AGU PUBLICATIONS

Water Resources Research

RESEARCH ARTICLE
10.1002/2015WR017971

Key Points:
- A procedure based on the regression technique ERP-MOGA is proposed for relevant input selection
- Relevant input selection is made based on variable occurrence statistics and engineering judgment
- Input selection improves data fitting in storm water quality modeling

Correspondence to:
E. Creaco,
creaco@unipv.it

Citation:
Creaco, E., L. Berardi, S. Sun, O. Giustolisi, and D. Savic (2016), Selection of relevant input variables in storm water quality modeling by multiobjective evolutionary polynomial regression paradigm, *Water Resour. Res.*, *52*, 2403–2419, doi:10.1002/2015WR017971.

Received 9 AUG 2015
Accepted 26 FEB 2016
Accepted article online 4 MAR 2016
Published online 1 APR 2016

Selection of relevant input variables in storm water quality modeling by multiobjective evolutionary polynomial regression paradigm

E. Creaco[1,2], L. Berardi[3], Siao Sun[4], O. Giustolisi[3], and D. Savic[5]

[1]Department of Civil Engineering and Architecture, University of Pavia, Pavia, Italy, [2]Formerly at College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, UK, [3]Department of Civil Engineering and Architecture, Technical University of Bari, Bari, Italy, [4]Key Laboratory of Regional Sustainable Development Modeling, Institute of Geographical Sciences and Natural Resource Research, Chinese Academy of Sciences, Beijing, People's Republic of China, [5]College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, UK

Abstract The growing availability of field data, from information and communication technologies (ICTs) in "smart" urban infrastructures, allows data modeling to understand complex phenomena and to support management decisions. Among the analyzed phenomena, those related to storm water quality modeling have recently been gaining interest in the scientific literature. Nonetheless, the large amount of available data poses the problem of selecting relevant variables to describe a phenomenon and enable robust data modeling. This paper presents a procedure for the selection of relevant input variables using the multiobjective evolutionary polynomial regression (EPR-MOGA) paradigm. The procedure is based on scrutinizing the explanatory variables that appear inside the set of EPR-MOGA symbolic model expressions of increasing complexity and goodness of fit to target output. The strategy also enables the selection to be validated by engineering judgement. In such context, the multiple case study extension of EPR-MOGA, called MCS-EPR-MOGA, is adopted. The application of the proposed procedure to modeling storm water quality parameters in two French catchments shows that it was able to significantly reduce the number of explanatory variables for successive analyses. Finally, the EPR-MOGA models obtained after the input selection are compared with those obtained by using the same technique without benefitting from input selection and with those obtained in previous works where other data-modeling techniques were used on the same data. The comparison highlights the effectiveness of both EPR-MOGA and the input selection procedure.

1. Introduction

In the last decades, water utilities have significantly increased the number and frequency of measured and collected data. This increase was mainly spurred by attempts to keep system processes under close control and to better understand their behavior. This fits nicely within the "smart city" paradigm where new information and communication technologies (ICTs) are used to achieve better, more livable cities. In addition, the development and application of ICT solutions for urban infrastructures increasingly facilitate the integration of heterogeneous data that might be collected for different purposes.

The increased availability of measured data has enabled development and training of data modeling, which can help practitioners and researchers in understanding complex phenomena and supporting management decisions. This is of particular relevance to civil engineering practice, where many complex and large-scale environmental phenomena cannot only be analyzed by means of physically based modeling. In fact, the extension of physically based models built at a laboratory scale to the real scale environment would usually require overly high costs associated with system monitoring. At the same time, practitioners and researchers need relationships between measurable variables in order to understand the problem and to support decisions.

In the field of storm sewer systems, traditional methods for collecting water quantity and quality data rely on sampling campaigns and laboratory analyses, which are typically costly and labor consuming. Therefore, measured field data are usually limited [e.g., *Bannerman et al.*, 1993]. Since the late 1990s, online continuous measuring programs have started to be used in drainage water monitoring [e.g., *Barraud et al.*, 2002]. As a

result, more and more data sets characterizing storm water are available due to the advancement of measuring technologies [e.g., *Lacour et al*., 2009; *Metadier and Bertrand-Krajewski*, 2012]. These available data sets can then be used for building data-driven models to help understand the complex phenomena or for prediction purposes. Some examples of storm water quality and sewer deterioration data-driven models are provided by *Vaze and Chiew* [2003], *Mourad et al*. [2005], *Aryal et al*. [2009], *Dotto et al*. [2011], *Sun and Bertrand-Krajewski* [2011, 2012, 2013], *Berardi et al*. [2006], *Savic et al*. [2006], *Berardi and Kapelan* [2007], and *Berardi et al*. [2008]. In particular, data-driven models are often preferred to physically based accumulation-erosion-transfer approaches when modeling storm water quality due to their conceptual simplicity and goodness-of-fit performance [*Métadier*, 2011; *Sun and Bertrand-Krajewski*, 2011]. Unlike physically based models, which have limited success in reproducing pollutographs in complex sewers, data-driven models are normally used to predict total loads (Sun and Bertrand-Krajewski). However, in a number of situations, the total load of pollutants may suffice to quantify the quality response of receiving waters [*Donigian and Huber*, 1991; *Dembélé et al*., 2010].

In the framework of data-driven models, identifying the most relevant variables to describe a given phenomenon is of interest from both the data collection and effective modeling viewpoints. On the one hand, understanding which information should be collected first can help in prioritizing what should be measured under budget constraints. On the other hand, the preliminary selection of inputs is essential for system identification through data modeling. As highlighted by *Galelli et al*. [2014], irrelevant input variables are uninformative about the underlying process and their eventual inclusion would cause noise and complexity to be added to the model. The inclusion of redundant, but relevant, input variables would instead increase the dimensionality of the model identification without providing any additional predictive benefit. One of the possible undesirable consequences of including irrelevant and redundant input variables is the construction of models that overfit training data, while showing poor generalization capabilities in other similar contexts. Such a drawback is undesirable in engineering fields where the transferability of identified data-driven models from one case to another (new) is required.

In the context of sewer systems, *Sun and Bertrand-Krajewski* [2012] investigated the extent to which the selection of the training data set affects the results of the training of data modeling for modeling storm water quality parameters COD (chemical oxygen demand) and TSS (total suspended solids). This was extended further by the same authors [*Sun and Bertrand-Krajewski*, 2013], who used a cross-validation method to improve input variable selection, i.e., the search for the most suitable number of explanatory variables to ensure that a model is neither overfitted nor underfitted. Though being one of the first attempts to consider input selection in this area, *Sun and Bertrand-Krajewski* [2013] used only simple linear regression models. More research is then needed in order to set up procedures aimed at input selection in nonlinear models, which can be more refined approximations of the complex phenomena in the urban and environmental systems [*Sun and Bertrand-Krajewski*, 2011].

The issue of input variable selection in nonlinear models of storm water quality is addressed in this paper. In particular, a procedure for relevant input selection was developed, based on the use of the multiobjective evolutionary polynomial regression (EPR-MOGA) and of the multi case strategy (MCS-EPR-MOGA). EPR-MOGA is a nonlinear regression technique, developed by *Giustolisi and Savic* [2009] as an extension of the original EPR paradigm proposed by *Giustolisi and Savic* [2006]. It has been used in many different contexts in the field of Hydroinformatics [e.g., *Berardi et al*., 2006; *Savic et al*., 2006; *Berardi et al*., 2008; *Laucelli and Giustolisi*, 2011]. Furthermore, *Berardi and Kapelan* [2007] and *Savic et al*. [2009] proposed and demonstrated a generalization of the EPR-MOGA paradigm, called MCS-EPR-MOGA. This technique is aimed at searching for model structures to represent the dependence of target variables on explanatory variables by simultaneously simulating various data sets (i.e., cases).

Applications of EPR-MOGA and MCS-EPR-MOGA to many contexts, including sewer systems [*Berardi et al*., 2006; *Berardi and Kapelan*, 2007], have proven their capability to yield robust models for the representation of real complex phenomena. However, the input selection performance of EPR-MOGA in the case of many candidate explanatory variables has never been tested. Furthermore, no systematic procedure for input selection though EPR-MOGA has ever been presented in the scientific literature. The novelty of this paper thus lies in the development of a systematic procedure for the identification of the most relevant inputs from a large number of candidates based on EPR-MOGA. In fact, this identification is as an essential

step prior to the construction of a data-driven model, when many potential explanatory variables are present.

Compared to other input selection procedures [e.g., *Bowden et al*., 2005a, 2005b; *D'Heygere et al*., 2006; *Giustolisi and Simeone*, 2006; *Yang and Ong*, 2011; *Wan Jaafar et al*., 2011; *Tirelli and Pessani*, 2011] presented in the scientific literature, the new procedure presents the advantage of being more simple. As highlighted by *Galelli et al*. [2014], input selection algorithms normally "involve three main steps: (1) generating a subset of inputs from the candidate input pool, (2) evaluating the subset of inputs in terms of their ability to predict the output, and (3) assessing whether the selected set inputs is optimal using a prespecified stopping criterion." As will be shown hereinafter, the new procedure, instead, is simply based on two steps: (1) application of EPR-MOGA or MCS-EPR-MOGA and (2) analysis of input occurrences in EPR-MOGA or MCS-EPR-MOGA models and selection of the most relevant inputs.

The remainder of the paper is organized as follows. First, the methodology is described. Then, the applications of the methodology to the modeling of storm water quality in two French sewer systems show how the input selection procedure works with a wide set of available data. Finally, EPR-MOGA and MCS-EPR-MOGA are applied to construct storm water quality models based on the selected input variables in the two sewer systems. The constructed models are then compared with those developed on the same data in a previous work [*Sun and Bertrand-Krajewski*, 2011].

2. Methodology

In this section, a brief description of the EPR-MOGA paradigm and its multicase strategy variant MCS-EPR-MOGA is first reported. The procedure developed for relevant input selection is then elaborated.

