Modelling energy performance using a new hybrid DE/MARS–based approach for fossil-fuel thermal power stations

Paulino José García–Nieto[a,*], Esperanza García–Gonzalo[a], José Pablo Paredes–Sánchez[b], Antonio Bernardo Sánchez[c]

[a]Department of Mathematics, Faculty of Sciences, University of Oviedo, 33007 Oviedo, Spain

[b]Department of Energy, College of Mining, Energy and Materials Engineering, University of Oviedo, 33004 Oviedo, Spain

[c]Department of Mining Technology, Topography and Structures, University of León, León 24071, Spain

Abstract Despite their environmental impact, fossil-fuel power plants are still commonly used due to their high capacity and relatively low cost compared to renewable energy sources. The aim of this paper is to assess the performance of such energy systems as a key element within a fossil-fuel energy supply network. The methodology relies on fossil-fuel power plant modelling to define an optimal energy management level. However, it can be difficult to model the energy management of Thermal Power Stations (TPS). Therefore, this paper shows an energy efficiency model found on a new hybrid algorithm that is a combination of multivariate adaptive regression splines (MARS) and differential evolution (DE) to estimate net annual electricity generation (NAEG) and carbon dioxide (CO_2) emissions (CDE) from economic and performance variables in thermal power plants. This technique requires the DE optimisation of the MARS hyperparameters during the development of the training process. In addition to successfully forecast net annual electricity generation (NAEG) and carbon dioxide (CO_2) emissions (CDE) (coefficients of determination with a value of 0.9803 and 0.9895, respectively), the mathematical

[*]Corresponding author. Tel.: +34-985103417; fax: +34-985103354.
E-mail address: lato@orion.ciencias.uniovi.es (P.J. García Nieto).

model used in this work can determine the importance of each economic and energy parameter to characterize the behaviour of thermal power stations.

Keywords Differential evolution (DE); Energy management; Multivariate adaptive regression splines (MARS); Regression analysis; Thermal power stations; Power network

Introduction

The world has recently been undergoing a major technological transition from the use of fossil energy sources to renewable ones, as high levels of fossil fuel consumption have severely affected government economic policies, environmental conditions and energy supply security (Behera et al. 2018). At the same time, however, the emergence of renewable energy sources within the electricity grid has progressively increased. The process of integrating renewable energy sources and fossil fuels into the electricity transmission grid creates challenges in managing energy production in real-time (Jayawardene and Venayagamoorthy 2015). The management of energy is currently one of the greatest challenges that applied energy studies have to face (Paredes-Sánchez et al. 2014). However, fossil fuels, such as coal, natural gas and oil, still account for 80% of global energy consumption while half of all the electricity is still produced in coal-fired power plants (Li et al. 2016). The resulting technological and market transformations across the power sector have translated into energy performance conditions that have yet to be quantified. Understanding the current level of energy performance is crucial to the development and application of different strategies worldwide (Paredes-Sánchez et al. 2015). The development of this energy strategy has established a correlation of sustainable energy production and environmental security around the world. Some of the main drive forces from sustainable energy production and energy efficiency point of view are identified in Vidadili et al.

(2017) to be: energy transition security, reduced carbon dioxide emissions, reduced energy production cost and massive integration of green energy technologies. In this sense, industrial revolution and environmental sustainability of any nation or community of people is the access to reliable and clean energy. In this context, US power plants are a driving force in a state of transition, which has resulted from significant changes in technological innovations, energy conversion, and management (Peer and Sanders 2018). Moreover, energy supply is a dynamic challenge based on energy management.

This paper analyses an innovative hybrid regressive model that uses adaptive multivariate regression splines (Friedman 1991; Sekulic and Kowalski 1992; Friedman and Roosen 1995; Hastie et al. 2003), also referred to globally as MARS, which is used alongside the method of evolutionary optimisation of Differential Evolution (DE) (Storn and Price 1997; Price et al. 2005; Rocca et al. 2011; Feoktistov 2006) in order to predict net annual electricity generation (NAEG) in fossil-fuel thermal power stations. In addition, carbon dioxide (CO_2) emissions (CDE) are thoroughly analysed in fossil-fuel thermal power stations. Fig. 1a shows an example of these thermal power stations, specifically an aerial photograph of the location of the fossil-fuel power plant of Soto de Ribera just a short distance from Oviedo city (Northern Spain) and Fig. 1b shows the same fossil-fuel power plant at a larger scale.

Fig. 1 a An aerial photograph of an example of fossil-fuel power plant (i.e., Soto de Ribera power plant) and **b** the same fossil-fuel power plant at a larger scale

Recent studies have analysed the performance of power plants and the trade-offs that follow more general shifts in fuel use. The global energy efficiency assessment considers both the reliability of the energy supply and the effective control of energy consumption in the different energy systems. However, these methodologies must take into account a number of factors in order to maximise production (Paredes-Sánchez et al. 2016). The following are four main methodologies for assessing energy efficiency: stochastic frontier analysis, data envelopment analysis, exergy analysis and benchmark comparison (Li and Tao 2017). Moreover, a multivariate adaptive regression splines (MARS) technique can be used to predict values for different applications (Friedman 1991; Sekulic and Kowalski 1992) and is a nonparametric regression technique which provides an analysis based on both nonlinearities and variable interactions by means of a complex linear model (Friedman 1991; Sekulic and Kowalski 1992; Friedman and Roosen 1995; Hastie et al. 2003).

Some advantages of applying the MARS method over other existing techniques include (Friedman and Roosen 1995; Hastie et al. 2003): (i) it provides more flexible models than linear regression models; (ii) it is easy to interpret and understand; (iii) it can manage continuous and categorical data; (iv) the hinge functions automatically partition the input data, so the effect of outliers is contained; (v) it does automatic *variable selection* (meaning that it includes important variables in the model and excludes unimportant ones); (vi) it tends to have a good bias-variance trade-off; and (vii) it offers an explicit mathematical expression of the dependent variable as a function of independent variables through an expansion of the base functions (*hinge functions and products of several hinge functions or interactions*).

In addition, the differential evolution (DE) technique was used in order to optimise the MARS hyperparameters in the training stage. DE is a global metaheuristic evolutionary method derived from genetic algorithms (GA) that is intrinsically capable of solving multidimensional optimisation problems involving continuous variables, as well as other calculation algorithms such as particle swarm optimisation (PSO) (Eberhart 2001; Clerc 2006; Olsson 2011) or ant colony optimisation (Dorigo and Stützle 2004). In addition, it is an inspired algorithm that results in remarkable quality solutions to optimisation problems by mutation, recombination and selection, that is to say, bio-inspired operators (Storn and Price 1997; Price et al. 2005; Simon 2013; Yang et al. 2013). Previous research has presented that MARS is a suitable tool in a variety of fields such as engineering, computing and medicine (Xu et al. 2004; Vidoli 2011; Zhang and Goh 2013; Zhang et al. 2015). Particularly, there are several studies in the literature on the use of the MARS approach in different energy research contexts (Sobri et al. 2018; Kisi et al. 2017; García Nieto and Álvarez Antón 2014). Mardani and coworkers (Mardani et al. 2017) and Sueyoshi and coworkers (Sueyoshi et al. 2017) reviewed the literature on the use of modelling techniques to measure energy efficiency. In recent years, MARS has been undertaken within energy systems due to its success in explaining complete systems (Cheng and Cao 2014; Sekhar Roy et al. 2018). Chen and coworkers (Chen et al. 2017) have studied an acquired meta-model of the thermal system assessment to develop a more accurate model that can provide performance predictions. Al-Musaylh and coworkers (Al-Musaylh et al. 2018) have also studied electricity demand in energy systems. However, nowadays there is a lack of performance assessment of fossil-fuel power plants that use the MARS technique.

