urban regions. Due to their impermeability and the large amount of pollutants that is generated or deposited on their surfaces, they can be a significant source of pollution to surface and ground water. To evaluate the risk stormwater pollution can pose to water bodies, natural habitats and eventually to human health, it is important to know the amount, concentration and the temporal variation with which highway-associated pollutants are released from the highway surface into the storm runoff during a storm event. Efforts have been made to build predictive models to predict the event mean concentrations (EMCs), or the full pollutographs of various constituents in storm runoff, using empirical (e.g. Barrett *et al.* 1998; Kayhanian & Stenstorm 2005), or mechanistic models (e.g. Kim *et al.* 2005a, b; Massoudieh *et al.* 2005, 2008; Kang *et al.* 2006). Since empirical models are based on limited data gathered from sites in a particular geographic region, their applicability is limited to those sites (Barrett *et al.* 1998). In contrast, mechanistic models developed based on physical site characteristics and theoretical assumptions associated with pollutant build-up and wash-off can theoretically be applied to all paved surfaces and chemical constituents.

To illustrate the application of the proposed evolutionary model identification algorithm, it was used for modeling highway contaminant accumulation, fate, and transfer. The following sections address this case study in detail.

The general model structure

The sheet flow over the highway is assumed to occur mainly perpendicular to the highway axis and the highway surface is assumed to be uniform in terms of slope,

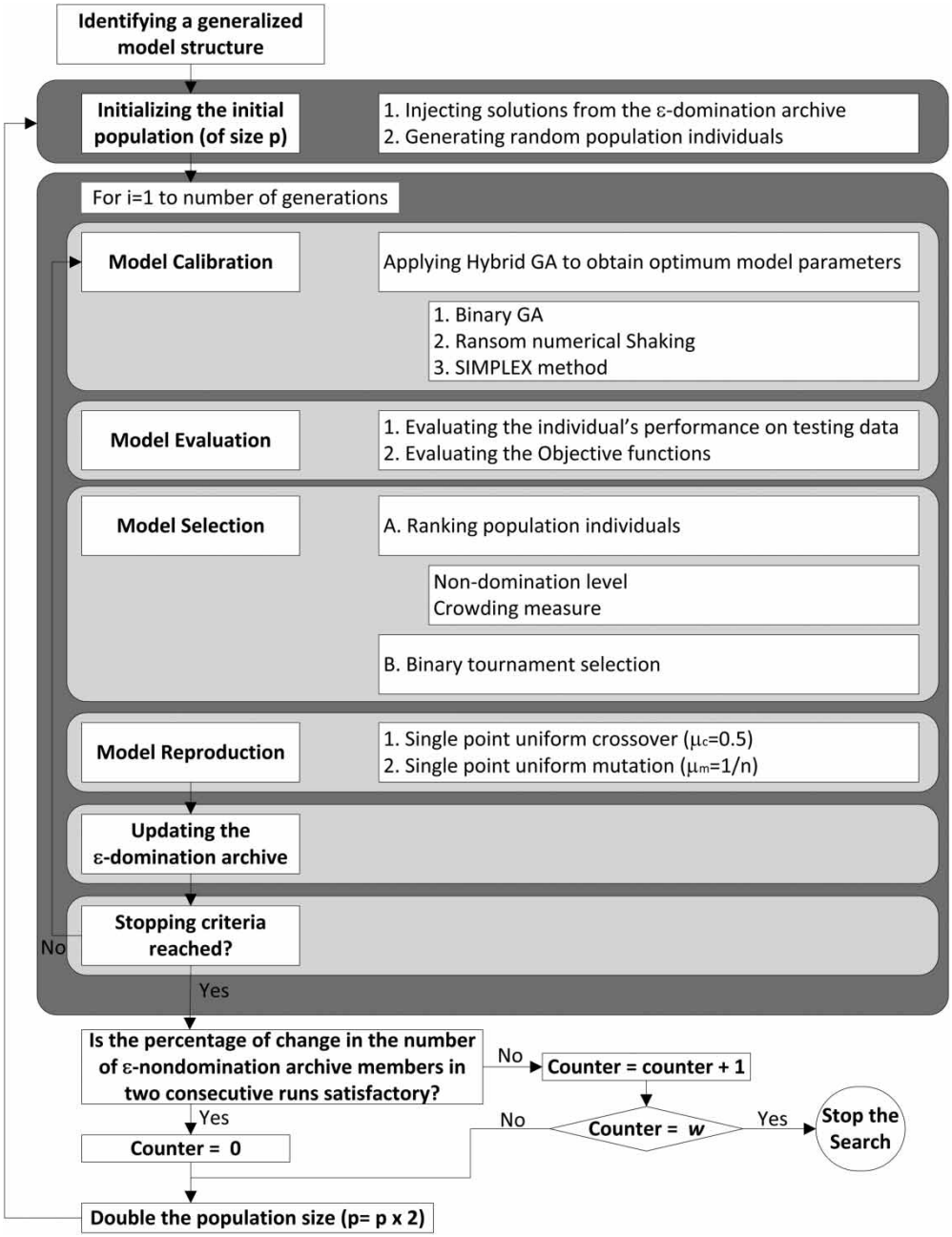

Figure 4 | Structure of the proposed DDM framework.

roughness and contaminant accumulation along the highway axis. With this assumption the problem can be represented as a one-dimensional system perpendicular to the highway axis. The contaminant model is composed of a flow and a transport component. The one-dimensional kinematic wave equation (Singh 1996) is used for modeling the runoff flow on the pavement:

$$\frac{\partial h}{\partial t} + \frac{\partial q}{\partial x} = i \tag{10}$$

where h [L] is the thickness of the sheet flow, and q [L^2/T] is the flow-rate per unit length of the highway surface that

is calculated using the Manning equation:

$$q = \frac{s^{1/2}}{n} h^{\alpha} \tag{11}$$

where n is the Manning's roughness coefficient, s is the slope of the pavement normal to the highway axis, and α is an exponent which was considered equal to 1.5 for thin sheet flow.