2.1. EPR-MOGA Paradigm for Selecting Input Variables

EPR is a hybrid data-modeling paradigm which combines the advantages of the heuristic search for optimal mathematical model expressions with numerical linear regression for parameter estimation. The EPR modeling scheme with its multiobjective enhancement and the more recent Multi-Case Strategy variant have been presented and discussed in a number of referenced works [e.g., *Giustolisi and Savic*, 2006; *Giustolisi and Savic*, 2009; *Savic et al*., 2009]. Thus, only the key features that are relevant to its applicability for input selection are discussed here. For the sake of clarity, these features are presented in separated subsections, although all of them are integrated in the same modeling paradigm.

2.1.1. Symbolic Expressions

The EPR paradigm operates on data to produce symbolic expressions, i.e., formulas that the user can easily interpret based on his/her expertise in the particular field. On the one hand, this enables scrutiny of the relationships found (e.g., in terms of direct/inverse linear/nonlinear dependence between inputs and output). On the other hand, symbolic expressions should contain combinations of variables that might have physical meaning for the process being modeled. In this case, EPR not only identifies significant variables but also is likely to facilitate knowledge discovery about the system [*Giustolisi and Savic*, 2009].

Equation (1) shows a general EPR model expression, where $\mathbf{X}_i$ are input vectors and $\mathbf{ES}(j,i)$ is the exponent of the ith input within the jth term of the pseudopolynomial:

$$\mathbf{Y} = a_0 + \sum_{j=1}^{m} a_j \times (\mathbf{X_1})^{\mathbf{ES}(j,1)} \cdots (\mathbf{X}_k)^{\mathbf{ES}(j,k)} \times f\left((\mathbf{X_1})^{\mathbf{ES}(j,k+1)} \cdots (\mathbf{X}_k)^{\mathbf{ES}(j,2k)}\right), \quad (1)$$

where a_j is the jth model parameter and function f is a user-defined function that has five possible settings: (1) no function, (2) natural logarithm, (3) exponential, (4) tangent hyperbolic, and (5) secant hyperbolic; m and k represent the maximum number of polynomial terms and the number of input variables (details can be found in *Giustolisi and Savic* [2006]).

The user starts by selecting a set of possible exponents, including the null value. During the search for the model expressions, those inputs that are assigned $\mathbf{ES}(j,i) = 0$ disappear from the expression. Such a strategy allows the user to include all available inputs as "candidate" explanatory variables, while the modeling strategy selects those that provide the best fit model.

For finding the most relevant inputs among many candidates, the simplest form, as given in equation (2) and obtained by considering the "no function" setting for f in equation (1), is generally suggested:

$$\mathbf{Y} = a_0 + \sum_{j=1}^{m} a_j \times (\mathbf{X_1})^{\mathbf{ES}(j,1)} \cdots (\mathbf{X}_k)^{\mathbf{ES}(j,k)}. \quad (2)$$

With regard to the adoption of the simple form in equation (2), it has to be noted that in this case EPR is simply used as a data analyzer, aimed at detecting the contribution of each input variable, and not for the construction of data-driven models.

2.1.2. Numerical Parameters (Coefficients)

Models in equations (1) and (2) are nonlinear structures. However, they are linear with respect to numerical parameters (a_i), which can be estimated by Least Squares (LS) or Nonnegative LS, or any linear parameter estimation procedure. From the system identification point of view, robust models are those that provide good fit to data by moving from one system to another while changing the numerical parameters only. This is particularly relevant to engineering applications where a relationship between measurable variables is needed, while numerical parameters are likely to represent specific system characteristics.

2.1.3. Multiobjective Modeling Paradigm

The multiobjective search paradigm implemented within EPR-MOGA shows a twofold advantage for selecting the relevant input variables by: (1) driving the search by selecting models with the objective functions to be optimized and (2) returning a set of optimal models, which provide a basis for selecting the most relevant input variables. Both features benefit from the user's expertise on the particular field.

The EPR model search can be driven toward those models that maximize the goodness of fit to data (i.e., in terms of Coefficient of Determination—CoD) while simultaneously minimizing the number N_x of inputs ($\mathbf{X}_i$) in the resulting model, the number N_a of polynomial terms (up to a maximum number m), or both. In the first case (goodness of fit versus number of inputs N_x), the number of monomial terms in resulting expressions is not limited (up to m terms decided by the user). In this case, the search tends to distribute candidate variables over different terms. In the second case (goodness of fit versus number of terms N_a), one additional term is justified only if it enables improved goodness of fit; this pushes the search toward more compact expressions where inputs are aggregated. In the third case (goodness of fit versus number of inputs N_x versus number of terms N_a), the addition of one input variable or one term in the final expressions depends on which model structure provides the best fit to the target output. Although establishing a rule about the preferred strategy is not easy, the last strategy (three-objectives) returns a larger number of models in comparison with the others, and thus is usually preferred. The result of the application of EPR-MOGA is then made up of a Pareto front (band or surface depending on two or three selected objectives) of models, featuring increasing levels of complexity and goodness-of-fit performance.

It is worth remarking that, due to the global exploration in the space of symbolic expressions, the persistence of some variables (or their combinations) in different models provides the user with increased confidence about their relevance in describing the phenomenon (i.e., target output).

2.2. MCS-EPR-MOGA Paradigm for Selecting Input Variables

In some real contexts, like that reported in this paper, data are collected over different systems (or come from different experiments on the same system). In this circumstance, there are two options to exploit the EPR paradigm for input selection: (1) testing the inputs selected for one system on the other and (2) searching for the relevant variables to describe the same phenomenon over all data sets simultaneously.

Option 1 requires performing the EPR-MOGA analysis for each data set and comparing the selected variables from different data sets. Option 2 can be easily performed using the MCS-EPR-MOGA strategy. The strategy basically searches for the common model structure (i.e., combination of exponents in equation (1) or (2)) but estimates different parameters (i.e., a_j in equation (1)) for each case (data set).

Therefore, MCS-EPR-MOGA can also be used in the global analysis of the relevant inputs of a physical phenomenon by using data coming from different data sets. The parameters (differing from case to case) catch the difference between one physical system and another, for a given physical phenomenon, or the changing condition over time of the same physical system.

It is worth noting that in MCS-EPR-MOGA [*Berardi and Kapelan*, 2007; *Savic et al*., 2009], the approach used in this paper assigns the same weight to each event in different data sets during the optimization. As a consequence, a larger data set (containing more events) has a larger impact on the calculation of the CoD describing the goodness of fit of the model to data. This is reasonable, as a larger data set is statistically

more significant than a shorter one in providing information to describe the system/phenomenon in hand [*Ljung*, 1999].

2.3. Selection of the Relevant Input Variables

A novel procedure that can be used for input selection in the framework of EPR paradigm is hereinafter described. In particular, a description is first provided for the procedure applied to a single case study. The extension to the case with a number of data sets featuring the same output and potential explanatory variables then follows.

It is worth noting that the reduction in the number of explanatory variables produced by the application of the procedure to any kind of problem reduces the size of the search space. Then, it improves the efficiency and effectiveness of any subsequent application of a data-driven modeling technique, such as EPR-MOGA, MCS-EPR-MOGA, artificial neural networks (ANNs) [*Haykin*, 1999], or genetic programming [*Koza*, 1992].

2.3.1. Single Data Sets

The procedure encompasses the following two steps:

1. Perform an EPR-MOGA run, in which a reduced set of candidate exponents is considered; the reduced set of candidate exponents **ES**(*j*, 1. . .*k*) aims to enable identification of the most relevant input variables, without necessarily yielding data-driven models with high fit to data. This reduced set of exponents has to be chosen in order to include positive values representing sublinear, linear, and superlinear relationships, and the negative values. The set must include the 0 value, which helps to consider the independence of the explained variable from a certain explanatory variable.
2. Explore the optimal EPR-MOGA models and analyze the occurrences of each potential explanatory variable, i.e., the number of times that it appears in the set of EPR-MOGA models obtained at the end of step 1, in order to identify relevant inputs, as explained below.

   For the selection of relevant inputs, the following statistical criteria can be adopted:

   a. Select the *Nsel* variables with highest number of occurrences, with *Nsel* being a parameter to be fixed a priori; this approach can be useful, for example, when only a limited budget is available to start a data collection campaign;
   b. Fix a threshold value and then consider, as relevant input variables, those that feature a number of occurrences over this threshold value.

The application of either of the statistical criteria has to be accompanied with engineering judgement. Such an approach, which is possible using the EPR paradigm, cannot be easily extended to other data-modeling techniques (e.g., artificial neural networks). In fact, EPR-MOGA has the advantage to produce symbolic mathematical expressions that can be easily read and interpreted by experts through physical insight.

2.3.2. Multiple Data Sets (Cases)

In this case, two possible options can be followed according to the application of EPR-MOGA or MCS-EPR-MOGA, as mentioned in section 2.2:

1. Option 1 consists in applying the EPR-MOGA-based procedure described in the previous subsection to each of the available case studies (i.e., separate data sets). For each potential explanatory variable, the total number of occurrences in the model expressions is considered as the sum of the numbers of occurrences in various case studies. This option requires separate EPR runs for different case studies;
2. Option 2 consists in applying MCS-EPR-MOGA to all case studies in a single run in order to derive the number of occurrences of each explanatory variable.

Once input variable occurrences have been evaluated, step 2 described in the previous subsection can be applied, based on occurrence analysis and expert engineering judgement.