This research study aims to cover that gap and demonstrate the potential of model fossil-fuel power plants through machine learning, and the importance of energy performance to ensure power supply. Moreover, the present study also discusses the overall main process parameters of energy production and includes a comprehensive discussion of the main parameters involved. Specifically, a hybrid DE optimised MARS (DE/MARS) model (Simon 2013; Yang et al. 2013; Fister et al. 2015) has been applied to estimate the net annual electricity generation NAEG and CDE from the economic and performance input variables in thermal power stations.

This research work is organised as follows: To start with, the necessary materials and methods to carry out this work are detailed. Secondly, the results obtained are shown and discussed. Lastly, the conclusions derived from the results are explained in detail. This work will hopefully help pave the way toward providing accurate energy performance analysis for safe energy production in power thermal systems.

Materials and methods

Experimental dataset

This study provides an innovative and systematic approach to evaluate the performance of fossil-fuel power plants. The case study is based on the available US database (Tajbakhsh and Hassini 2018). The selected databases include Clean Air Task Force (CATF 2018), Energy Information Administration (EIA 2018), United States Environmental Protection Agency (USEPA 2018) and Federal Energy Regulatory Commission (FERC 2018). In total, over 1,080 sources of real data containing quantitive information about power plant performance have been categorised.

This research work is aimed to establish a means of estimating the electrical power generation and the resulting carbon dioxide pollution indicators using easily measurable variables. The two MARS models use the same economic and performance variables and the output variables include the net annual electricity generation (TWh) and carbon dioxide (CO_2) emission (Mton).

The net annual electricity generation (NAEG) is relevant because it is directly related to both the total gross power generation as well as the amount of electricity generated by a power plant, which is transmitted and distributed for consumer use. This dependent variable is directly linked to energy management in order to produce electric power from primary energy sources. Likewise, the second dependent variable is carbon dioxide (CO_2) (CDE), which is a greenhouse gas that absorbs and emits infrared radiation on its two active vibration frequencies. This process causes the carbon dioxide to heat both the surface of the earth and the lower atmosphere as well as to, simultaneously, cool the upper atmosphere. The vast majority of climatologists agree that the increase in the atmospheric concentration of CO_2, and the greenhouse effect it creates, is the main reason for the increasing average global temperatures since the middle of the 20th century. Although the main greenhouse gas responsible for this warming is carbon dioxide, there are other long-lasting greenhouse gases (such as methane, nitrous oxide and ozone) which also contribute to the process. However, CO_2 remains the greatest cause of concern, since it plays a much greater role in total warming than all the other gases combined and has a long atmospheric life. Localised concentrations of carbon dioxide invariably reach extremely high levels near fossil-fuel power plants. Therefore, it is necessary to monitor and control such emissions to the atmosphere. Energy models have been established for energy systems and power plants from different points of view. To be more precise, according to underlying assumptions, relevant literature includes studies that show their energy potential, pollutant emissions or energy performance (Paredes-Sánchez et al. 2019). This innovative machine-learning model involves techno-economic parameters ranging from energy resources to management (Díez et al. 2005; Wang et al. 2020; Zhao et al. 2013). For all of these reasons, these main parameters (or input variables) have been identified as economic and performance variables indicated below.

➢ Economic variables:

- Book value of plant and land (BVPL) (in millions): the physical assets, including land and buildings, of the fossil-fuel power plant.
- Annual production expenses (APE) (in millions): the necessary costs to keep the plant operating. The operating costs of fossil-fuel power plants include those associated with fuel, labour, and maintenance.
- Plant nameplate capacity (PNC) (in MW): the intended sustained full-load output of a facility.
- Annual number of employees (ANE): the number of workers that operate and maintain the plant in operating conditions, i.e.: plant operators, distributors, and dispatchers.
- Annual revenue (AR) (in millions): the income generated by normal business activity, usually from the sale of power supply to customers.

➢ Performance variables:

- Annual fuel consumed (AFC) (in MJ): the amount of fossil-fuel energy consumed by the power plant to produce electricity per year.

- Nitrous oxide (N_2O) emissions (N2OE) (in kton): Fossil-fuel powered plants release small amounts of CH_4 and N_2O during combustion at low temperatures (less than 950°C). These plants are affected by the type of fuel and operating conditions such as the fraction of excess air. Since it, too, is a greenhouse gas, nitrous oxide significantly contributes to global greenhouse gas emissions.

- Methane (CH_4) emission (ME) (in kton): CH_4 is typically formed as a result of incomplete combustion, i.e. a lean fuel-comburent mix and early quenched oxidation reactions in the combustion process. While CH_4 emissions are relatively unusual in large, well-functioning furnaces, they are much more common in smaller-scale combustion, e.g. in heating stoves and open fireplaces. Global methane emissions are another important factor that contributes to global greenhouse gas emissions within the global challenge of climate change.

- Sulfur dioxide (SO_2) emissions (SDE) (in kton): SO_2 is produced during the burning of sulfur or materials containing sulfur. The heat generated in this process is recovered by steam generation and is subsequently converted into electricity. SO_2 emissions are a major air pollutant and a precursor to acid rain and atmospheric particulates that have a significant impact on human health.

- Nitrogen oxides (NO_x) emissions (NOxE) (in kton): NO_x is the generic term for nitrogen oxides that are usually generated in the air by the reaction between nitrogen and oxygen during the combustion process based on fossil fuels (e.g. coal) at high temperatures. NO_x emissions are a direct contributor to the greenhouse effect and, together with sulphur dioxide (SO_2), is one of the main chemicals responsible for the phenomenon known as acid rain.

In short, this study constructs innovative machine learning models for thermal power stations from the aforementioned experimental dataset. Thus, two separate hybrid MARS models have been built using the DE-based parameter optimizer (DE/MARS) for the prediction of the net annual electricity generation (NAEG) and carbon dioxide (CO_2) emissions (CDE), respectively.

Multivariate adaptive regression splines technique

In statistical machine learning, multivariate adaptive regression splines (MARS) is a regression method conceived by Friedman in 1991 that is appropriate for problems with a large number of input variables (Friedman 1991; Sekulic and Kowalski 1992; Friedman and Roosen 1995; Hastie et al. 2003; García Nieto et al. 2015). The technique uses a nonparametric approach that could be understood as a prolongation of linear models that allows tackling *interactions* among input variables and nonlinearities. Interactions, in the MARS models, are composites of the linear piece-wise functions being, piece-wise higher-order polynomial functions that are able to model more complex relationships between the variables. Moreover, MARS models tend to have a good bias-variance trade-off. The models are flexible enough to model nonlinearity and variable interactions (thus MARS models have fairly low bias), yet the constrained form of MARS basis functions prevents too much flexibility (thus MARS models have fairly low variance).

The MARS technique constructs models according to the following expansion:

$$\hat{f}(x) = \sum_{i=0}^{M} c_i B_i(x) \tag{1}$$

Therefore, this technique approximates the dependent output variable y by means of an averaged addition of $B_i(x)$ so that the coefficients c_i are constant. $B_i(x)$ can be (Friedman 1991; Sekulic and Kowalski 1992; Friedman and Roosen 1995; Hastie et al. 2003 Chou et al. 2004; Zhang et al. 2015; García Nieto et al. 2015):

- constant and equal to 1. This term is called intercept and corresponds to the term c_0;
- a hinge or hockey stick function: this function is $\max(0, \text{constant} - x)$ or $\max(0, x - \text{constant})$. The constant value is termed knot. The MARS technique chooses variables and knots values for these ones according to the procedure indicated later;
- the multiplication of hinge functions: in this case, these functions model nonlinear relationships between variables.

For instance, Fig. 2 shows a couple of splines for q=1 at the node t=3.5.