A rate-limited release and reattachment of contaminants is assumed between the pavement surface and the runoff, and the movement of contaminants in the surface runoff is considered to be controlled by advection, dispersion and attachment and detachment to the pavement surface. The particles and chemical constituents attached to the pavement surface were considered to be adsorbed to the pavement surface to at most two different sorption sites classes, including fast release and slow release sorption sites (Figure 5). Accordingly, the mass balance equation for transport in the runoff is:

$$\frac{\partial hC}{\partial t} + \frac{\partial S_f}{\partial t} + \frac{\partial S_s}{\partial t} + \frac{\partial qC}{\partial x} - \frac{\partial}{\partial x}\left(\partial h \frac{\partial C}{\partial x}\right) = 0 \tag{12}$$

where C is the concentration of mobile constituents such as suspended solids (TSS), dissolved chemicals, and the chemicals bound to mobile particles in the runoff, and S is the concentration of attached or adsorbed constituents to the pavement surface, expressed in mass per unit surface area of the pavement. The subscripts f and s denote the type of site, i.e. fast release and slow release sites, respectively. So, for example, S_f is the mass per unit area of constituents attached to fast sites on the pavement. Consequently, the mass balance for the attached contaminants is obtained by assuming rate controlled attachment/deposition on the surfaces and detachment/erosion from the surfaces, where the rate of attachment and detachment are considered a linear function of concentrations in each phase:

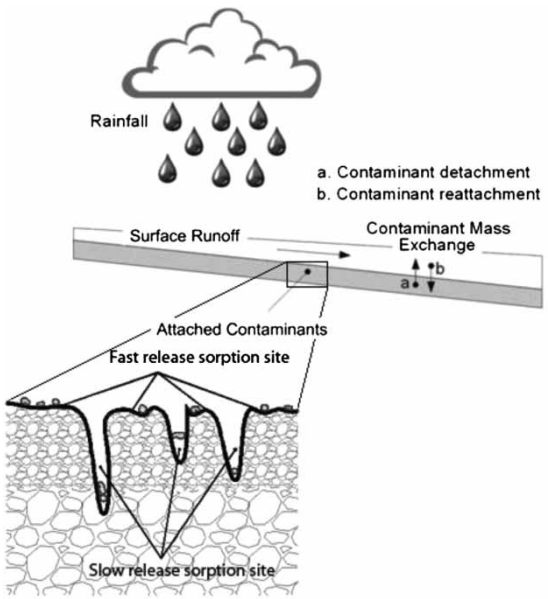

Figure 5 | Conceptual stormwater runoff and pollutant transport model for small drainage area.

$$\frac{\partial S_f}{\partial t} - k_{rf} C + k_{af} S_f = 0 \tag{13}$$

$$\frac{\partial S_s}{\partial t} - k_{rs} C + k_{as} S_s = 0 \tag{14}$$

where k_a is the attachment mass exchange rate coefficient and k_r is the release rate of particles from the pavement surface. The second subscripts denote the site types, (i.e. f for fast release sites and s for slow release).

The attachment and detachment of constituents are complex processes that are unknown functions of diffusive mass exchange between the phases, the rain drop impact, and the sheet flow shear stress (Massoudieh *et al.* 2005, 2008). Here, the main goal of the DDM algorithm is to identify the optimal form of functions representing the mass exchanges so that the model can best reproduce the measured mobile concentrations (C) at the downstream boundary of the basin. The attachment and release coefficients (k_a and k_r) for both sorption sites are deemed to be influenced by the rain intensity (i), flow shear stress and consequently flow velocity (u). Therefore, the following general functional structure is considered for these coefficients:

$$k = \beta_0 + \beta_1 i^{\mu_1} + \beta_2 u^{\mu_2} \tag{15}$$

where k can represent any of the four k coefficients in Equations (13) and (14) and β and μ are the coefficients and exponents, respectively. So, Equation (14) in fact represents four equations for k_{rf}, k_{af}, k_{rs}, and k_{as} that are encapsulated in one equation for the sake of brevity.

The accumulation of the attached contaminants on the surfaces during the dry period is controlled by the rates of pollutant deposition and removal which are mainly due to traffic and wind, and can be generally assumed as power functions of attached constituents mass per unit surface area. Therefore, the following equations are considered to govern the build-up and removal during the dry periods between storm events:

$$\frac{\partial S_f}{\partial t} - \omega_f + \varphi_f S_f^{\eta_f} = 0 \tag{16}$$

$$\frac{\partial S_s}{\partial t} - \omega_s + \varphi_s S_s^{\eta_s} = 0 \tag{17}$$

where ω, φ and η are unknown constants, coefficients and exponents, respectively. The accumulated constituents at the beginning of the season were considered as calibration parameters. Observed values of aqueous phase concentration at the end of the basin $C(x = L)$ have been measured at various times during the event. This quantity is used for model calibration, testing and model selection. To calculate C, Equations (10)–(14) are solved during the wet period and Equations (16) and (17) are solved during the dry periods between the events. A finite volume discretization method is used to solve coupled Equations (10)–(14). The model is run continuously throughout the wet season. It is worth noting that the right hand sides of Equations (12)–(14) during the rain event and Equations (16) and (17) during the dry season represent the vector function G in Equation (3) and the coefficients k_{rf}, k_{af}, k_{rs}, k_{as}, ω_f, ω_s, φ_f, φ_s, η_f, η_s comprise the K vector (Equation (4)) while the first four coefficients are considered to be functions of external factors rain intensity and shear stress and the last six are considered constant.

It should be noted that the forms of the attachment and release rate coefficients and contaminant accumulation rates presented in Equations (15)–(17) are the most general forms of the model's process functions. This general form contains 26 parameters to be estimated and, therefore, is over-parameterized considering the amount of data available on aqueous phase concentration at the downstream boundary of the basin during the events. The goal here is to allow the DDM algorithm to determine what specific subset of the unified model structure is the optimal model structure. So the DDM approach selects the terms of Equations (15)–(17) that are to be included in the final model and their forms in terms of being linear or power relationships. More complicated general forms such as multiplicative or other functions of the environmental variables could have been included in the general model, however the form was kept simple in order to keep the selected model practical and sensible.

Running the proposed evolutionary DDM framework

Hydrologic data including rain intensity, hydrographs and constituent's pollutographs were obtained from a first flush highway runoff characterization study of highly urbanized highway sites in west Los Angeles, California (Stenstrom & Kayhanian 2005). The data gathered from one site in a rain season was used as the 'calibration set' (training set) and the data for another site during the same year was used as the 'testing set'. Since the transport of a large number of contaminants is mainly due to adsorption to TSS, to demonstrate the modeling methodology, it was decided to conduct the modeling only for TSS. Therefore, the variable C in Equation (12) in this case study represents the TSS concentration in the runoff while S_f and S_s represent the mass per surface area of suspended solids trapped in fast and easy sites on the pavement surface.