3. Applications

3.1. Case Studies

Two urban catchments, located in the West and East part of Lyon (France), respectively, namely the Ecully and Chassieu catchments [*Métadier*, 2011; *Metadier and Bertrand-Krajewski*, 2012; *Sun and Bertrand-Krajewski*, 2012], were considered. These catchments were two experimental sites in the OTHU project (Field Observatory for Urban Hydrology—www.othu.org). In particular, the Ecully catchment is a low-density residential catchment, with a surface area of 245 ha and an imperviousness coefficient of 42%. The Chassieu catchment

Table 1. Variables Measured in the Ecully and Chassieu Catchments[a]

Variables	Variable Symbol	Unit	Ecully			Chassieu		
			Range	Mean Value	Standard Deviation	Range	Mean Value	Standard Deviation
TSS event load	TSS	kg	2–4294	557	617	5–6,404	371	685
COD event load	COD	kg	2–7,573	976	1098	7–5,823	352	619
Rainfall duration	TRD	h	0.09–49.4	5.8	6.8	0.1–51.2	5.1	6.2
Total rainfall depth	TRH	mm	0.2–70.0	6.3	10.0	0.2–65.4	5.6	9.2
Rainfall intensity	RI	mm/h	0.1–29.1	2.0	3.6	0.1–27.6	1.5	2.5
Maximum rainfall intensity	MRI	mm/h	0.3–172	12.9	22.3	0.2–137	11.3	17.0
Maximum rainfall intensity in 5 min	MRI5	mm/h	0.3–36.8	4.4	5.7	0.2–40.3	3.9	4.8
Number of the day in a year when an event occurs (1–365)	NDAY		2.0–354	166	102	1.4–366	183	111
Number of the hour in a day when an event occurs (1–24)	NHOUR		0–23	12.4	7.0	0–23	11.1	6.8
Antecedent dry period	ADP	day	0.2–23.6	2.7	3.6	0.1–21.9	2.5	3.6
Antecedent dry period from the last rainfall event exceeding *J* mm (*J* = [5:5:40])	ADP*J*	day		11.2–84.1	13.3–83.7		10.1–116.2	10.8–80.6
Cumulative rainfall during *M* hours before the rainfall event (*M* = 4, 8, 12, 24, 36, 48, 52, 72, 96)	CR*Y*	mm		0.02–9.6	0.07–13.3		0.02–9.7	0.2–12.9
Maximum rain intensity in *Y*-minutes during *Z* hours before the event (*Y* = 5,10, 30; *Z* = 4, 8, 12, 24, 36, 48, 52, 72, 96)	MRI(*Y*,*Z*)	mm/h		2.2–10.7	3.2–20.2		1.9–11.5	3.4–17.9
Average flow rate	AFR	L/s	21.8–1,018	118	112	4.0–880	102.1	121.7
Maximum flow rate	MFR	L/s	28.5–3,860	417	589	9.4–4,574	472	630
Duration of the runoff flow	DRF	h	0.03–2.38	0.3	0.3	0.1–2.54	0.32	0.27
Total flow volume	TFV	m^3	120–42,198	3,313	5,727	84–42,506	2,993	5,386

[a]Adapted from *Sun and Bertrand-Krajewski* [2012].

is an industrial catchment, with a surface area of 185 ha and an imperviousness coefficient of 75%. In the Ecully and Chassieu catchments, numerous variables (see Table 1) were measured during 239 and 263 rain events, respectively, in the period from 2004 to 2008.

Two typical storm water quality indicators, i.e., the total amount (kg) of chemical oxygen demand (COD) and total suspended solids (TSS) recorded during a certain rain event, were chosen as target variables (model outputs). In addition, peak concentrations of COD and TSS, which are representative of interevent characteristics, were also tentatively used as model outputs. However, the performance of models simulating peak pollutant concentrations was generally poor (see Appendix A). These variables, which are assessed on short time intervals, are more prone to random factors and measurement uncertainty, and thus are less related to event rainfall and runoff variables. Therefore, further analysis on modeling peak pollutant concentrations was not performed.

The potential explanatory variables (model inputs) included 56 variables associated with rainfall and runoff, such as total amount, mean rate, duration, etc. These variables, listed in Table 1, were obtained starting from rainfall and runoff time series recorded in the Ecully and Chassieu catchments during the OTHU project. Among the potential explanatory variables listed in Table 1, there are three groups. The first group is ADP*J* (antecedent dry period from the last rainfall event exceeding *J* mm). The second group is CR*M* (cumulative rainfall during *M* hours before a rainfall event, i.e., in other eventual rain events preceding the event considered inside an *M* hour-long time slot). The third group is MRI(*Y*,*Z*) (maximum rainfall intensity over a *Y* minute time interval during *Z* hours before an event, i.e., in other eventual rain events preceding the event considered inside a *Z* hour-long time slot). In the ADP*Y* group, eight variables are present corresponding to *Y* values ranging from 5 to 40. In the CR*Y* group, nine variables are present corresponding to *Y* values ranging from 4 to 96. Finally, 27 variables are present in the MRI(*Y*,*Z*) group corresponding to *Y* and *Z* values ranging from 5 to 30 and from 4 to 96, respectively.

In the applications, the first two third of the events were used for selecting the most relevant variables. The three-objective approach (see sections 2.1 and 2.2) was used in both EPR-MOGA and MCS-EPR-MOGA. All modeling runs were performed using the EPR-MOGA-XL tool (www.hydroinformatics.it).

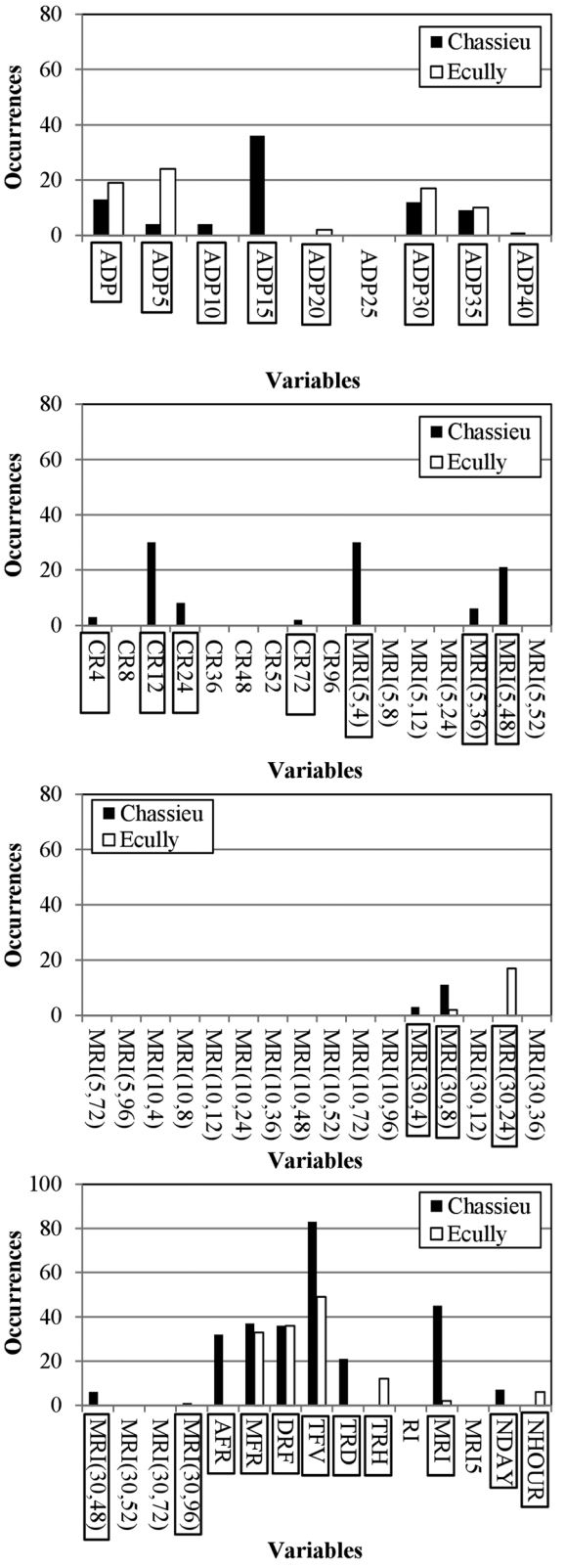

Figure 1. Modeling COD in the two catchments. Occurrences of the potential explanatory variables listed in Table 1 inside EPR-MOGA models. Selected variables highlighted with a rectangle.

For the selection of relevant input variables for either TSS or COD, both options reported in section 2.3 were used, i.e., option 1 entailing the comparison between two separate EPR-MOGA modeling runs (the cases in Ecully and Chassieu) and option 2 using MCS-EPR-MOGA.

For both EPR-MOGA and MCS-EPR-MOGA runs, the maximum number of monomials in the EPR structure in equation (2) was set to 5 and the set of exponents for candidate variables (**ES**) was set to [−2, −1, −0.5, 0, 0.5, 1, 2] following a brief sensitivity analysis. It is worth noting that the choice of the exponents aims to represent all the possible kinds of relationship between output and input variables, i.e., the aim was not to search for the function structure for fitting the data. Apart from the 0 value, the values of 0.5, 1, and 2 account for sublinear (exponent within range of 0–1), linear (exponent equal to 1), and superlinear (exponent larger than 1) dependences, respectively. The negative values consider inverse dependencies. The selection of a larger number of candidate exponents inside EPR-MOGA does not conceptually improve the selection of attributes/inputs, while exponentially increases the search space of the MOGA optimization, which is undesirable. Both EPR-MOGA and MCS-EPR-MOGA were applied using the Nonnegative LS regression strategy to estimate coefficients in equation (2). As shown by *Giustolisi and Savic* [2009], this numerical regression strategy is a way to hinder the introduction of alternate monomials, which are likely to reproduce noises in data and overfit the target variables; ultimately it helps to increase the generalization of the models (i.e., selection of relevant variables).

EPR-MOGA and MCS-EPR-MOGA runs were performed with population and generation parameters equal to 10 and 40, respectively, in the EPR-MOGA-XL software.