Fig. 2 An example of linear basis functions

Two steps are the base of the MARS. First, it constructs a very complex model in the forward phase and then it simplifies it in the backward stage (Chou et al. 2004; Xu et al. 2004; Freedman et al. 2007; Cheng and Cao 2014; Zhang et al. 2015):

- Forward stage: MARS starts with the intercept term, calculated averaging the values of the dependent variable. Next, MARS sums linear combinations of pairs of hinge functions with the aim of minimizing the least-squares error. These new hinge functions depend on a knot and a variable. Thus, to add new terms MARS has to try all the different combinations of variables and knots with the previous terms, called *parent terms*. Then, MARS determines the coefficients c_i using linear regression. It adds terms until a certain threshold for the residual error or a maximum number of terms is reached.

- Backward stage: the previous stage usually constructs an overfitted model. In order to construct a better model with greater generalization skill, this new stage simplifies the model removing terms, taking as a criterion the generalised cross-validation (*GCV*) criterion described below, removing first the terms that add more *GCV* to the model.

Generalised cross-validation (*GCV*) is the goodness-of-fit index utilised to assess the suitability of the terms of the model in order to prune it from the model. *GCV* does not only take into account the residual error but also how complex the model is. High values of *GCV* mean high residual error and complexity. The formula of this index is (Friedman 1991; Sekulic and Kowalski 1992; Friedman and Roosen 1995; Hastie et al. 2003; Cheng and Cao 2014; Zhang et al. 2015):

$$GCV(M) = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}_M(\mathbf{x}_i)\right)^2}{\left(1 - C(M)/n\right)^2} \tag{2}$$

where the parameter $C(M)$ grows with the number of terms in the regression function and thus, increases the value of the *GCV* index. It is given by (Friedman 1991; Sekulic and Kowalski 1992; Friedman and Roosen 1995; Hastie et al. 2003):

$$C(M) = (M+1) + d\,M \tag{3}$$

where d is a coefficient that determines the importance of this parameter and M is the number of terms in Eq. (1).

The relative importance of the independent variables that appear in the regression function (as only some of these variables remain in the final function) can be assessed using different criteria (Chou et al. 2004; Xu et al. 2004; Cheng and Cao 2014; Zhang et al. 2015): (a) the *GCV* attached to a variable can be one of these criteria and it is measured taking into account how much this index increases if this variable is erased from the final function; (b) the same criterion can be applied using the *RSS* index; and (c) another criterion is the number of subsets (Nsubset) of which the variable is a part. If it is part of more terms, its importance is greater.

Differential evolution (DE) algorithm

The differential evolution (DE) method, which was initially discovered by Storn and Price (Storn and Price 1997), optimises the problem iteratively by attempting to improve a candidate solution concerning a well-known quality measurement. The total population $\mathrm{X} = \left[\mathrm{x}_1, \mathrm{x}_2, ..., \mathrm{x}_M\right]^T$ involves M individuals in such a way that the n-th individual is represented by an objective vector designating an individual's position in the search space. The objective vector pertaining to the n-th individual at the t-ith iteration of the optimisation is expressed as (Storn and Price 1997; Price et al. 2005):

$$x_n^p(0) = L^p + \mathrm{rand}(0,1) \times \left(U^p - L^p\right) \tag{4}$$

The differential evolution (DE) technique requires five stages to be able to build the optimisation algorithm; which are indicated as follows (Storn and Price 1997; Price et al. 2005, Feoktistov 2006; Rocca et al. 2011):

- *Initialization*

Firstly, the initial objective vectors of *M* individuals are created (or produced) in the design space *H*. This task is carried out in a random way so that the initial objective vector (0-th iteration: corresponds to the initial location of the particle) of dimension *p* $\left(p \in \{1, \ldots, P\}\right)$ is given via:

$$x_n^p\left(0\right) = L^p + \mathrm{rand}\left(0,1\right) \times \left(U^p - L^p\right) \tag{5}$$

so that rand(0,1) represents a random number evenly distributed in the interval [0,1].

- *Mutation*

A benefactor vector is produced using differences of scale among individuals for each individual of the population. The *n*-th benefactor vector is generated by means of the following mutation strategy, and is written as the expression:

$$\mathbf{v}_n = \mathbf{x}_{r_1}\left(t\right) + G\left(\mathbf{x}_{r_2}\left(t\right) - \mathbf{x}_{r_3}\left(t\right)\right) \tag{6}$$

where r_1, r_2 and r_3 are random integers evenly distributed in the interval $\left[1, M\right]$ such that $r_1 \neq r_2 \neq r_3 \neq n$, and *G* is a scaling factor.

- *Crossover*

In order to ensure diversity, a crossover stage must be carried out. Indeed, in order to generate a trial vector with guarantees, we will do a crossover of the individual elements from both the objective vector and the benefactor vector. In this study, we have used the binomial crossover to create the *n*-th trial vector in the *p*-th dimension governed by the expression:

$$u_n^p = \begin{cases} v_n^p & \text{if rand}(0,1) < CR \text{ or } r_n = p \\ x_n^p(t) & \text{otherwise} \end{cases} \tag{7}$$

where $CR \in [0,1]$ is a parameter called crossover probability and r_n is a random integer spread evenly in the interval $[1, P]$. Consequently, elements of the trial vector are taken from the benefactor vector with a probability *CR* so that at least one element of the benefactor vector is accepted.

- *Selection*

Next, the trial vector is checked and the *n*-th objective vector is computed at the next iteration via:

$$\mathbf{x}_n(t+1) = \begin{cases} \mathbf{u}_n & \text{if } f(\mathbf{u}_n) \le f(\mathbf{x}_n(t)) \\ \mathbf{x}_n(t) & \text{otherwise} \end{cases} \tag{8}$$

Therefore, the objective vector is replaced by the trial vector if its performance is greater than or equal to the performance of the objective vector.

- *Stopping criterion*

The algorithm is stopped if the permitted maximum number of function evaluations is reached and after all *M* objective vectors have been upgraded. Otherwise, the above steps, from the second one to the fifth, are repeated.

The pseudocode of the DE algorithm can be written as:

Random initialization of the individuals and calculate the objective

while Current_number_of_function_evaluations < Max_function_evaluations **do**

for n = 1:M **do**

Carry out the mutation according to Eq. (6)

Carry out the binomial crossover according to Eq. (7)

Calculate the objective taking into account the constraints of the trial vector

end for

for n = 1:M **do**

Upgrade the *n*-th objective vector according to Eq. (8)

end for

end while

Accuracy measure

Tables 1 and 2 show the independent economic and environmental variables used in this work, respectively (Tajbakhsh and Hassini 2018; CATF 2918; EIA 2018; USEPA 2018; FERC 2018). Thus, there were a total of 11 predictors in both the DE/MARS and M5 models. The units of the obtained variables (net annual electricity generation, and carbon dioxide (CO_2) emissions) are TWh and Mton (Tajbakhsh and Hassini 2018; CATF 2918; EIA 2018; USEPA 2018; FERC 2018), respectively.

Table 1 Economic parameters used in this work, with their mean and standard deviation

Table 2 Environmental variables in this work to investigate the power plant case study with their mean and standard deviation

The goodness-of-fit criterion applied in this article to foretell the output variables from the variables left over was the determination coefficient R^2, which is a ratio that shows the relationship between the variation in the predicted variable defined by the analysis model and the behaviour in the same variable across the data set. Let's name the observed values t_i and the estimated value of the respective model y_i. These variabilities can be evaluated with the following expressions (Wasserman 2003; Freedman et al. 2007):

- $SS_{err} = \sum_{i=1}^{n} \left(t_i - y_i\right)^2$
- $SS_{tot} = \sum_{i=1}^{n} \left(t_i - \bar{t}\right)^2$

where $\bar{t}$ is the average value of the observed n samples:

$$\bar{t} = \frac{1}{n}\sum_{i=1}^{n} t_i \tag{9}$$

Then, the coefficient of determination is:

$$R^2 \equiv 1 - \frac{SS_{err}}{SS_{tot}} \tag{10}$$

The coefficient of determination defines the suitability of the regression values come close to the actual values (i.e. the closer to 1 the better).