The framework was applied to the two datasets, i.e. calibration data for model parameter estimation, and a testing data for evaluating the transferability of the model. All weighting coefficients, m, in the objective functions (Equations (7) and (8)) were set to 1 to reflect equal importance of models' performance on both datasets. Equations (12)–(14) were solved using an implicit finite volume method in order to regenerate pollutograph data and to evaluate the model structures. In the training phase, all model constants were allowed to be determined by the hybrid GA algorithm, while in the model testing phase, only the initial constituent concentration build-up, $S_f|_{t=0}$ and $S_s|_{t=0}$, were allowed to be determined by the hybrid GA search algorithm.

The termination criterion δ was set as 10% and the threshold w was set equal to 5. The ε precision vector was

chosen to be (0.2, 0.1). For the hybrid GA used in parameter estimation, the following parameters were used: population = 40, crossover probability = 100%, mutation rate = 0.008, number of step is the SIMPLEX enhancement 100, maximum standard deviation for random shaking = 0.05.

MODELING RESULTS

Figure 6 shows the obtained values of the objective functions for the best performing models and the ultimate Pareto front. In this figure, each point represents a model which, compared to its neighbor models, performs better according to at least one objective function. Table 1 shows the mathematical structure and objective function values of the obtained models. To further assess the validity and usefulness of the models, a third set of unseen data (validation dataset) was used. The r^2_{XY} for each of the models and datasets is also presented in Table 1.

Among the final models on the Pareto front, model 1 is the best model in terms of transferability by showing no decline in model fit when applied to unseen data. On the other hand, the small r^2_{XY} for all datasets indicate that this model has the lowest performance in predicting the observed data which can paint the model unreliable for prediction purposes. Conversely, at the other end of the table, model 6 presents the best match collectively with respect to the calibration and test data, while relative to the calibration data presents the largest relative decline of matching between the model and measured data when applied to the test data (i.e. the lowest transferability). Models 2–5 are sorted based on the values of the first objective function from low to high. So as we move towards the end of the table, the match between measured data and modeled results improves while the decline of the model performance when applied to unseen data decreases. It should be noted that r^2_{XY} is only an indicator of how good the model captures the trends in the data and not the difference between the magnitudes of modeled and observed data. So a decline in r^2_{XY} for models with a high $f_2(X)$ is compromised by a larger COD for the testing data. Figures 7–9 show the modeled vs. observed pollutographs in selected rain events by models 4 and 6, for the calibration, testing and validation datasets, respectively. Figure 10 illustrates the predicted vs. measured TSS concentrations at the discharge outlet for the testing, calibration and validation datasets for both models.

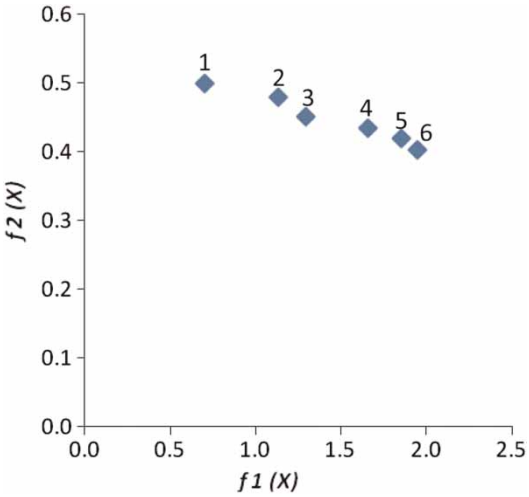

Figure 6 | Ultimate Pareto front of the solution archive ($f_1(X)$ and $f_2(X)$ are the first and second objective functions, respectively).

Model 4 has a more complex structure (seven parameters) and results in similar r^2_{XY} values for calibration, testing and validation datasets, which is an indication of its good transferability in terms of capturing the pollutographs' trends. A closer look at Figure 7 shows that pollutographs generated by this model have several spikes throughout the rain event. This is due to the sensitivity of this model to the rain intensity and the fact that rain data used here (as in most rain data available) is in the form of accumulation in given time intervals, and therefore contains abrupt changes.

Acquiring the highest value for the first objective function ($f_1(X) = 1.945$), model 6 has the best overall performance for calibration and testing datasets. This model also has a relatively high r^2_{XY} for validation data, which is counter-intuitive since we expect this model to present the lowest transferability. This unexpected result can be attributed to the accidental similarity between the datasets (or site characteristics/rain patterns) used for calibration and validation. The simple model structure of model 6 indicates that representing the pavement surface using only a single sorption site type is sufficient. Furthermore, the obtained model structure implies that the attachment of TSS particles during the storm (k_a) is not significantly important in the overall process and the detachment of suspended

Table 1 | Mathematical forms of the obtained models

	Model structure		$f_1(X)$	$f_2(X)$	r^2_{XY} calibration	r^2_{XY} testing	r^2_{XY} validation
1	$k_{rs} = 1.728i^{0.700}$	$\frac{\partial S_s}{\partial t} = 3.351$	0.699	0.500	0.170	0.157	0.199
	$k_{as} = 0.001$	$k_{rf} = 415.139i^{1.222}$					
2	$\frac{\partial S_f}{\partial t} = 0.4847$	$\frac{\partial S_s}{\partial t} = 0.172$	1.133	0.480	0.379	0.284	0.019
	$k_{rs} = 0.290u^{2.279}$	$k_{rf} = 1.750u^{1.156}$					
		$k_{af} = 0.071$					
3	$\frac{\partial S_f}{\partial t} = 0.009$	$\frac{\partial S_s}{\partial t} = 2.302$	1.297	0. 451	0.415	0.276	0.288
	$k_{rs} = 227.094i^{1.295}$	$k_{rf} = 0.112u^{0.538}$					
	$k_{as} = 2.323u^{3.000}$	$k_{af} = 103.337i^{1.983}$					
4	$k_{rs} = 4.415u^{2.689}$	$\frac{\partial S_s}{\partial t} = 0.641$	1.656	0.435	0.463	0.411	0.467
		$k_{rf} = 26.298i^{1.140}$					
		$k_{af} = 0.192u^{2.333}$					
5	$\frac{\partial S_f}{\partial t} = 0.288$	$\frac{\partial S_s}{\partial t} = 0.039$	1.850	0.419	0.493	0.238	0.642
	$k_{rs} = 0.003i^{0.444}$	$k_{rf} = 1.328u^{2.029}$					
	$k_{as} = 0.002u^{1.363}$	$k_{af} = 987.936i^{1.489}$					
6	$\frac{\partial S_f}{\partial t} = 0.085 - 0.071S_f^{0.670}$		1.945	0.403	0.519	0.301	0.534
	$k_{rf} = 494.8i^{1.076}$						

particles can be assumed to be irreversible during the storm (Equation (12)). Figure 8 shows that the predicted pollutographs follow the same trend for most storm events, and for some events, the prediction is fairly close to measured values. Although showing the highest overall accuracy compared to other models, the low r^2_{XY} value for testing data (0.301) and the plot of modeled vs. observed TSS concentrations (Figure 10(d)) indicate significant bias in the model's predictions for testing data.