For selection of the relevant inputs based on the number of occurrences, criterion b in section 2.3 was adopted considering a threshold value equal to 0. This means that only variables that never appeared in the EPR-MOGA and MCS-EPR-MOGA models were excluded from the list of the relevant input variables. Two lists of relevant

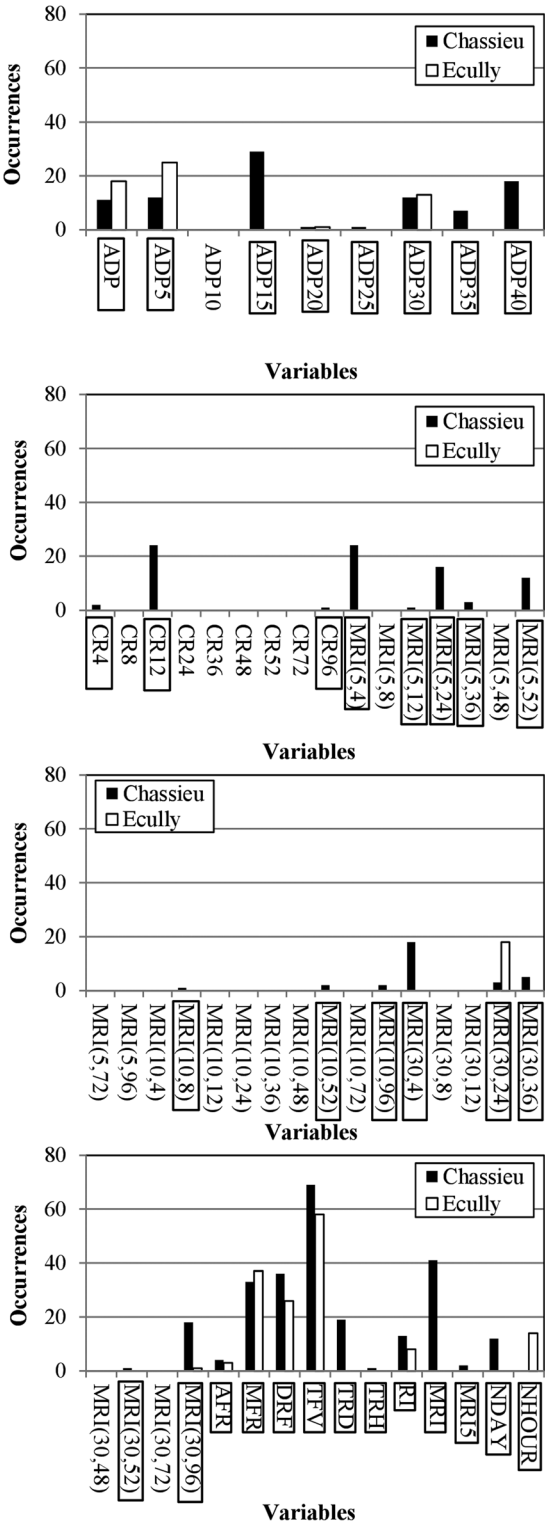

Figure 2. Modeling TSS in the two catchments. Occurrences of the potential explanatory variables listed in Table 1 inside EPR-MOGA models. Selected variables highlighted with a rectangle.

explanatory variables (i.e., one for EPR-MOGA models and the other for MCS-EPR-MOGA) were then created for TSS and COD, respectively.

The results of the application of the procedure for input selection are reported in section 3.2.

Subsequent analyses based on the selected explanatory variables for the development and test of data-driven models are then reported in section 3.3.

3.2. Analysis of Input Selection Results

The application of EPR-MOGA to the modeling of COD and TSS in the Chassieu catchment yielded 49 and 42 models, respectively. The application to the modeling of COD and TSS in the Ecully catchment yielded 28 and 29 models. The simultaneous application of MCS-EPR-MOGA to the two catchments led to 53 and 58 models for the modeling of COD and TSS, respectively.

The occurrences of the various input variables in the EPR models were calculated and reported in Figures 1–3. In particular, Figures 1 and 2 show the results of option 1 based on EPR-MOGA, for the modeling of COD and TSS, respectively. Figure 3 shows the results of option 2 based on MCS-EPR-MOGA.

Figures 1 and 2 indicate that 29 and 35 explanatory (out of the original 56 candidates) were included by EPR-MOGA in the list of relevant variables for modeling COD and TSS, respectively. Figure 3 shows that MCS-EPR-MOGA selects 30 and 31 variables as relevant input variables for modeling COD and TSS, respectively. The results indicate that selected variables are very similar whether option 1 or option 2 is used. The advantage of using MCS-EPR-MOGA is that it requires only one run to analyze all the cases simultaneously.

However, the application of the input selection procedure helped in reducing significantly the list of explanatory variables. Most of the eliminated variables belong to the three groups, i.e., (ADP*Y*, CR*Y*, and MRI(*Y*,*Z*)). An explanation for this is that the variables in each group are strongly interrelated; therefore, the suitable selection of one or some of them is sufficient to fully describe the explanatory effects of the whole group, making the others redundant. As an example, let us consider the results obtained through EPR-MOGA and, in particular, the group MRI(*Y*,*Z*) of candidate explanatory variables for the representation of COD (Figure 1). Among the 27 variables, only 8

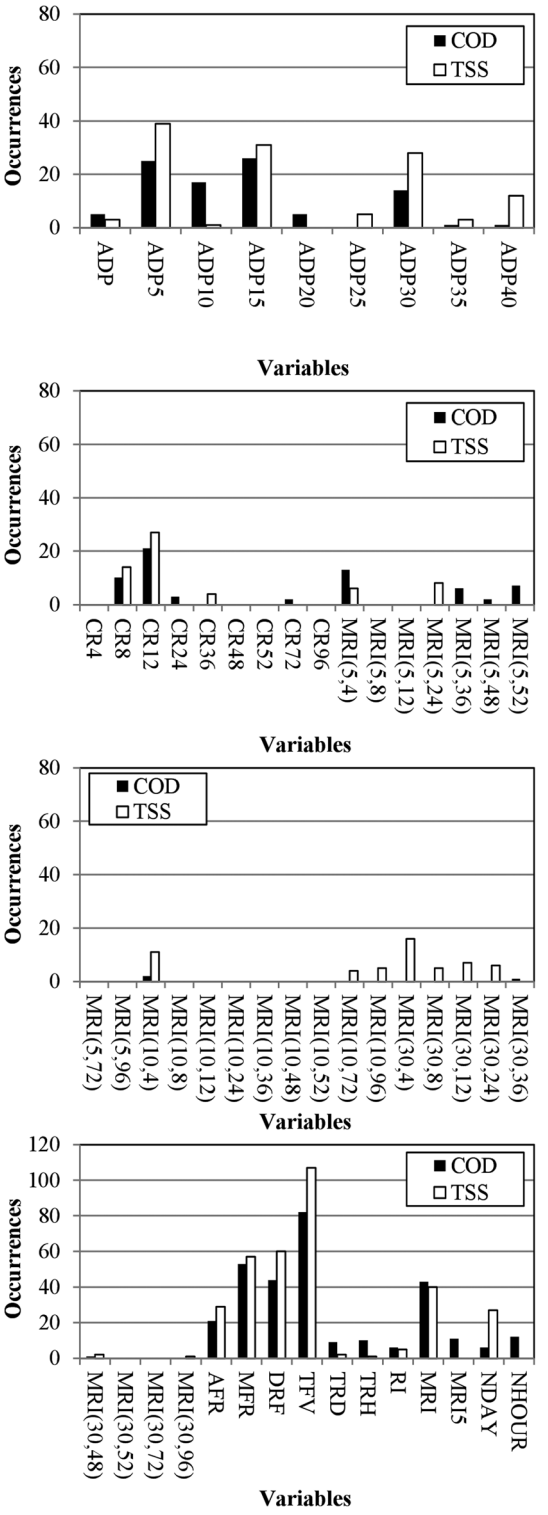

Figure 3. Output COD and TSS. Occurrences of the potential explanatory variables listed in Table 1 inside the MCS-EPR-MOGA models. Selected variables are those featuring a number of occurrences larger than 0.

are simultaneously present in the models for both catchments. Consequently, they are considered relevant inputs.

Figures 1–3 provide an insight into the most relevant explanatory variables (in terms of number of occurrences in the models) for the transport of pollutants in sewers. If, for instance, the five most relevant explanatory variables are required for modeling COD, both EPR-MOGA and MCS-EPR-MOGA select TFV, DRF, MRI, MFR, and ADP15. At first sight, it seems to be unexpected that, in the context of the EPR-MOGA applications, some variables have many occurrences in one catchment and are almost absent in the other. For instance, MRI and ADP15 have more occurrences in Chassieu than in Ecully. This may be due to the different characteristics of the catchments. However, in the Ecully catchment, ADP30 and ADP35, which are closely related to ADP15, have more occurrences than in the Chassieu catchment and may thus compensate for the fewer occurrences of ADP15.

The results of EPR-MOGA and MCS-EPR-MOGA also agree when the five most relevant explanatory variables for TSS are considered, which are TFV, DRF, MRI, MFR, and ADP5. At this stage, it has to be noted that four out of five most relevant explanatory variables, i.e., TFV, DRF, MRI, and MFR, are the same for modeling COD and TSS. The other variables, ADP15 (for COD) and ADP5 (for TSS), are interrelated since they both characterize the antecedent dry period. The selection of TFV, DRF, MRI, MFR, and ADP, as the most relevant explanatory variables for pollutant quantities in sewers, is in agreement with the recent findings in the scientific literature [*Ashley et al*., 2004]. A relationship is expected to exist between the COD and TSS masses and all these variables.