Furthermore, the MARS method depends on the following hyperparameters (Friedman and Roosen 1995; Xu et al. 2004; Vidoli 2011):

- Interactions: It describes a maximum number of basis functions included in a situation where the effect of one causal variable on an outcome depends on the state of another causal variable by basis functions, multiplying in a term.
- Maximum number of base functions (Maxfuncs): it defines the number of base functions used before the previous phase. However, some of them are filtered.
- Penalty parameter (*d*): it is considered for each node there is one GCV penalty related to that knot.

Thus, the novel hybrid DE/MARS and M5 models have been constructed with predicted variables the net annual electricity generation and carbon dioxide (CO_2) emissions and the other eleven economic and performance parameters as predictor variables (input variables). The coefficient of determination R^2 is used as a criterion to assess the success of each model.

As mentioned previously, the success of the methodology is largely due to the MARS model parameters, which makes the right fit to be absolutely essential. Therefore, each set of parameters in relation with a model must be evaluated so that they can be compared, which takes up the greater part of the computer processing time. Accordingly, the fewer models are evaluated in the process, the better, as fewer model evaluations are required. Consequently, the hyperparameters have been optimized by divers models (Friedman and Roosen 1995; Xu et al. 2004; Vidoli 2011) and the DE optimisation technique was found to be the most appropriate, effective and simple tool in order to estimate the net annual electricity generation and carbon dioxide (CO_2) emissions from the other eleven economic and performance parameters (input variables). Fig. 3 shows the flowchart of the DE/MARS method used in this work.

Fig. 3 DE/MARS method flowchart

Furthermore, cross–validation was used for computing the coefficient of determination (R^2) (Picard and Cook 1984; Wasserman 2003; Freedman et al. 2007). Thus, a ten–fold cross–validation algorithm was used to estimate the predictive capacity of the DE/MARS model.

The DE module for the validation process was used to optimise the MARS parameters. The validation process seeks the best parameters for the data, using the mean value of the coefficient of determination, making R^2 the objective function in this condition.

Initially, sets of three parameters (30 different, random ones with members of the population) were calculated within the search space (Table 3). A model was defined with each of these parameter sets, and the corresponding R^2 cross-validation was estimated. As DE looks for a minimum and we are trying to obtain the highest R^2, the fitness function associated to a given model (set of three parameters) was minus R^2 because the minimum of this function would be the maximum of R^2. A new population was created by using mutation and recombination; once again, the suitability of each model was estimated. The aptitude of the models was selected, each new member of the population was compared with his previous partner, and the worst was discarded. Subsequently, a new population of 30 members was created containing a combination of the new and old populations. The steps of mutation, recombination and selection were repeated until the stop criterion was reached. Finally, if the best fitness of the last iteration differs less than 10^{-8} with the best fitness of the previous one, with a condition of allowing for a maximum of 200 iterations, the process stops. The MARS regression was performed with the Earth library (Milborrow 2014) along with the DE function from DEOptim library (Mullen et al. 2011; Ardia et al. 2016) from the R project. The model with the best fitness in the last iteration was then selected to define the result.

Table 3 Search for space in the DE tuning process for each of the MARS parameters

Results and discussion

Effective conservation efforts and the transition to an energy future are both essential to avoid catastrophic climate change (Vidadili et al. 2017). Fossil fuel-based thermal power generation is one of the earliest forms of large-scale energy generation. Over consumption of electricity calls for an urgent need to improve the efficiency of fossil-fuel thermal power generation in order to save fossil fuel resources while minimizing contaminants. Energy losses from energy conversion and distribution in the energy supply network are very significant. Energy efficiency enables the reduction of resources used for energy extraction, transformation, transportation and use. In this regard, active actions in the energy and environmental policy based on strategies about economic development, environmental protection and social development are important (Rehbein et al. 2020). In order to make these kinds of changes,it is needes to have the right policies, regulations and energy performance analysis of the energy systems. These will target on overall energy policy, demand and supply-side measures, energy tariff regulations, power sector reform and promotion of energy efficiency auditing.

The optimal parameters for the obtained models (DE/MARS) by means of the differential evolution (DE) technique for the net annual electricity generation (NAEG) and carbon dioxide (CO2) emission (CDE) are shown in Table 4. In addition to that, Tables 5 and 6 show these parameters to define the final models. Particularly, they define a list of nine and seven main functions for the two adjusted DE/MARS models, just as their coefficients for the dependent variables net annual electricity generation and carbon dioxide emissions, respectively. Observe that $h(x)=x$ if $x>0$ and $h(x)=0$ if $x \geq 0$. Moreover, a graphical representation of the two best DE/MARS models for the net annual electricity generation and carbon dioxide is shown in Figs. 4 and 5, respectively.

Fig. 4 Representation of the partial relations within the MARS model for the NAEG according to its independent variables: **a** first-order relation with AFC; **b** first-order relation with CDE; **c** first-order relation with AR; **d** second-order relation with AP and AR; **e** second-order relation with ANE and CDE; and **f** second-order relation with ANE and AR

These graphs are an idea of which predictors in the MARS equation have the greatest effect on the predicted value when considering other predictor variables in their median values. Fig. 4 shows how variable AENG changes according to the model when all the variables but one (first-order relation) or but two (second-order relation) are fixed at its median value. A sharp change in the gradient of the straight line (first-order term) or curvature of the surface (second-order term) is indicative of different influence. Accordingly, Fig. 4a shows the dependent variable NAEG (y-axis) as a function of the annual fuel consumed (AFC) (x-axis) considering the other input variables to be constant. It should be noted that the NAEG increases very slightly and then peaks at 4 TWh at about 2000 MJ. Next, NAEG stabilizes at 4TWh, that is, its value remains constant for all AFC values. Fig. 4b shows the NAEG (*y*-axis) as a function of the carbon dioxide (CO_2) emissions (CDE) (*x*-axis) so that NAEG remains constant up to a CDE value of 2.5 Mton and then NAEG grows following a straight line reaching a maximum value of 10 TWh for a CDE value of 20 Mton. Fig. 4c shows the line chart corresponding to NAEG (*y*-axis) as a function of AR (*x*-axis) follows a straight line. From AR 0 to 250 million dollars, NAEG increases slightly and then, from an AR value 250 to 1,600 million dollars, NAEG increases gradually with a maximum of 7.5 TWh at AR 1,600 million dollars. Similarly, the surface chart (see Fig. 4e) shows that NAEG rises dramatically and peaked as ANE and CDE increase. The combined effect of variables ANE and CDE have quite an impact on the dependent variables that is greater than the one obtained separately.

Fig. 5 Description of the partial relations within the MARS model for the CDE with respect to independent variables. In this case, all of the partial relationships are first-order terms

Fig. 5 shows only first-order relationships as the DE-MARS model does not have second-order basis functions. In this case, a simple model with only three independent variables, N2OE, ME, and AENG, was created. The variable CDE increases when any of these independent variables increase. Fig. 5a indicates the CDE as a function of the nitrous oxide (N_2O) emissions (N2OE), keeping the remaining independent variables as constants. It should be observed that CDE increases in a straight line from 0.5 Mton at 0 kton to 12 Mton at 100 kton. In this case, all the partial relationships are first-order terms: there are no second-order terms. Fig. 5b represents the dependent variable CDE (y-axis) as a function of the methane (CH_4) emissions (kton) (ME) (x-axis). From ME value 0 to 20 kton, CDE decreases slightly reaching a minimum of 2 Mton. From ME value 20 to 65 kton, CDE increases substantially in a straight line with a maximum of 6 Mton. Finally, Fig. 5c shows the CDE dependent variable as a function of the NAEG variable. From NAEG value 0 to 7 TWh, CDE grows substantially following a straight line up to a CDE value of 4.5 Mton. From NAEG value 7 to 17 TWh, CDE increases gradually in a straight line up to reach a maximum value of 6.5 Mton.