It should be noted that the resulting optimal model structures are only valid for the amount of data available for this research. A larger dataset might result in a different optimal model structures. Furthermore, the reason for relatively low r^2_{XY} values for calibration and testing data can be attributed to the complexity of the system and, more specifically, the high spatial and temporal heterogeneity in accumulation of pollutants on the highway surface, the dependence of pollution accumulation to factors not included in the model such as wind, traffic volume and their temporal variations as well as the inherent simplifications in the overall model structure such as the one-dimensional assumption, and neglecting the non-uniformity of the particle sizes.

The goodness of fit values reported in Table 1 may seem relatively low. In order to evaluate the method, we compared the results of the models discovered by the data-driven method with two versions of the wash-off model proposed by Kang *et al.* (2006) using the same datasets. These models are one-dimensional deterministic models that use the kinematic wave equation to calculate flow, and mass transport and erosion equations to calculate pollutant concentrations.

In the first version of their model, Sartor & Boyd's (1972) assumption was implemented, assuming that the road is a long-term source capable of generating various products of decomposition and aggregate materials. In the second version, it was assumed that pollutant loadings increase with time, and based on initial calibration results, equations were derived for the buildup of constituents. Hereafter, the first and second versions of the model are referred to as 'Kang *et al.* (2006) (S-B)' and 'Kang *et al.* (2006)',

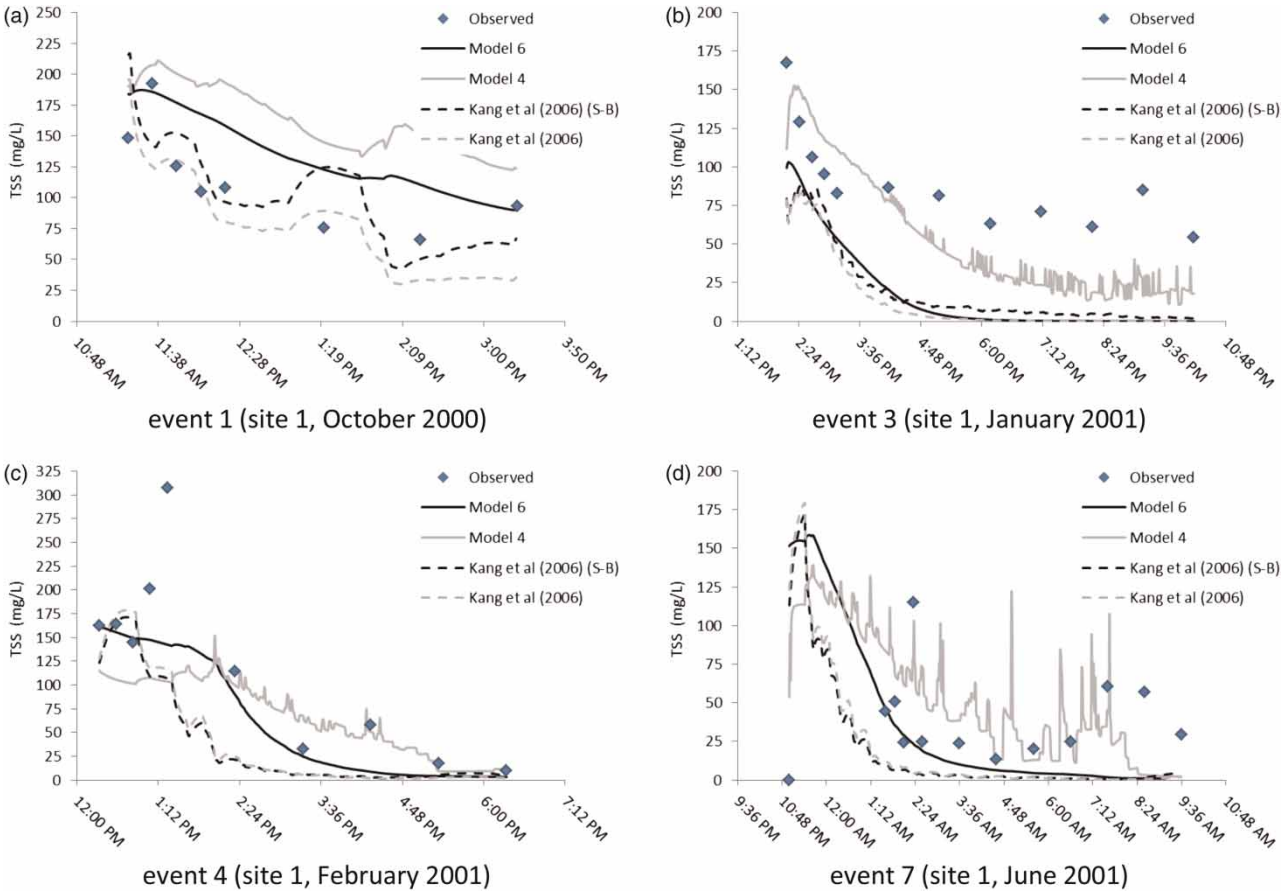

Figure 7 | Modeled vs. observed TSS pollutographs for four selected rain events of the calibration dataset.

respectively. In the Kang *et al.* (2006) (S-B) model, the rate of attachment and detachment are governed by the following equations:

$$\begin{cases} \dfrac{\partial S_f}{\partial t} + \beta_1 u^2 S_f = 0 \\ \dfrac{\partial S_s}{\partial t} + \beta_2 u^2 S_s = 0 \end{cases} \tag{18}$$

In the Kang *et al.* (2006) model an assumption is made that the change in the source of the pollutants attached to the slow sites as a result of erosion can be ignored:

$$\begin{cases} \dfrac{\partial S_f}{\partial t} + \beta_1 u^2 S_f = 0 \\ E = \beta_2 u^2 : (E \text{ replaces } \partial S_s/\partial t \text{ in Eq. 12}) \end{cases} \tag{19}$$

Equations (18) or (19) are solved along with the transport equation (Equation (12)) for the concentration of pollutants in the runoff. In the Kang *et al.* (2006) model, the term $\partial S_s/\partial t$ is replaced by E which represents the release rate of pollutants from the long-term (slow) sites. For in-depth description of these models' structures, the reader is referred to Kang *et al.* (2006).