3.3. Modeling Storm Water Quality Indicators by EPR-MOGA

The selected inputs were then used to construct models for simulating storm water quality parameters COD and TSS in the Ecully and Chassieu catchments. These applications were performed by applying EPR-MOGA with the following settings. In particular, the first two third of the events, previously used for input selection, were used hereinafter to develop and train data-driven models; the second one third of the events were used for testing the data-driven models. The upper boundary of N_a was set to 5. Compared to input selection, a wider set of exponents, with values in the range between −3 and 3, spanned with a 0.5 long step, was used in order to improve the fitting performance of the data-driven models,

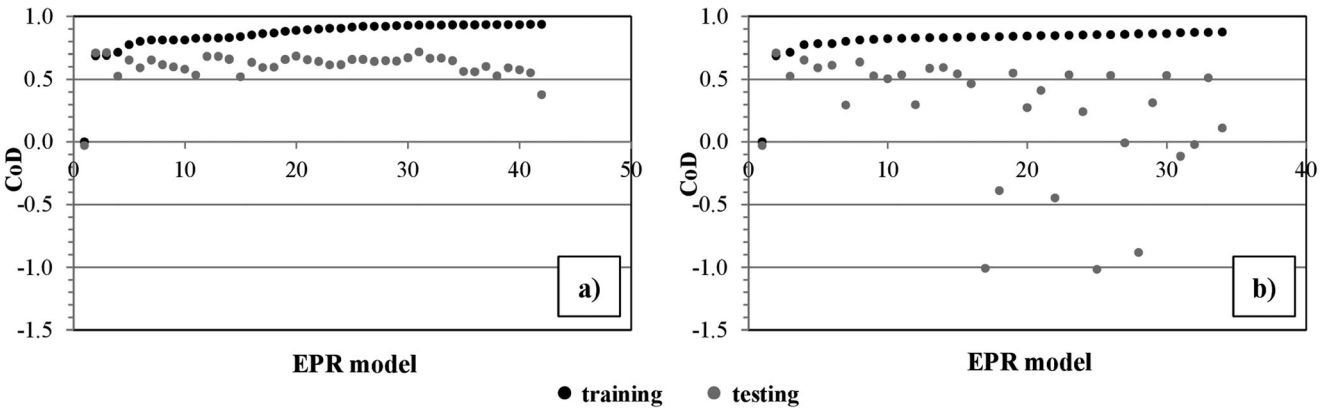

Figure 4. Values of coefficient of determination CoD obtained by the EPR models in the Ecully catchment in the training (black dots) and testing (grey dots) phases for the representation of COD (kg); results of (a) application of EPR following input selection procedure and (b) application of EPR alone. In the graphs, EPR models are sorted according to ascending values of CoD.

provided that the most significant model inputs/attributes have been identified in the previous step of work. A shorter step could have been used to span the range of exponents to improve the goodness of fit in the training phase. However, results of simulations not reported in the paper showed that this leads to two problems for the EPR-MOGA models:

1. Decrease in the generalization performance of the model, i.e., decrease in the model prediction ability for test data;
2. Decrease in the understanding ability of EPR-MOGA model structures, which is a relevant purpose of a paradigm searching for symbolic structures of data.

In particular, EPR-MOGA was applied to each catchment for modeling COD and TSS with the selected explanatory variables of Figures 1 and 2, respectively. In these runs, the population and generation parameters were also set to 10 and 40, respectively, in the EPR-MOGA-XL software. At the end of each run, the final optimal models were sorted in terms of increasing values of coefficient of determination (CoD) computed on the basis of the training data. This coefficient increases from 0 to 1, when the model complexity in terms of N_a and N_x increases. The lowest value of 0 is obtained by the simplest model which always returns a value equal to the mean of the output in the training phase. The value equal to 1 would be reached by a theoretical model that fully explains the phenomenon under consideration. The training and testing results (CoD_{train} and CoD_{test}) are reported in Figures 4a and 5.

In particular, Figure 4a reports results for the modeling of COD in the Ecully catchment. The graph shows that, for less complex models (corresponding to the lower values of CoD_{test}), CoD_{test} is quite close to CoD_{train}. When a certain model complexity is reached, the distance between CoD_{test} and CoD_{train} tends to increase. In fact, the graph shows that the CoD_{test} and CoD_{train} dots are further and further when CoD_{test} (and then the model complexity) grows. Graphs such as that in Figure 4a, where CoD_{train} and CoD_{test} are plotted together, then help in making distinction between robust models (with CoD_{test} close to CoD_{train}) and fragile models (with CoD_{test} much lower than CoD_{train}). Figure 4b reports the results of a benchmark run, where EPR was applied to the whole set of 56 input variables, without any prior application of the input selection procedure. In this benchmark run, the wider set of exponents of [−3, −2.5 −2, −1.5, −1, −0.5, 0, 0.5, 1, 1.5, 2, 2.5, 3] was adopted. Furthermore, population and generation parameters in the EPR-MOGA-XL software were set to 6 and 20, respectively. This resulted in the same computational burden as when applying first the input selection procedure and then EPR-MOGA in the other case. The comparison of the CoD_{train} and CoD_{test} values in Figure 4a with the corresponding values in Figure 4b highlights better fitting performance for the case of EPR applied following input selection in Figure 4a. Some negative values are even observed for COD_{test} in Figure 4b. Furthermore, the graphs in Figure 4 show that the models obtained in the benchmark MOGA-EPR run (Figure 4b) feature, on average, a larger CoD decay from the training to the testing phases, than those obtained using a relevant input selection strategy (Figure 4a). Ultimately, this attests to the efficiency of the input selection procedure herein proposed, which helps in obtaining robust models with better fitting performance.

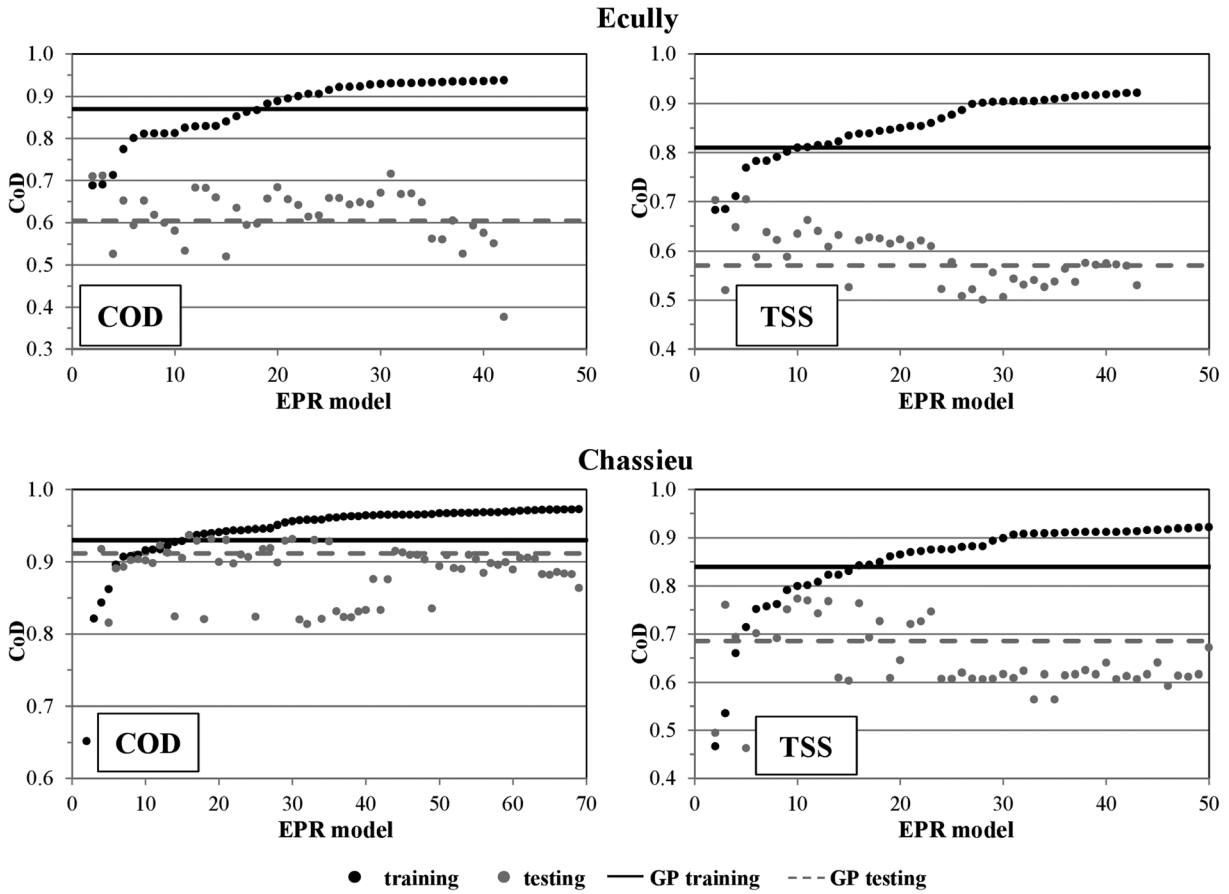

Figure 5. Values of coefficient of determination CoD obtained by the EPR models for the representation of COD (kg) and TSS (kg) in the Ecully and Chassieu catchments in the training and testing phases. In the graphs, EPR models are sorted according to ascending values of CoD. CoD values of the models obtained by *Sun and Bertrand-Krajewski* [2011], using the GP technique in the testing and training phases.

The graphs in Figure 5 show the results obtained by applying EPR-MOGA to the set of relevant input variables, in both catchments for modeling both COD and TSS. The graphs also display, as a benchmark, the CoD_{train} and CoD_{test} values reported by *Sun and Bertrand-Krajewski* [2011] obtained by using the genetic programming (GP) technique. The graphs highlight that the runs performed were able to yield a number of models with better performance than GP, as indicated by higher CoD values and lower differences between training and testing for EPR-MOGA than for GP. This result may be ascribed to the effective search for data-driven models in the EPR-MOGA technique as well as from the efficient selection of the relevant input variables.