Table 4 Optimal MARS model hyperparameters found with DE

Table 5 DE/MARS model basis functions and their coefficients c_i for the Net annual electricity generation (NAEG)

Table 6 DE/MARS model basis functions and their coefficients c_i for the carbon dioxide (CO_2) emission (CDE)

The correlation and determination coefficients for the DE/MARS models based on the data are shown in Table 7.

Table 7 Cross-validation coefficients of determination (R^2) and correlation (r) for DE/MARS models used in this analysis to determine the NAEG and CDE

Consequently, the MARS technique in combination with DE optimisation proved to be the best model for estimating NAEG and CDE, since the two applied DE/MARS models have coefficients of determination R^2 equal to 0.9803 and 0.9901, and correlation coefficients equal to 0.9895 and 0.9947, respectively. Therefore, the results demonstrate a reliable goodness-of-fit and show a good level of agreement between our predicted, and observed data.

Additionally, the significance ranking of the economic and performance parameters (input variables), and the NAEG and CDE as dependent variables (output variables) are presented. Tables 8 and 9, and Figs. 6 and 7 show these variables, respectively. Figs. 8 and 9 show that both AENG and CDE DE/MARS models can predict with great accuracy the value of the dependent variable, as the observed values (blue line) are very close to the predicted values (red line) and there exist only small discrepancies between the two values. That means that the chosen independent variables provide enough information to ascertain the variations of the dependent variables that the DE/MARS model is able to replicate.

Table 8 Results of the variables in relation with each variable to define their significance in the DE/MARS NAEG model

Table 9 Results of the variables in relation with each variable to define their significance in the DE/MARS CDE model

Fig. 6 Level of importance of the parameters in the DE/MARS NAEG model in accordance with Nsubsets criterion

Fig. 7 Level of importance of the parameters in the DE/MARS CDE model in accordance with Nsubsets criterion

Consequently, *Annual revenue* is the most representative variable in predicting NAEG for the MARS model, followed by carbon dioxide (CO_2) emissions, annual number of employees, annual production expenses and book value of plant and land.

The annual revenue indicates the direct cost of the net annual electricity generation, which makes it the major parameter in the model. The second most important factor is Carbon dioxide emissions, as it is a direct effect of combustion. Finally, the annual number of employees and annual production expenses are of less importance as they are related to the NAEG only in an indirect way, through the operation management of a power plant. The book value of the plant is related with the plant size, which involves different aspects of the conversion technology, not only the power capacity.

Similarly, the relative level of importance of the independent variables in the Carbon dioxide (CO_2) emission model is shown in Fig. 7 and Table 9. In this case, this second MARS model defines the emission of nitrous oxide as the first variable (N_2O), followed by net annual electricity generation and, finally, Methane (CH_4) emission.

In this case, the GHG emissions are related with the combustion process. The main factor is Nitrous oxide which, together with Carbon dioxide, depends on proper operating conditions in the boiler. Ultimately, the greatest concern surrounding the net annual electricity generation from fossil fuels is the level of carbon dioxide emissions. Finally, Methane emissions have a lesser effect on the model as a result of its residual participation in the combustion process or energy management capacity.

Thus, the present research allows for the prediction of the first dependent variable (net annual electricity generation (NAEG)) with results that correlate with the actual observed values. Fig. 8 compares the observed NAEG with that which was predicted using this hybrid DE/MARS model (see Fig. 8).

Fig. 8 Experimental and foretold NAEG values from the cross-breed DE/MARS model (R^2 = 0.9803)

Similarly, Fig. 9 compares observed and predicted Carbon dioxide (CO_2) emission values by the DE/MARS model.

Fig. 9 Experimental and foretold Carbon dioxide (CO_2) emission values with the hybrid DE/MARS–based model (R^2 = 0.9895)

Conclusions

Comparing power generation and environmental assets is a very challenging problem due to differences in technology, operation and size. Energy efficiency and fossil-fuel thermal power station are becoming an increasingly important part of the changing energy system in the sustainable energy context, a fact that raises challenges for the current technologies. Investments in energy efficiency and environmental protection can reduce electricity demand and allow the early decommissioning of old remaining coal and fossil fuel plants.

Computational fluid dynamics (CFD) techniques represented a very important advance, modelling in detail the energy and environmental parameters management (Díez et al. 2005; Sankar et al. 2019). However, CFD techniques maintained important shortcomings: a huge computing time and modelling complexity. In this sense, the overall performance in thermal power plants developed in this work was accurately predicted and modelled by using this new hybrid DE/MARS model with success. In addition, this newer, less-expensive method will provide an excellent alternative to the more expensive traditional methods. Furthermore, MARS models produce, through an expansion of functions (known as hinge functions and products of two or more hinge functions), an explicit mathematical formula. This formula shows the dependent variable as a function of the independent variables. This last feature is a fundamental difference compared to alternative automatic learning methods since most of them behave like a black box.

A high result of R^2 ($R^2 = 0.9803$) was obtained when the hybrid DE/MARS model was tested with the experimental data from the net annual electricity generation (see Fig. 8). Similarly, the model for the experimental dataset of the Carbon dioxide (CO_2) emissions also obtained a high coefficient of determination ($R^2 = 0.9895$), and the predicted results agreed with the observed Carbon dioxide (CO_2) emission dataset values (see Fig. 9).

This method also allows the ranking of the input variables involved in predicting the energy performance in thermal power stations. Thus, the annual revenue is the most influential parameter in the NAEG model, whilst the Nitrous oxide (N_2O) emission is the most relevant to CDE concentration. Also, in the case of the CDE model, a particularly simple and highly effective model was created: only three variables remain and the model is piece-wise linear. The AENG model is more complex but it is still quite simple, reducing the initial number of variables from ten to six. Both NAEG and CDE show their dependence from the input variables. The energy analysis, in particular, can improve the efficiency of fossil-fuel thermal power by either energy conservation within the system or through the differences between initial parameters and final performance parameters. NAEG is directly related to both the total gross power generation as well as the amount of electricity generated by a power plant. It depends on AR, CDE and ANE, parameters dependent on the power plant size and operation conditions.

Finally, the hybrid DE/MARS regression method seems to considerably improve the predictive capacity obtained by the MARS-based regressor, with no need to optimise its parameters.

However, the model as DE/MARS is a data-based model, it is relevant to be careful not to extrapolate the results for different situations. The DE/MARS model depends strictly on the results provided from the available data and is conceptually different from a deterministic mathematical model based on a set of integral or differential equations.

In conclusion, this highly effective new DE/MARS model should prove to be a valuable tool for estimating and predicting energy performance in thermal power stations.

Acknowledgements

The authors wish to thank the United States Environmental Protection Agency (USEPA), Clean Air Task Force (CATF), Energy Information Administration (EIA) and Federal

Energy Regulatory Commission (FERC) for providing the experimental dataset used in this research. Moreover, this research was funded by the Foundation for the Promotion of Applied Scientific Research and Technology in Asturias (FICYT) through the GRUPIN project Reference IDI/2018/000221, co-financed with EU FEDER funds.