The pollutographs predicted by the two versions of Kang *et al.*'s model for the three datasets and selected rain events are plotted in Figures 7–9. Table 2 also compares the values of the goodness of fit criteria of these models with models 4 and 6 obtained using the DDM framework. Models 4 and 6 slightly outperform both versions of Kang's models in terms of reproducing the calibration data. This is an expected outcome since the complexity of both versions of Kang's models are almost similar with that of models 4 and 6. On the other hand, in terms of transferability (i.e. accuracy of prediction of the testing and validation data) models 4 and 6 outperform both versions of Kang's model significantly as it is expected.

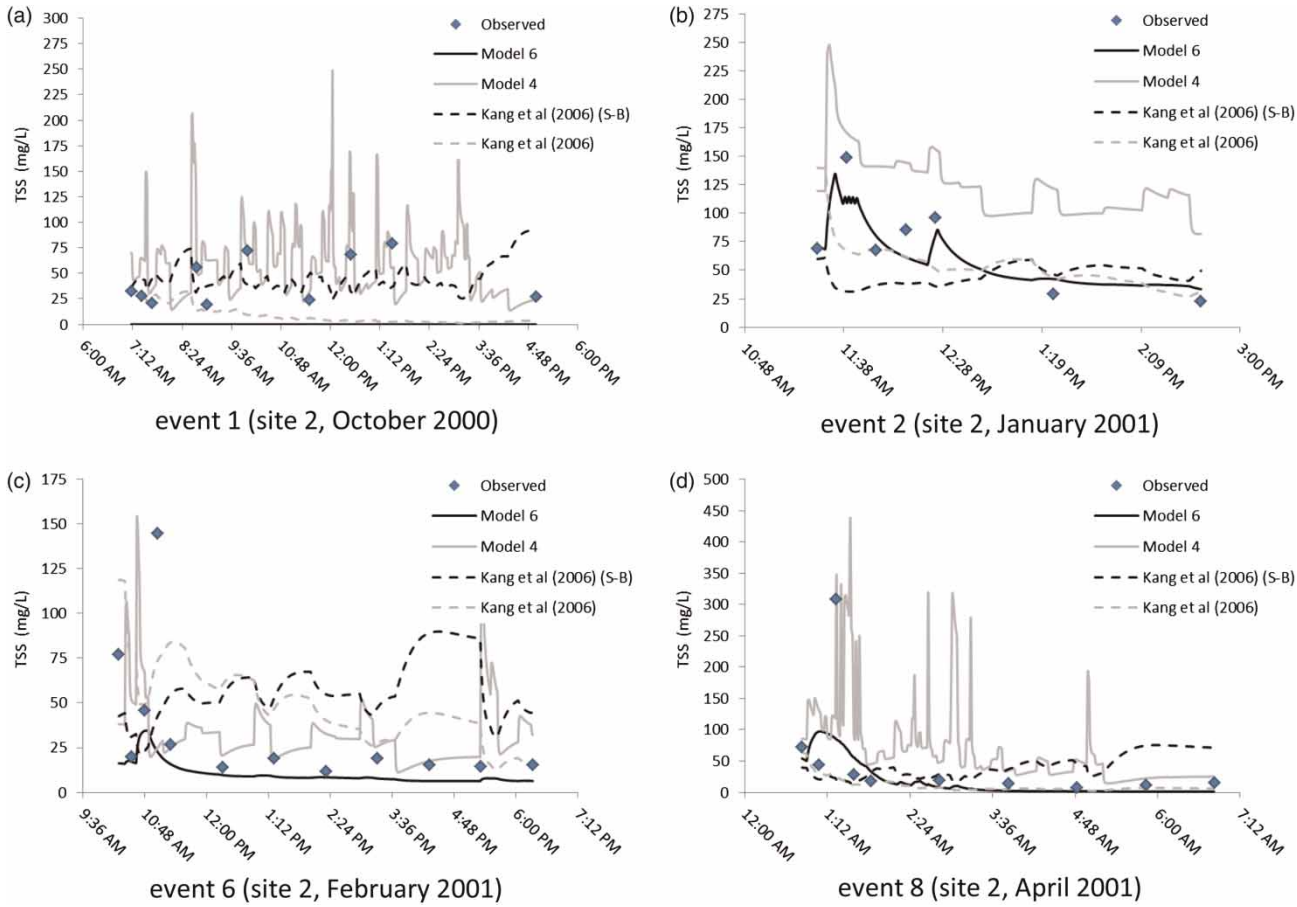

Figure 8 | Modeled vs. observed TSS pollutographs for four selected rain events of the testing dataset.

CONCLUSIONS

This paper presents a novel framework for DDM of environmental phenomena. This framework has been designed so as to be able to incorporate physically based or constitutive assumptions of the system into the modeling process. The framework uses a MOGA called NSGA-II to find optimum model structures and a hybrid GA to calibrate the models. The key features of the proposed framework are as follows:

1. This framework allows the modeler to construct the general structure of the model based on the physical processes controlling the system while allowing the data driven algorithm to discover the relationships controlling some components of the model. This feature extends the abilities and flexibility of the method, and makes it more suitable for application in modeling complex environmental systems. Furthermore, linking the model with the physical processes deemed to govern a system, this modeling framework can be used as an effective tool for understanding the underlying physical processes.
2. Employing a MO approach, the suggested framework allows the modeler to automatically incorporate different aspects of goodness of a model in the model selection process. These aspects include, but are not necessarily limited to, the capability of the model to reproduce the observed data after parameter estimation, and the transferability of the model, i.e. the capability of the model to reproduce datasets other than the data used for parameter estimation, using the parameters obtained in the parameter estimation step. This forces the algorithm to avoid overfitting the data and search for more parsimonious models.