Table 2. EPR-MOGA Models Chosen to Have CoD_{train} Values Close to the GP Models Obtained by *Sun and Bertrand-Krajewski* [2011]

Catchment	EPR-MOGA-XL model	$CoD_{train} - CoD_{test}$ by EPR	$CoD_{train} - CoD_{test}$ by GP
Ecully	$COD = 0.020106 \frac{MRI(30,48)TFV^{2.5}}{TRD^{0.5}ADP15^{0.5}MRI(30,96)^{2}MFR^{1.5}} + 0.37998MFR + 0.093824ADP30^{0.5}MFR + 176.0661TRD^{0.5}$	0.88 − 0.66	0.87 − 0.60
Ecully	$TSS = 0.073534TFV + 0.000037636MFR^{2} + 3.9991ADP5^{0.5}MFR^{0.5} + 61.3158$	0.81 − 0.66	0.81 − 0.57
Chassieu	$COD = 0.056207TFV + 5.5718MFR^{0.5} + 0.00035637MRI\ ADP15^{0.5}DRF^{0.5}\ TFV$	0.94 − 0.94	0.93 − 0.91
Chassieu	$TSS = 0.016373ADP20^{0.5}MRI(30,4)^{0.5}MFR + 0.00024997\frac{MRI\ TFV^{1.5}}{MRI5} + 87.9863$	0.85 − 0.73	0.84 − 0.69

Table 3. EPR-MOGA Models in the Run Related to the Ecully Catchment and to COD as Output Variable

EPR-MOGA-XL Model	$CoD_{train} - CoD_{test}$
$COD = 22.8704\ TFV^{0.5}$	0.69 − 0.71
$COD = 0.13184\ TFV + 28.7659 MFR^{0.5}$	0.78 − 0.65
$COD = 0.13144\ TFV + 0.19267 ADP5^{0.5} MFR + 264.7627$	0.80 − 0.59
$COD = 0.13185\ TFV + 0.00024437 ADP30^{2.5} + 0.19258 ADP5^{0.5} MFR + 224.4755$	0.81 − 0.60

In the graphs in both Figures 4 and 5, CoD_{test} is always noticeably lower than CoD_{train}. This may be due to the way the data were originally divided into training (first two third of the data) and testing (last one third of the data) phases and could then be possibly improved by different data subdivision. In this study, a similar data subdivision strategy to *Sun and Bertrand-Krajewski* [2011] was used in order to facilitate comparison with the Authors' models in terms of goodness of fit. Nevertheless, the issue of data subdivision is out of the scope of the paper.

For either catchment and for either target variable, an EPR-MOGA model was then selected in such a way as to have similar CoD_{train} values to *Sun and Bertrand-Krajewski* [2011]. Selected models are reported in Table 2, as well as their CoD_{train} and CoD_{test} values. The analysis of Table 2 mainly highlights two aspects:

1. The EPR-MOGA models are compact and easy to interpret due to their simple polynomial structure. This is an advantage compared to other techniques, such as GP, which may produce very complex structures [*Luke and Panait*, 2006];
2. The performance of the selected EPR-MOGA models does not deteriorate significantly when passing from training to testing phases, unlike the GP models by *Sun and Bertrand-Krajewski* [2011]. This demonstrates the improved robustness of EPR models, benefitting from both the preliminary input selection and the modeling paradigm itself.

In many cases, the EPR models are so simple that they can show the dependence of the output on the explanatory variables in a very straightforward way.

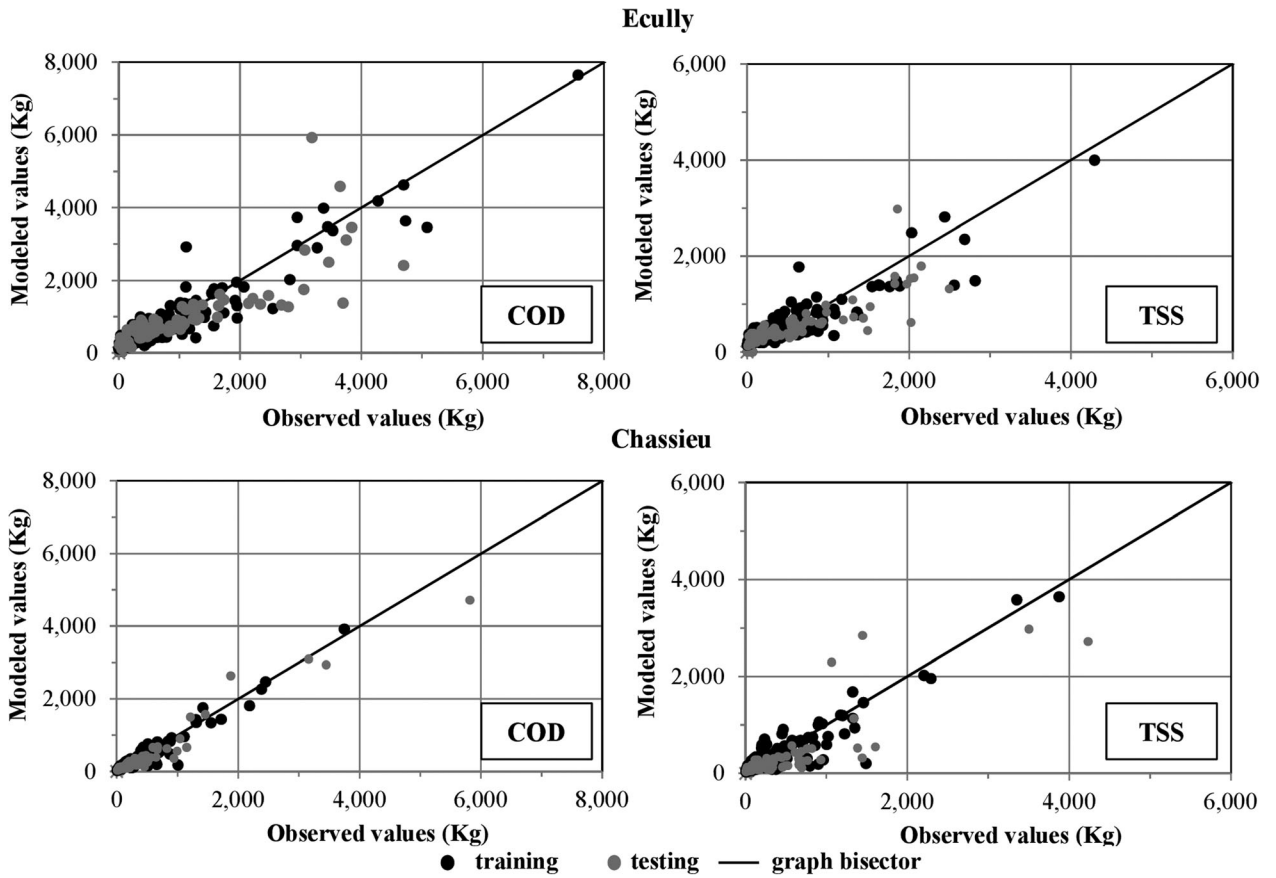

Figure 6. Values of COD and TSS in kg (dots) obtained through the EPR-MOGA models reported in Table 2 as a function of the observed values in the Ecully and Chassieu catchments and graph biceptor (line).

Table 4. Indices Used to Represent the Goodness of Fit

Name	Formula
Coefficient of Determination	$\mathrm{CoD}=1-\frac{\sum_{i=1}^{N}(y_i-f_i)^2}{\sum_{i=1}^{N}(y_i-\overline{y})^2}$
Mean Squared Error	$\mathrm{MSE}=\frac{1}{N}\sum_{i=1}^{N}(y_i-f_i)^2$
Correlation	$\rho=\frac{\sum_{i=1}^{N}(y_i-\overline{y})(f_i-\overline{f})}{\sqrt{\sum_{i=1}^{N}(y_i-\overline{y})^2}\sqrt{\sum_{i=1}^{N}(f_i-\overline{f})^2}}$
Index of agreement	$d=1-\frac{\sum_{i=1}^{N}(y_i-f_i)^2}{\sum_{i=1}^{N}(\lvert y_i-\overline{y}\rvert+\lvert f_i-\overline{y}\rvert)^2}$

The analysis of the structure highlights that, in the Ecully catchment, there is a direct relationship between output TSS and explanatory variables TFV, MFR, and ADP5. In the Chassieu catchment, output COD is directly related to TFV, MFR, MRI, ADP15, and DRF.

In the case of the model representing TSS in the Chassieu catchment, in addition to TFV, MFR, ADP20, MRI(30,4), and AFR, the output also relates directly to the dimensionless MRI/MRI5.

The structure of the model for representing COD in the Ecully catchment is more difficult to interpret. In order to have a cleaper insight into the dependency between the output variable and the various explanatory variables in this case, the models with lower fitting performance belonging to the same EPR-MOGA run can be referred to. In the case of output COD in the Ecully catchment, some models with lower fitting performance than that in Table 2 are reported in Table 3.

The simpler polynomial models in Table 3 highlight the direct relationship of COD with TFV, MFR, ADP5, and ADP30. The table also shows how these variables are gradually included in the model structure as the complexity increases. This helps in understanding which explanatory variables influence COD most. In particular, TFV is present starting from the simplest model and this variable alone is able to explain the phenomenon with a CoD = 0.69. MFR is added to the second model whereas ADP5 appears in the third model. Adding MFR and ADP5 to TFV increased CoD to 0.8. Finally, ADP30 appears only in the most complex model