References

Al-Musaylh MS, Deo RC, Adamowski JF, Li Y (2018) Short-term electricity demand forecasting with MARS, SVR and ARIMA models using aggregated demand data in Queensland, Australia. Adv Eng Inform 35:1–16

Ardia D, Mullen KM, Peterson BG, Ulrich J (2016) *DEoptim*: Differential evolution in R, version 2.2-4

Behera MK, Majumder I, Nayak N (2018) Solar photovoltaic power forecasting using optimized modified extreme learning machine technique. Eng Sci Tech, Int J 21(3):428–438

Chen X, Yang H, Sun K (2017) Developing a meta-model for sensitivity analyses and prediction of building performance for passively designed high-rise residential buildings. Appl Energ 194:422–439

Cheng MY, Cao MT (2014) Accurately predicting building energy performance using evolutionary multivariate adaptive regression splines. Appl Soft Comput 22:178–188

Chou S–M, Lee S–M, Shao YE, Chen I–F (2004) Mining the breast cancer pattern using artificial neural networks and multivariate adaptive regression splines. Expert Syst Appl 27(1):133–142

Clean Air Task Force (CATF) (2018) Database, https://www.catf.us/work/power-plants/

Clerc M (2006) Particle swarm optimization. Wiley-ISTE, London (UK)

Díez LI, Cortés C, Campo A (2005) Modelling of pulverized coal boilers: review and validation of on-line simulation techniques. Appl Thermal Eng 25(10):1516–1533

Dorigo M, Stützle T (2004) Ant colony optimization. Bradford Publisher, The MIT Press, Cambridge, Massachusetts (USA)

Eberhart RC, Shi Y, Kennedy J (2001) Swarm intelligence. Morgan Kaufmann, San Francisco

Energy Information Administration (EIA) (2018) Database, https://www.eia.gov/electricity/data.php#gencapacity

Federal Energy Regulatory Commission (FERC) (2018) Database, https://www.ferc.gov/industries/

Feoktistov V (2006) Differential evolution: in search of solutions. Springer, New York

Fister I, Stranad D, Yang X–S, Fister Jr. I (2015) Adaptation and hybridization in nature-inspired algorithms. In: Fister I, Fister Jr I (eds) Adaptation and hybridization in computational intelligence, Springer, New York, vol 18, pp 3–50

Freedman D, Pisani R, Purves R (2007) Statistics. W.W. Norton & Company, New York

Friedman JH (1991) Multivariate adaptive regression splines. Ann Stat 19:1–141

Friedman JH, Roosen CB (1995) An introduction tomultivariate adaptive regression splines. Stat Methods Med Res 4:197–217

García Nieto PJ, Álvarez Antón JC (2014) Nonlinear air quality modeling using multivariate adaptive regression splines in Gijón urban area (Northern Spain) at local scale. Appl Math Comput 235:50–65

García Nieto PJ, García–Gonzalo E, Alonso Fernández JR, Díaz Muñiz C (2015) Hybrid PSO–MARS–based model for forecasting a successful growth cycle of the Spirulina platensis from experimental data in open raceway ponds. Ecol Eng 81:534–542

Hastie T, Tibshirani R, Friedman J (2003) The elements of statistical learning. Springer, New York

Jayawardene I, Venayagamoorthy GK (2015) Reservoir based learning network for control of two-area power system with variable renewable generation. Neurocomputing 170:428–438.

Kisi O, Parmar KS, Soni K, Demir V (2017) Modeling of air pollutants using least square support vector regression, multivariate adaptive regression spline, and M5 model tree models. Air Qual Atmos Hlth 10(7):873–883

Li Y, He Y, Su Y, Shu L (2016) Forecasting the daily power output of a grid-connected photovoltaic system based on multivariate adaptive regression splines. Appl Energ 180:392–401

Li MJ, Tao WQ (2017) Review of methodologies and polices for evaluation of energy efficiency in high energy-consuming industry. Appl Energ 187:203–215

Mardani A, Zavadskas EK, Streimikiene D, Jusoh A, Khoshnoudi M (2017) A comprehensive review of data envelopment analysis (DEA) approach in energy efficiency. Renew Sustain Energy Rev 70:1298–1322

Milborrow S (2014) Earth: Multivariate adaptive regression spline models. R Package, version 3.2–7

Mullen KM, Ardia D, Gil D, Windover D, Cline J (2011) *DEoptim*: An R package for global optimization by differential evolution. J Stat Softw 40(6):1–26. URL http://www.jstatsoft.org/v40/i06/

Olsson AE (2011) Particle swarm optimization: theory, techniques and applications. Nova Science Publishers, New York

Paredes-Sánchez JP, Gutiérrez-Trashorras AJ, Xiberta-Bernat J (2014) Energy potential of residue from wood transformation industry in the central metropolitan area of the Principality of Asturias (northwest Spain). Waste Manage Res 32(3):241–244

Paredes-Sánchez JP, Gutiérrez-Trashorras AJ, Xiberta-Bernat J (2015) Wood residue to energy from forests in the Central Metropolitan Area of Asturias (NW Spain). Urban for Urban Gree 14(2):195–199

Paredes-Sánchez JP, López-Ochoa LM, López-González LM, Xiberta-Bernat J (2016) Bioenergy for District Bioheating System (DBS) from eucalyptus residues in a European coal-producing region. Energy Convers Manage 26: 960–970

Paredes-Sánchez JP, López-Ochoa LM, López-González LM, Las-Heras-Casas J, Xiberta-Bernat J (2019). Evolution and perspectives of the bioenergy applications in Spain. J Clean Prod 213:553–568

Peer RA, Sanders KT (2018) The water consequences of a transitioning US power sector. Appl Energ 210:613–622

Picard R, Cook D (1984) Cross-validation of regression models. J Am Stat Assoc 79(387):575–583

Price K, Storn RM, Lampinen JA (2005) Differential evolution: a practical approach to global optimization. Springer, Berlin

Rehbein JA, Watson JE, Lane JL, Sonter LJ, Venter O, Atkinson SC, Allan JR. (2020) Renewable energy development threatens many globally important biodiversity areas. Glob Change Biol 26(5):3040 –3051

Rocca P, Oliveri G, Massa A (2011) Differential evolution as applied to electromagnetics. IEEE Antennas Propag 53(1):38–49

Sankar G, Kumar DS, Balasubramanian, KR (2019) Computational modeling of pulverized coal fired boilers–A review on the current position. Fuel 236:643–665

Sekhar Roy S, Roy R, Balas VE (2018) Estimating heating load in buildings using multivariate adaptive regression splines, extreme learning machine, a hybrid model of MARS and ELM. Renew Sust Energ Rev 82(3):4256–4268

Sekulic SS, Kowalski BR (1992) MARS: A tutorial. J Chemometr 6:199–216.

Simon D (2013) Evolutionary optimization algorithms. Wiley, New York

Sobri S, Koohi-Kamali S, Rahim NA (2018) Solar photovoltaic generation forecasting methods: A review. Energy Convers Manage 156:459–497.

Storn R, Price K (1997) Differential evolution - a simple and efficient heuristic for global optimization over continuous spaces. J Global Optim 11:341–359.