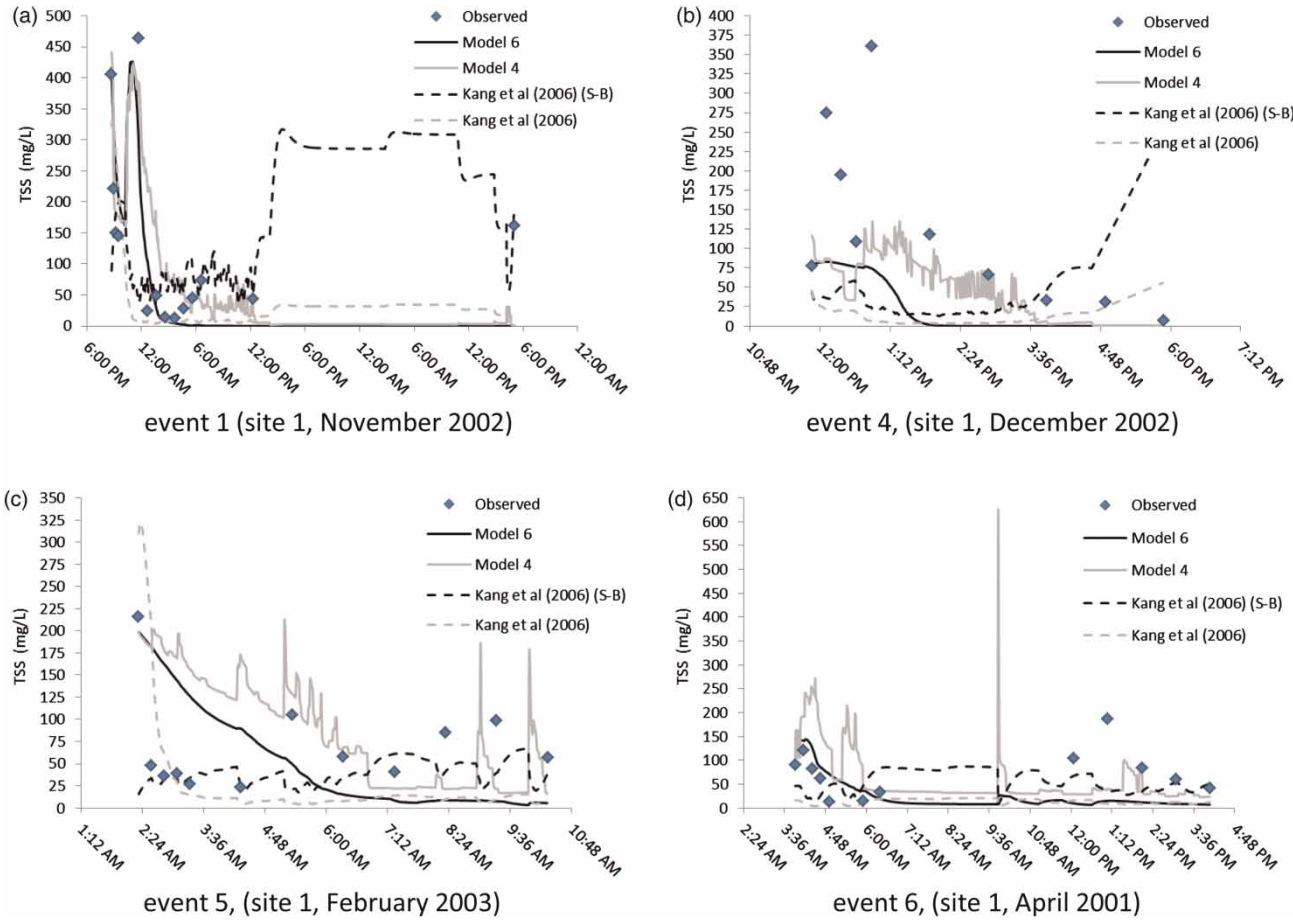

Figure 9 | Modeled vs. observed TSS pollutographs for four selected rain events of the validation dataset.

3. The proposed method is able to find a set of Pareto optimum solutions which all have acceptable performances. This set of equally 'fit' models enables the modeler to make an intelligent choice regarding the model structure based on the goal of the modeling practice, e.g. (a) mere ability to predict the system, (b) simple model structure, (c) model with physical meaning (scientific knowledge discovery) and so on.

In order to test the applicability of the proposed framework to environmental modeling, the method was applied to observed data from highway sites in California to discover mathematical structures of TSS wash-off, build-up and transport from highway surfaces. Six different models were obtained, each being superior to others in at least one aspect of model goodness. The models' objective function values and the r^2_{XY} for three datasets (i.e. calibration, testing and validation) along with the predicted pollutographs and modeled vs. observed plots, were investigated to make a final evaluation of the models' performance. It was concluded that the model that had the best performance in terms of transferability (model 1) lacked sufficient precision in predicting TSS pollutographs. Furthermore, the simplest model (model 6) was found to be the overall most precise model but showed significant bias in modeling the testing data. Another model (model 4) was found to show consistent accuracy in modeling all three datasets.

The same datasets were used to calibrate two versions of the model proposed by Kang *et al.* (2006). Comparison of the predicted pollutographs (Figures 7–9) and the calculated goodness of fit values (Table 2) indicate that the models obtained by the DDM framework perform better than Kang's models both in terms of the capability to reproduce the calibration data and when applied to the unseen testing and validation datasets.