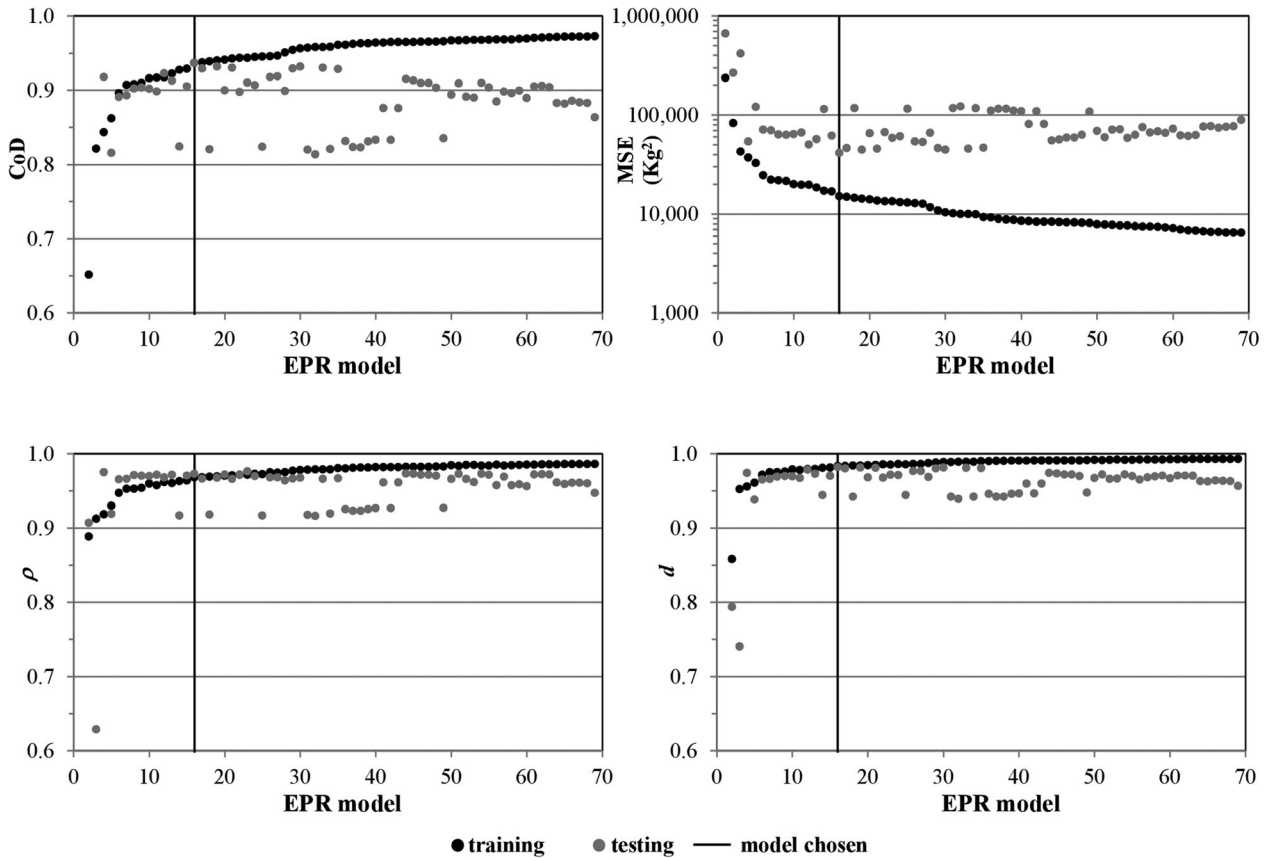

Figure 7. Values of the coefficient of determination (CoD), Mean Squared Error (MSE), correlation ρ, and index of agreement d for the various EPR-MOGA models obtained to simulate COD in the Chassieu catchment. Values for the model chosen in Table 2 are highlighted by the vertical line.

in Table 3, yielding only a marginal CoD increase. In order to further improve CoD, other variables need to be added. However, this results in more complex polynomial forms (see Table 2).

Figure 6 displays the scatterplots of modeled and observed values in both catchments for the EPR-MOGA models reported in Table 2. Globally, these graphs highlight the good alignment of data values along the graph bisector, attesting to the good agreement of modeled and observed values in both the training and testing phases.

Since various authors [e.g., *Gupta et al*., 2009] proved that the choice of a particular goodness-of-fit indicator within the optimization may affect the results, an analysis was carried out to assess model performance by a number of indicators in addition to CoD, including the Mean Squared Error (MSE), the correlation coefficient (ρ), and the index of agreement (d). This was done by postprocessing the results of the optimal EPR-MOGA models obtained using COD in the Chassieu catchment (see Figure 5). The formulas used to assess the goodness-of-fit indicators as a function of N modeled (f) and observed (y) values are shown in Table 4. In the graphs in Figure 7, the values of the indicators in both the training and testing phases are shown. Similarly, to the trend of CoD plotted against model complexity in the training phase, the corresponding MSE is, as expected, strictly monotonous. The trend of the other goodness-of-fit indicators is globally monotonous, despite some very small oscillations, which can hardly be spotted in the graphs. Therefore, the graphs show a high agreement between the four goodness-of-fit indicators. The fact that CoD generally agrees well with the other indicators then corroborates the choice of CoD as an objective function in the optimization process. The graphs in Figure 7 also show that the model chosen in Table 2 features a small decay from the training to the testing phases in terms of all the goodness-of-fit indicators.

4. Discussion

Most algorithms in the scientific literature [e.g., *Bowden et al*., 2005a, 2005b; *D'Heygere et al*., 2006; *Giustolisi and Simeone*, 2006; *Yang and Ong*, 2011; *Wan Jaafar et al*., 2011; *Tirelli and Pessani*, 2011] address input selection by iteratively including or excluding potential inputs toward models with high goodness-of-fit performance. In these cases, a model structure needs to be predefined, which is usually challenging for a complex problem such as storm water quality. In this study, an input selection method is developed through explorative runs of EPR-MOGA or MCS-EPR-MOGA, which enable model construction without predefining the model structure. By considering frequency with which each potential input appears in the EPR models, the user can get an estimate of the extent to which that variable is influential in modeling the output. Based on this methodology, input selection can be easily performed. A major advantage of the method thus lies in its ability to perform input selection without predefining a model structure. Furthermore, in comparison with other data-driven models (e.g., ANN and EP models), many near-optimal models are generated in one run of EPR-MOGA or MCS-EPR-MOGA, while other methods need a number of runs to identify multiple optimal models. Furthermore, the simple form of EPR-MOGA or MCS-EPR-MOGA models allows interpretation of their mathematical expressions and parameters thus facilitating explanation of the phenomenon. As demonstrated in this study, the EPR-MOGA and MCS-EPR-MOGA models are capable of simulating total loads of pollutants with better goodness-of-fit performance than other data-driven methodologies [*Sun and Bertrand-Krajewski*, 2011; *Sun and Bertrand-Krajewski*, 2013] or physically based models [*Métadier*, 2011].

Like previous data-driven studies, this work does not use the EPR-MOGA and MCS-EPR-MOGA models to analyze the dynamics of storm water events, that is, the interevent process is not simulated. It is worth mentioning that, as an attempt, EPR-MOGA was also applied to model the peak pollutant concentration. However, the peak values, which are assessed on short time intervals, are more prone to random factors and measurement uncertainty, and are thus difficult to model using only rainfall and runoff variables (see Appendix A). The modeling of pollutant dynamics in the interevent process remains challenging for either data-driven or physically based models [e.g., *Métadier*, 2011].

The potential input variables for simulating pollutant loads include rainfall and runoff variables collected on the field. Even when runoff measurements are not available, a calibrated rainfall-runoff model can be used to estimate runoff from rainfall data. Input selection and EPR-MOGA or MCS-EPR-MOGA can then be applied to simulate total loads of pollutants using rainfall and simulated runoff data.

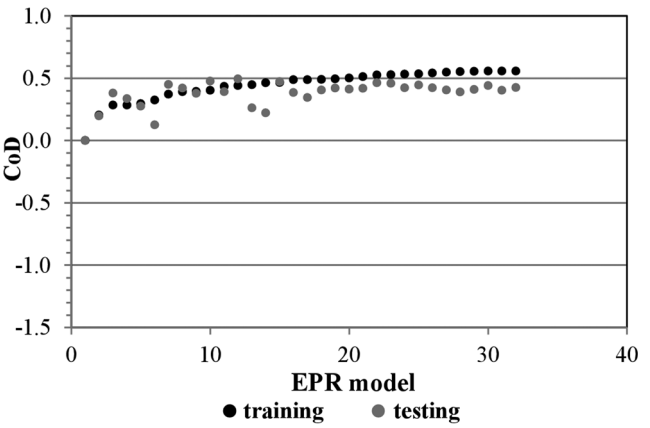

Figure 8. Values of coefficient of determination CoD obtained by the EPR models in the Ecully catchment in the training (black dots) and testing (grey dots) phases for the representation of COD_{max} (mg/L); results of application of EPR alone. In the graph, EPR models are sorted according to ascending values of CoD_{max}.

5. Conclusions

This paper presented a procedure for the selection of the relevant input variables in the context of the multiobjective evolutionary polynomial regression (EPR-MOGA) and of its multi case strategy (MCS-EPR-MOGA) extension. EPR-MOGA and MCS-EPR-MOGA are nonlinear regression techniques yielding data-driven models able to express the relationship between the target and explanatory variables of a certain phenomenon with growing structure complexity and fitting performance.

The input selection procedure consists in performing an exploratory run of EPR-MOGA and MCS-EPR-MOGA. At the end of this run, a statistical and engineering judgement-based analysis of the occurrences of the potential explanatory variables in the models is performed.

Applications concerned derivation of the relevant input variables to describe storm water quality in two French catchments. Results showed that the input selection procedure reduced significantly the number of explanatory variables in modeling two indicators of storm water quality (i.e., TSS and COD). A proof of the effectiveness of the procedure was supplied by the results of subsequent EPR-MOGA runs using different sets of potential inputs. Results proved that

1. The proposed input selection procedure enabled the fitting performance and robustness of the models obtainable by EPR-MOGA to be improved;
2. The EPR-MOGA models obtained following input selection featured simple structures and high model robustness compared to the results using other data-driven modeling techniques in previous studies.

It has to be noted that the input selection procedure described in this paper could also be applied to other data-driven modeling techniques. The advantage of using EPR-MOGA or MCS-EPR-MOGA lies in their multiobjective modeling approach. Thanks to this, these techniques are able to produce many models in a single run, based on which analysis of variable occurrences can be easily made. Should another technique, such as the genetic programming, be used instead, numerous runs would be required involving a much larger computational cost. Another advantage of EPR-MOGA and MCS-EPR-MOGA is that they produce models that are easier to interpret and check using engineering judgement.

Appendix A

In a preliminary analysis, EPR-MOGA optimizations were performed to analyze the feasibility of the maximum concentrations of COD and TSS as target variables in EPR-MOGA models as an alternative option to the total loads. As an example, the graph in Figure 8 reports the results of an optimization performed to model the maximum concentration of COD, COD_{max}, in the Ecully catchment. In this optimization, the population and generation parameters in the EPR-MOGA-XL software were set to 6 and 20, respectively, and the wider set of exponents of [−3, −2.5 −2, −1.5, −1, −0.5, 0, 0.5, 1, 1.5, 2, 2.5, 3] was adopted. Furthermore, all the 56 explanatory variables were considered. The results of the optimization show that the CoD values obtained in the training and testing phases are always lower than 0.6. The comparison with the optimization related to the total load of COD, which had produced CoD values up to 0.88 (see Figure 4b), highlights the fact that such regression models as the EPR-MOGA models, perform better when event mean concentrations, or total loads, have to be reproduced. This led to the choice of the total loads of COD and TSS as target variables in all the optimizations of this paper.