Sueyoshi T, Yuan Y, Goto M (2017) A literature study for DEA applied to energy and environment. Energy Econ 62:104–124

Tajbakhsh A, Hassini E (2018) Evaluating sustainability performance in fossil-fuel power plants using a two-stage data envelopment analysis. Energ Econ 74:154–178

United States Environmental Protection Agency (USEPA) (2018) Database, https://www.epa.gov/research/methods-models-tools-and-databases

Vidadili N, Suleymanov E, Bulut C, Mahmudlu C (2017) Transition to renewable energy and sustainable energy development in Azerbaijan. Renew Sustain Energy Rev 80:1153 –1161

Vidoli F (2011) Evaluating the water sector in Italy through a two stage method using the conditional robust nonparametric frontier and multivariate adaptive regression splines. Eur J Oper Res 212(13):583–595

Wang G, Deng J, Zhang Y, Zhang Q, Duan L, Hao J, Jiang J (2020) Air pollutant emissions from coal-fired power plants in China over the past two decades. Sci Total Environ 741:140326

Wasserman L (2003) All of statistics: a concise course in statistical inference. Springer, New York

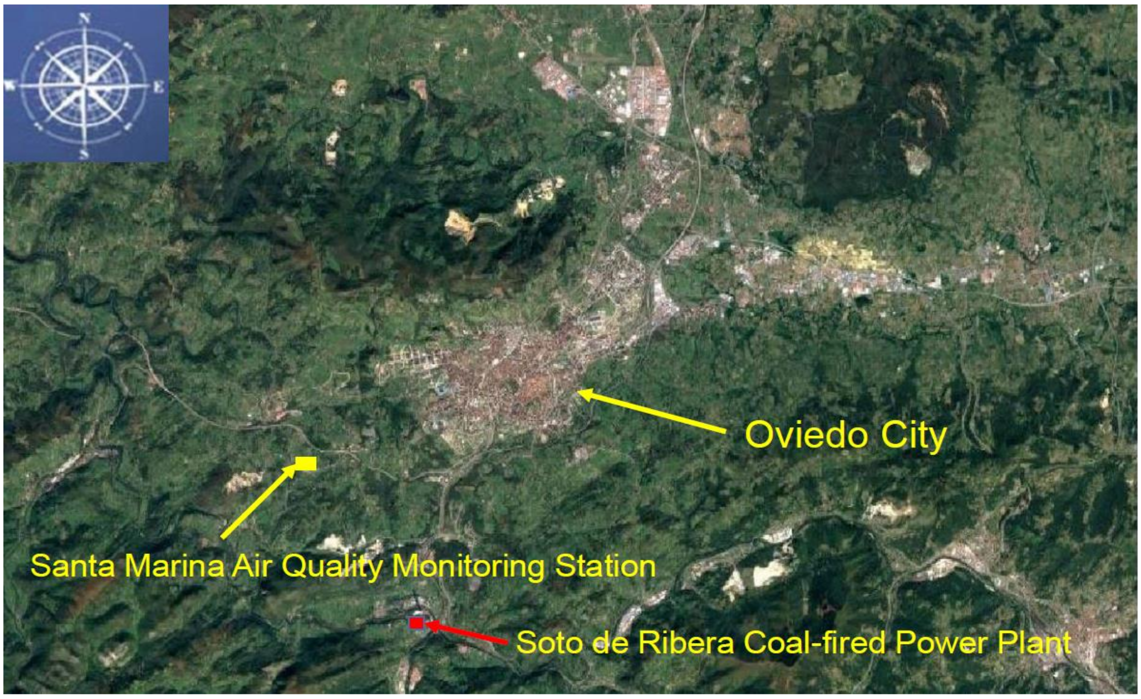
Xu QS, Daszykowski M, Walczak B, Daeyaert F, De Jonge MR, Heeres J, Koymans LMH, Lewi PJ, Vinkers HM, Janssen PA, Massart DL (2004) Multivariate adaptive regression splines—studies of HIV reverse transcriptase inhibitors. Chemometr Intell Lab 72(1):27–34

Yang X–S, Cui Z, Xiao R, Gandomi AH, Karamanoglu M (2013) Swarm intelligence and bio-inspired computation: theory and applications. Elsevier, London

Zhang WG, Goh ATC (2013) Multivariate adaptive regression splines for analysis of geotechnical engineering systems. Comput Geotech 48:82–95

Zhang W, Goh ATC, Zhang Y, Chen Y, Xiao Y (2015) Assessment of soil liquefaction based on capacity energy concept and multivariate adaptive regression splines. Eng Geol 188:29–37

Zhao M, Minett AI, Harris AT (2013) A review of techno-economic models for the retrofitting of conventional pulverised-coal power plants for post-combustion capture (PCC) of CO_2. Energ Environ Sci 6(1):25–40



(a)

(b)

Fig. 1 a An aerial photograph of an example of fossil-fuel power plant (i.e., Soto de Ribera power plant) and **b** the same fossil-fuel power plant at a larger scale