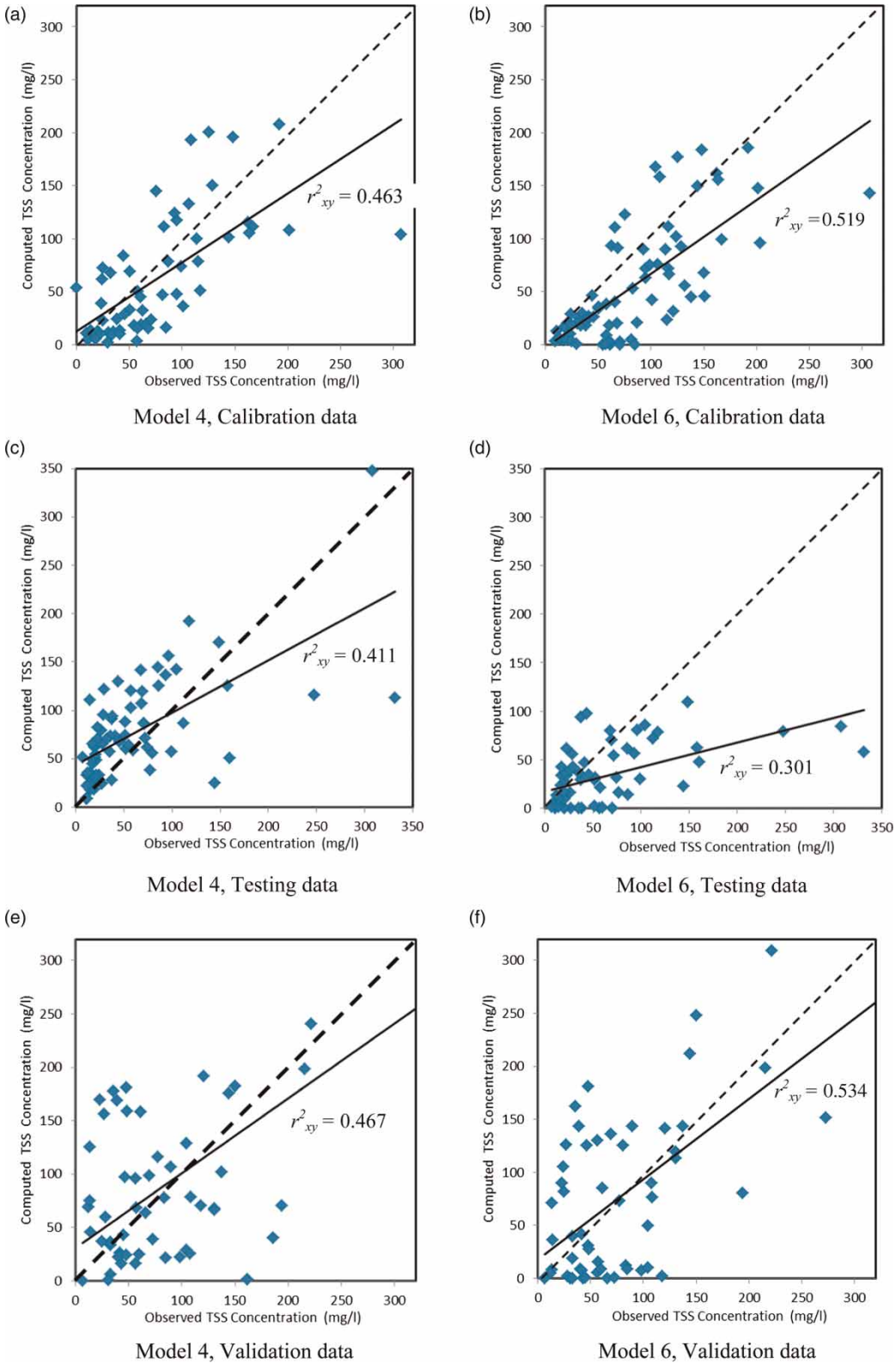

Figure 10 | Modeled vs. measured TSS concentrations for calibration, testing and validation datasets using models 4 and 6.

Table 2 | Goodness of fit values of models 4, 6, Kang *et al*. (2006) (S-B) and Kang *et al*. (2006)

	Calibration		Testing		Validation	
	r^2_{XY}	COD	r^2_{XY}	COD	r^2_{XY}	COD
Model 4	0.4630	0.4731	0.4110	0.3094	0.4670	0.5211
Model 6	0.5190	0.6419	0.3010	0.4833	0.5340	0.4502
Kang *et al*. (2006) (S-B)	0.3917	0.7203	0.1599	0.2586	0.0056	0.2304
Kang *et al*. (2006)	0.3591	0.6783	0.0269	0.3669	0.2916	0.4845

Our results confirm the potential for the use of the proposed evolutionary DDM framework in environmental modeling. However, more case studies are required to confirm its efficiency and ease of use for a variety of problems. We believe the described methodology fills a need in the environmental research community for a tool for modeling complex environmental systems.

ACKNOWLEDGEMENTS

Partial funding for this study was provided by the District of Columbia Water Resources Research Institute (DC-WRRI). Also, the authors gratefully acknowledge Dr Masoud Kayhainian, Professor Michael Stenstrom, and graduate students and research staff from UC Davis and UCLA Departments of Civil and Environmental Engineering for sharing the highway pollutograph data used in this study.

REFERENCES

Babayan, A. V., Savic, D. A. & Walters, G. A. 2005 Multiobjective optimization for the least-cost design of water distribution systems under correlated uncertain parameters. In: *Proceedings of the 2005 World Water and Environmental Resources Congress* (R. Walton, ed.). Reston, United States, pp. 36.

Back, T., Hammel, U. & Schwefel, H. 1997 Evolutionary computation: comments on the history and current state. *IEEE Trans. Evol. Comp.* **1** (1), 3–17.

Back, T. & Schwefel, H. P. 1993 An overview of evolutionary algorithms for parameter optimization. *Evol. Comp.* **1** (1), 1–23.

Bardossy, A., Bogardi, I. & Duckstein, L. 1990 Fuzzt regression in hydrology. *Water Resour. Res* **26** (7), 1497–1508.

Barrett, M. E., Irish, L. B., Malina, J. F. & Charbeneau, R. J. 1998 Characterization of highway runoff in Austin, Texas, area. *J. Environ. Eng.-ASCE* **124** (2), 131–137.

Bekele, E. G. & Nicklow, J. W. 2007 Multi-objective automatic calibration of SWAT using NSGA-II. *J. Hydrol* **341** (3–4), 165–176.

Chipman, J. 2006 Pareto and contemporary economic theory. *Int. Rev. Econ.* **53** (4), 451–475.

Coley, D. A. 1999 *An Introduction to Genetic Algorithms for Scientists and Engineers*. World Scientific Publishing, Singapore.

Davidson, J. W., Savic, D. & Walters, G. A. 1999 Method for the identification of explicit polynomial formulae for the friction in turbulent pipe flow. *J. Hydroinform.* **1** (2), 115–126.

Davidson, J. W., Savic, D. A. & Walters, G. A. 2000 Approximators for the Colebrook-White formula obtained through a hybrid regression method. In: *Computational Methods in Water Resources Vol 2 Computational Methods, SurfaceWater Systems and Hydrology* (L. R. Bentley, J. F. Sykes, C. A. Brebbia, W. G. Gray & G. F. Pinder, eds). Balkema, Rotterdam, pp. 983–989.

Deb, K., Agrawal, S., Pratap, A. & Meyarivan, T. 2000 A fast elitist non-dominated sorting genetic slgorithm for multi-objective optimization: NSGA-II. In: *Proceedings of the Parallel Problem Solving from Nature VI Conference* (M. Schoenauer, K. Deb, G. U. Rudolph, X. Yao, E. Lutton, J. J. Merelo & H.-P. Schwefel, eds). Paris, France, pp. 849–858.

Deb, K., Mohan, M. & Mishra, S. 2003 *A Fast Multi-Objective Evolutionary Algorithm for Finding Well-spread Pareto-optimal Solutions*. KanGAL Report No. 2003002. Indian Institute of Technology, Kanpur, India.

Deb, K., Pratap, A., Agarwal, S. & Meyarivan, T. 2002 A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Trans. Evol. Comp.* **6** (2), 182–197.

De Jong, K. A. 2006 *Evolutionary Computation: A Unified Approach*. MIT Press, Cambridge, MA.