The reason for the bad results of EPR-MOGA is that the maximum pollutant concentration, which is assessed on short time intervals, is more prone to random factors and is thus possibly less related to the event rainfall

and runoff characteristics listed in Table 1. In addition, measurement uncertainty in the maximum pollutant concentration values is significant in comparison to that in event loads due to accumulated effect of uncertainty.

Notation

a	coefficient in the monomials in the generic EPR model.
ADP	antecedent dry period (day).
ADPJ	antecedent dry period from the last rainfall event exceeding J mm (Y = [5:5:40]) (day).
AFR	average flow rate (L/s).
CoD	coefficient of determination.
COD	chemical oxygen demand event load (kg).
COD_{max}	maximum concentration of the chemical oxygen demand (mg/L).
CRM	cumulative rainfall during M hours before the rainfall event (M = 4, 8, 12, 24, 36, 48, 52, 72, 96) (mm).
d	index of agreement.
DRF	duration of the runoff flow (h).
EPR	acronym for evolutionary polynomial regression.
ES	set of exponents which can be selected in EPR.
f	function in the EPR polynomial structure.
$\overline{f}$	average value of f.
GP	acronym for genetic programming.
MCS	multicase study.
MFR	maximum flow rate (L/s).
MOGA	acronym for multiobjective genetic algorithm.
MRI(Y,Z)	maximum value of the average rainfall intensity over a Y minute long time interval during Z hours before an event (Y = 5, 10, 30; Z = 4, 8, 12, 24, 36, 48, 52, 72, 96) (mm/h).
MRI	maximum rainfall intensity (mm/h).
MRI5	maximum rainfall intensity in 5 min (mm/h).
MSE	mean squared error (unit of measurement of the variable considered, raised to 2).
N	number of data.
N_a	number of monomials in the generic EPR model.
NDAY	number of the day in a year when an event occurs (1–365).
NHOUR	number of the hour in a day when an event occurs (1–24).
$Nsel$	number of relevant input variables to be selected.
N_x	number of input occurrences in the generic EPR model.
RI	rainfall intensity (mm/h).
TFV	total flow volume (m^3).
TRD	total rainfall duration (h).
TRH	total rainfall depth (mm).
TSS	total suspended solids event load (kg).
X	explanatory variable in the generic EPR model.
Y	explained variable in the generic EPR model.
y	generic observed variable.
$\overline{y}$	average value of y.
ρ	correlation.

Acknowledgments
The data sets of the Chassieu and Ecully catchments have been collected and made available by OTHU—Field Observatory on Urban Hydrology (www.othu.org) with the financial support of the Greater Lyon and the Rhône-Méditerranée & Corse Water Agency.

References

Aryal, R., J. Kandasamy, S. Vigneswaran, R. Naidu, and S. H. Lee (2009), Review of stormwater quality, quantity and treatment methods. Part 1: Stormwater quantity modeling, *Environ. Eng. Res.*, *14*(2), 71–78.

Ashley, R. M., J.-L. Bertrand-Krajewski, T. Hvitved-Jacobsen, and M. Verbanck (Eds.) (2004), *Solids in Sewers, Sci. Tech. Rep. 14*, 360 pp., IWA Publ., London.

Barraud, S., J. Gibert, T. Winiarski, and J.-L. Bertrand-Krajewski (2002), Implementation of a monitoring system to measure impact of stormwater runoff infiltration, *Water Sci. Technol.*, *45*(3), 203–210.

Bannerman, R. T., D. W. Owens, R. B. Dodds, and N. J. Hornewer (1993), Sources of pollutants in Wisconsin stormwater, *Water Sci. Technol.*, *28*(3-5), 241–259.

Berardi, L., and Z. Kapelan (2007), Multi-case EPR strategy for the development of sewer failure performance indicators, in *Proceedings of World Environmental and Water Resources Congress 2007*, pp. 1-12, American Society of Civil Engineers, Reston, Va., doi:10.1061/40927(243)162.

Berardi, L., Z. Kapelan, D. A. Savic, and O. Giustolisi (2006), Modelling sewer performance indicators, in *Proceedings of the 7th International Conference on HydroInformatics, HIC 2006*, vol. IV, pp. 2829–2836, Research Publishing Services, Singapore.

Berardi, L., Z. Kapelan, O. Giustolisi, and D. A. Savic (2008), Development of pipe deterioration models for water distribution systems using EPR, *J. Hydroinformatics*, *10*(2), 113–126.

Bowden, G. J., H. R. Maier, and G. C. Dandy (2005a), Input determination for neural network models in water resources applications. Part 1. Background and methodology, *J. Hydrol.*, *301*, 75–92.

Bowden, G. J., H. R. Maier, and G. C. Dandy (2005b), Input determination for neural network models in water resources applications. Part 2. Case study: Forecasting salinity in a river, *J. Hydrol.*, *301*, 93–107.

Dembélé, A., J.-L. Bertrand-Krajewski, and B. Barillon (2010), Calibration of stormwater quality regression models: A random process?, *Water Sci. Technol.*, *62*(4), 875–882.

D'Heygere, T., P. L. M. Goethals, and N. De Pauw (2006), Genetic algorithms for optimisation of predictive ecosystems models based on decision trees and neural networks, *Ecol. Model.*, *195*, 20–29.

Donigian, A. S., and W. C. Huber (1991), Modeling of nonpoint source water quality in urban and non-urban areas, *Rep. EPA/600/3-91/039*, U.S. Environ. Prot. Agency, Washington, D. C.

Dotto, C. B. S., M. Kleidorfer, A. Deletic, W. Rauch, D. T. McCarthy, and T. D. Fletcher (2011), Performance and sensitivity analysis of stormwater models using a Bayesian approach and long-term high resolution data, *Environ. Model. Software*, *26*(10), 1225–1239.

Galelli, S., G. B. Humphrey, H. R. Maier, A. Castelletti, G. C. Dandy, and M. S. Gibbs (2014), An evaluation framework for input variable selection algorithms for environmental data-driven models, *Environ. Modell. Software*, *62*, 33–51.

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

Giustolisi, O., and D. A. Savic (2009), Advances in data-driven analyses and modelling using EPR-MOGA, *J. Hydroinformatics*, *11*, 225–236.

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

Gupta, H. V., H. Kling, K. K. Yilmaz, and G. F. Martinez (2009), Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, *J. Hydrol.*, *377*, 80–91.

Haykin, S. (1999), *Neural Networks: A Comprehensive Foundation*, Prentice Hall, Upper Saddle River, N. J.

Koza, J. R. (1992), *Genetic Programming: On the Programming of Computers by Meams of Natural Selection*, 813 pp., MIT Press, Cambridge, Mass.

Lacour, C., C. Joannis, and G. Chebbo (2009), Assessment of annual pollutant loads in combined sewers from continuous turbidity measurements: Sensitivity to calibration data, *Water Res.*, *43*(8), 2179–2190.

Laucelli, D., and O. Giustolisi (2011), Scour depth modelling by a multi-objective evolutionary paradigm, *Environ. Modell. Software*, *26*, 498–509.

Ljung, L. (1999), *System Identification: Theory for the User*, 2nd ed., Prentice Hall, Upper Saddle River, N. J.

Luke, S., and L. Panait (2006), A comparison of bloat control methods for genetic programming, *Evol. Comput.*, *14*(3), 309–344.

Métadier, M. (2011), Traitement et analyse de séries chronologiques continues de turbidité pour la formulation et le test demodèles des rejets urbains par temps de pluie (Processing and analysis of turbidity time series for determination and testing of stormwater quality models) [in French], PhD thesis, 400 pp., INSA, Lyon, France. [Available at http://www.safege.com/wp-content/uploads/2014/06/these-mm-072011.pdf.]

Metadier, M., and J.-L. Bertrand-Krajewski (2012), The use of long-term on-line turbidity measurements for the calculation of urban stormwater pollutant concentrations, loads, pollutographs and intra-event fluxes, *Water Res.*, *46*(20), 6836–6856.

Mourad, M., J.-L. Bertrand-Krajewski, and G. Chebbo (2005), Calibration and validation of multiple regression models for stormwater quality prediction: Data partitioning, effect of dataset size and characteristics, *Water Sci. Technol.*, *52*(3), 45–52.

Savic, D. A., O. Giustolisi, L. Berardi, W. Shepherd, S. Djordjevic, and A. Saul (2006), Modelling sewer failure by evolutionary computing, *Water Manage. J.*, *159*(2), 111–118.

Savic, D. A., O. Giustolisi, and D. Laucelli (2009), Asset deterioration analysis using multi-utility data and multi-objective data mining techniques, *J. Hydroinformatics*, *11*(3), 212–224.

Sun, S., and J.-L. Bertrand-Krajewski (2011), The calibration of urban storm water quality models using genetic programming (GP), in *Urban Water Management: Challenges and Opportunities*, edited by D. A. Savic, Z. Kapelan, and D. Butler, pp. 663–668, Cent. for Water Syst., Univ. Exeter, Exeter, U. K.

Sun, S., and J.-L. Bertrand-Krajewski (2012), On calibration data selection: The case of stormwater quality regression models, *Environ. Modell. Software*, *35*, 61–73.

Sun, S., and J.-L. Bertrand-Krajewski (2013), Input variable selection and calibration data selection for storm water quality regression models, *Water Sci. Technol.*, *68*(1), 50–58.

Tirelli, T., and D. Pessani (2011), Importance of feature selection in decision-tree and artificial-neural-network ecological applications. Alburnus alburnus alborella: A practical example, *Ecol. Informatics*, *6*, 309–315.

Vaze, J., and F. H. S. Chiew (2003), Comparative evaluation of urban storm water quality models, *Water Resour. Res.*, *39*(10), 1280, doi:10.1029/2002WR001788.

Wan Jaafar, W. Z., J. Liu, and D. Han (2011), Input variable selection for median flood regionalization, *Water Resour. Res.*, *47*, W07503, doi:10.1029/2011WR010436.

Yang, J. B., and C. J. Ong (2011), Feature selection using probabilistic prediction of support vector regression, *IEEE Trans. Neural Networks*, *22*, 954–962.