El-Baroudy, I., Elshorbagy, A., Carey, S. K., Giustolisi, O. & Savic, D. 2010 Comparison of three data-driven techniques in modelling the evapotranspiration process. *J. Hydroinform.* **12** (4), 365–379.

Evsukoff, A. G., Fairbairn, E. M. R., Faria, E. F., Silvoso, M. M. & Filho, R. D. T. 2006 Modeling adiabatic temperature rise during concrete hydration: a data mining approach. *Comput. Struct.* **84** (31–32), 2351–2362.

Fan, S. K. S., Liang, Y. C. & Zahara, E. 2006 A genetic algorithm and a particle swarm optimizer hybridized with Nelder-Mead simplex search. *Comput. Ind. Eng.* **50**, 401–425.

Fonseca, C. M. & Fleming, P. J. 1995 An overview of evolutionary algorithms in multiobjective optimization. *Evol. Comput.* **3**, 1–16.

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

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

Giustolisi, O. & Savic, D. A. 2009 Advances in data-driven analyses and modeling using EPR-MOGA. *J. Hydroinform.* **11**, 225–236.

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

Hirschen, K. & Schafer, M. 2006 A study on evolutionary multi-objective optimization for flow geometry design. *Comput. Mech.* **37** (2), 131–141.

Holland, J. K. 1975 *Adaptation in Natural and Artificial Systems*. MIT Press, Cambridge, MA.

Hsu, K., Gupta, H. & Sorooshian, S. 1995 Artificial neural network modeling of the rainfall-runoff process. *Water Resour. Res.* **31** (10), 2517–2530.

Kang, J. H., Kayhanian, M. & Stenstrom, M. K. 2006 Implications of a kinematic wave model for first flush treatment design. *Water Res.* **40** (20), 3820–3830.

Kayhanian, M. & Stenstrom, M. 2005 Mass loading of first flush pollutants with treatment strategy simulations. *Transport. Res. Rec.: J. Transport. Res. Board* **1904** (1), 133–143.

Khare, V., Yao, X. & Deb, K. 2003 Performance scaling of multi-objective evolutionary algorithms. In: *Proceedings of the Second International Conference on Evolutionary Multi-Criterion Optimization, EMO 2003. (Lecture Notes in Computer Science Vol.2632)* (C. M. Fonseca, P. J. Fleming, E. Zitzler & K. Deb, eds). Berlin, Germany, pp. 376–390.

Kim, L. H., Kayhanian, M., Lau, S. L. & Stenstrom, M. K. 2005a A new modeling approach for estimating first flush metal mass loading. *Water Sci. Technol* **51** (3–4), 159–167.

Kim, L. H., Kayhanian, M., Zoh, K. D. & Stenstrom, M. K. 2005b Modeling of highway stormwater runoff. *Sci. Total Environ* **348** (1–3), 1–18.

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

Laumanns, M., Thiele, L., Deb, K. & Zitzler, E. 2002 Combining convergence and diversity in evolutionary multiobjective optimization. *Evol. Comput.* **10** (3), 263–282.

Lee, C. C. & Doong, S. H. 2008 Evolutionary selection of model for time constrained decision problems: a GA approach. *Expert Syst. Appl.* **34**, 1886–1892.

Lemke, F., Benfenati, E. & Müller, J. 2006 Data-driven modeling and prediction of acute toxicity of pesticide residues. *ACM SIGKDD Explor. News* **8** (1), 71–79.

Massoudieh, A., Abrishamchi, A. & Kayhanian, M. 2008 Mathematical modeling of first flush in highway storm runoff using genetic algorithm. *Sci. Total Environ.* **398** (1–3), 107–121.

Massoudieh, A., Huang, X. J., Young, T. M. & Marino, M. A. 2005 Modeling fate and transport of roadside-applied herbicides. *J. Environ. Eng.-ASCE* **131** (7), 1057–1067.

Michaelwicz, Z. 1992 *Genetic Algorithms+Data Structures=Evolution programs*, 3rd edition. Springer-Verlag, New York.

Nazemi, A., Xin, Y. & Chan, A. H. 2006 Extracting a set of robust Pareto-optimal parameters for hydrologic models using NSGA-II and SCEM. In *Proceedings of 2006 IEEE Congress on Evolutionary Computation (CEC'06)*. Vancouver, Canada, pp. 6792–6799.

Nelder, J. & Mead, R. 1965 A simplex method for function minimization. *Comput. J.* **7** (4), 308.

Osyczka, A. 2002 *Evolutionary Algorithms for Single and Multicriteria Design Optimization*. Physica-Verlag, Heidelberg, New York.

Poli, R., Langdon, W. B. & McPhee, N. F. 2008 *A field guide to genetic programming*. Online. Available from http://lulu.com and freely available from http://www.gp-field-guide.org.uk (accessed 12 April 2012).

Reed, P., Kollat, J. B. & Devireddy, V. K. 2007 Using interactive archives in evolutionary multiobjective optimization: a case study for long-term groundwater monitoring design. *Environ. Model. Softw.* **22** (5), 683–692.

Reed, P., Minsker, B. S. & Goldberg, D. E. 2003 Simplifying multiobjective optimization: an automated design methodology for the nondominated sorted genetic algorithm-II. *Water Resour. Res.* **39** (7), 1196.

Sartor, J. D. & Boyd, G. B. 1972 *Water Pollution, Aspects of Street Surface Contaminants*. EPA-R2-72-081, USEPA, Washington, DC.

Sharifi, S. 2009 *Application of Evolutionary Computation to Open Channel Flow Modelling*. PhD Thesis, University of Birmingham, Birmingham, UK.

Sharifi, S., Sterling, M. & Knight, D. W. 2011 Prediction of end-depth ratio in open channels using genetic programming. *J. Hydroinform.* **13** (1), 36–48.

Singh, V. 1996 *Kinematic Wave Modeling in Water Resources: Surface-water Hydrology*. Wiley-Interscience, New York.

Solomatine, D. P. & Ostfeld, A. 2008 Data-driven modelling: some past experiences and new approaches. *J. Hydroinform.* **10** (1), 3–22.

Sousa, S., Martins, F., Alvim-Ferraz, M. & Pereira, M. 2007 Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations. *Environ. Model. Softw.* **22** (1), 97–103.

Stenstrom, M. K. & Kayhanian, M. 2005 *First Flush Phenomenon Characterization*. California Department of Transportation Division of Environmental Analysis, Sacramento, CA.

First received 4 March 2011; accepted in revised form 14 September 2011. Available online 6 March 2012