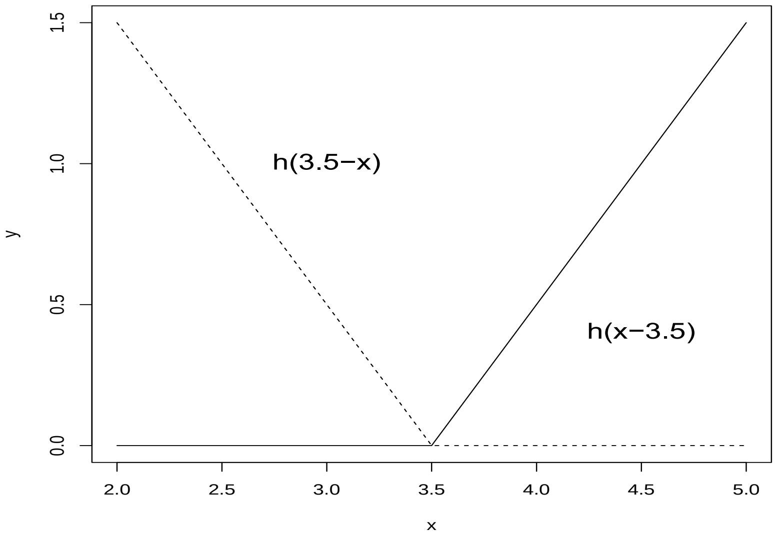


Fig. 2 An example of linear basis functions

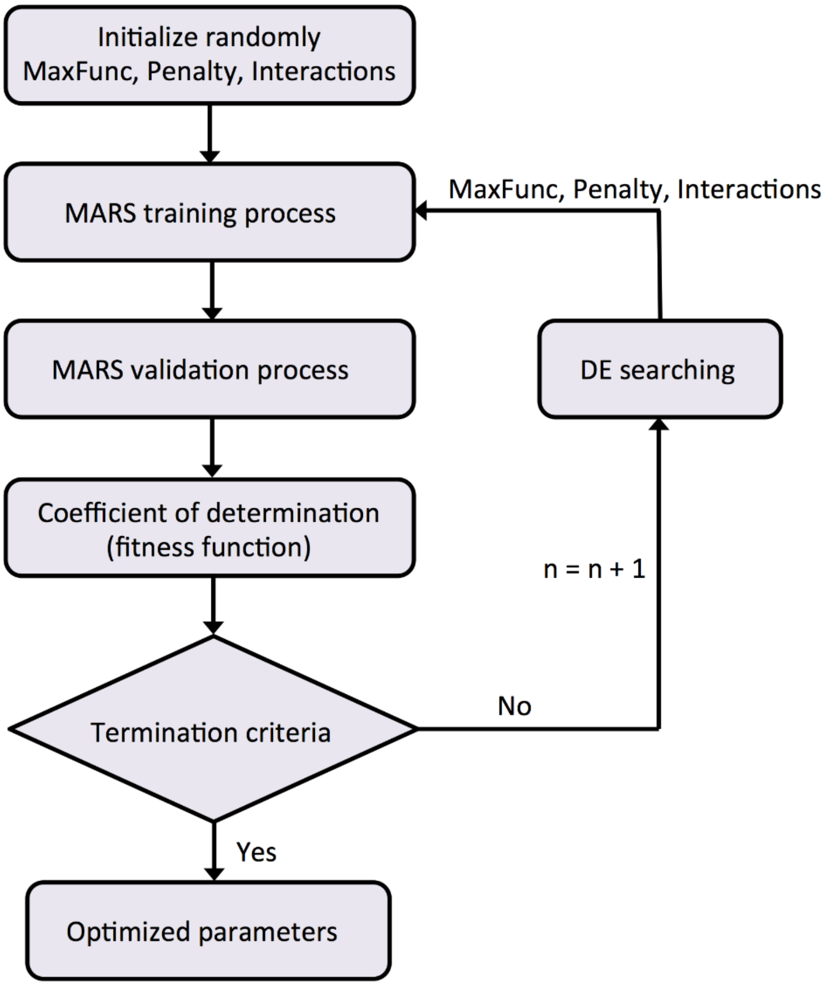


Fig. 3 DE/MARS method flowchart

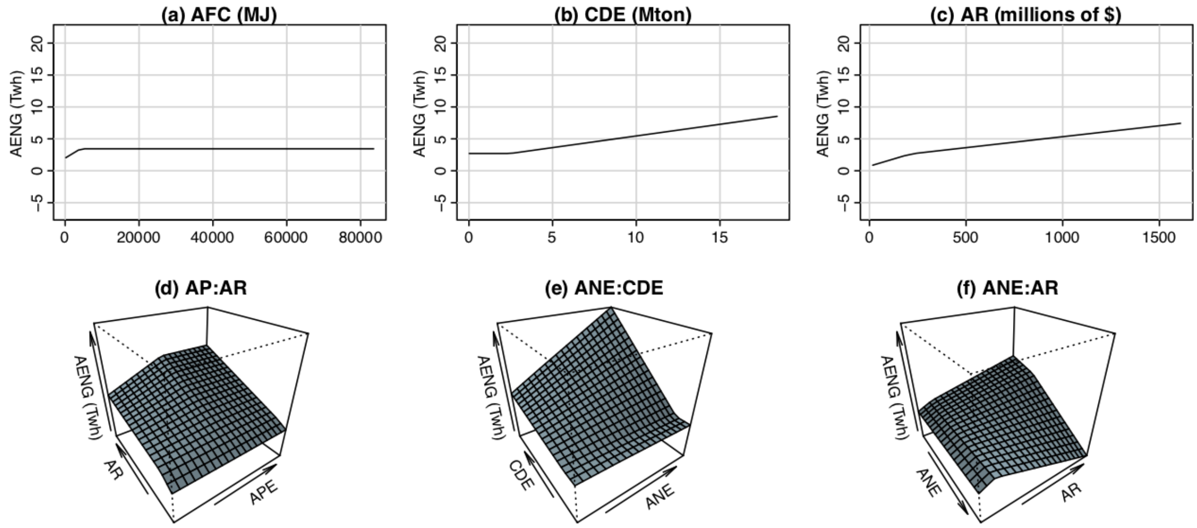


Fig. 4 Representation of the partial relations within the MARS model for the NAEG according to its independent variables: **a** first-order relation with AFC; **b** first-order relation with CDE; **c** first-order relation with AR; **d** second-order relation with AP and AR; **e** second-order relation with ANE and CDE; and **f** second-order relation with ANE and AR

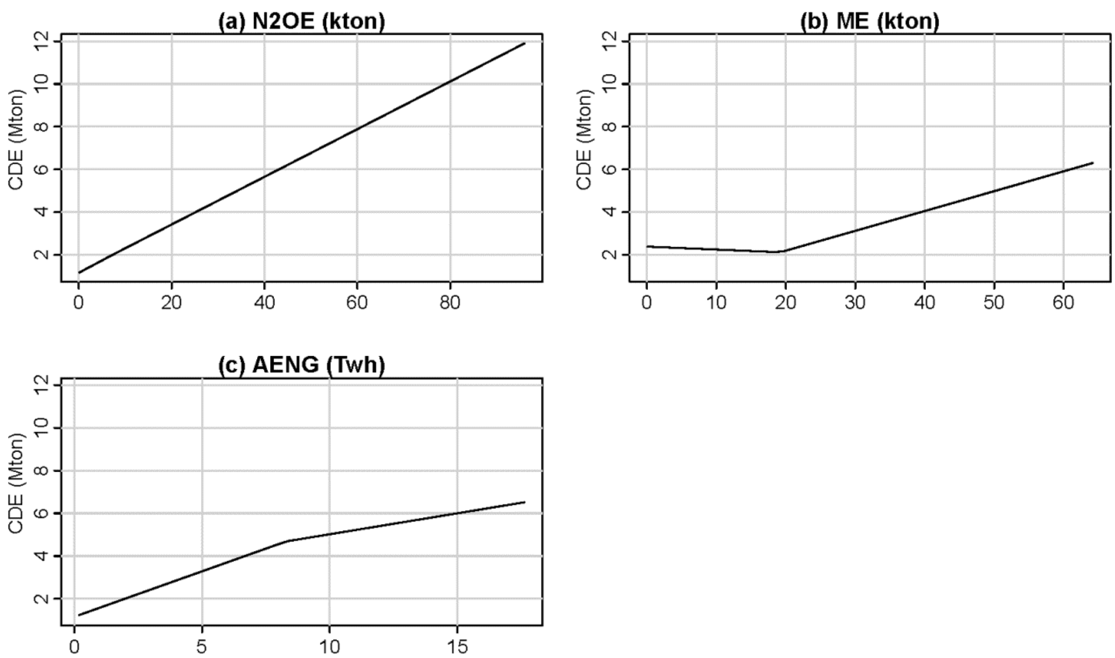


Fig. 5 Description of the partial relations within the MARS model for the CDE with respect to independent variables. In this case, all of the partial relationships are first-order terms

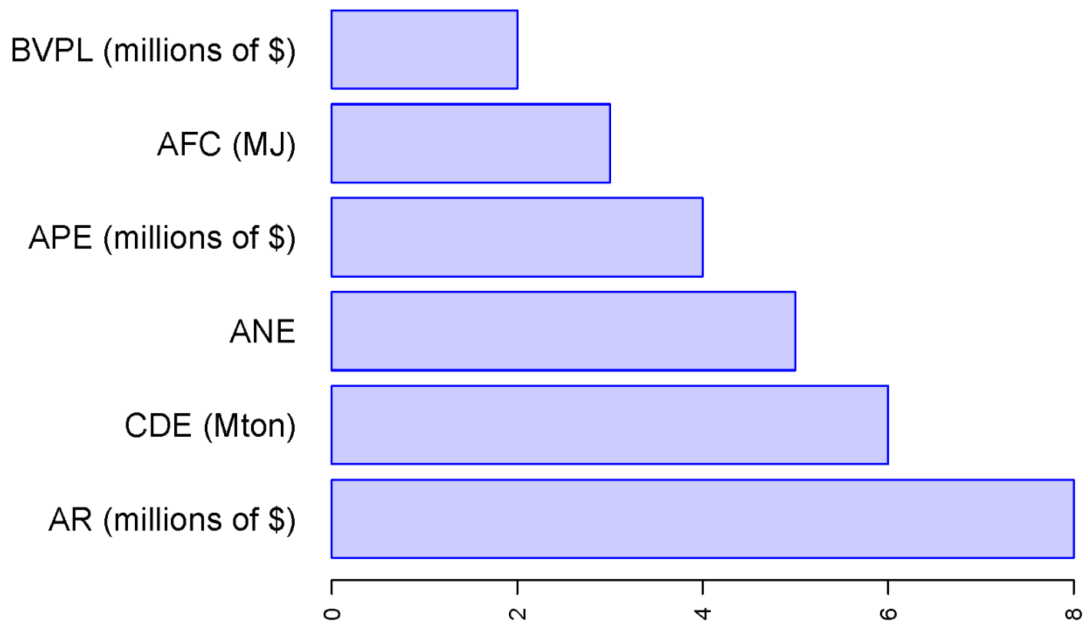


Fig. 6 Level of importance of the parameters in the DE/MARS NAEG model in accordance with Nsubsets criterion

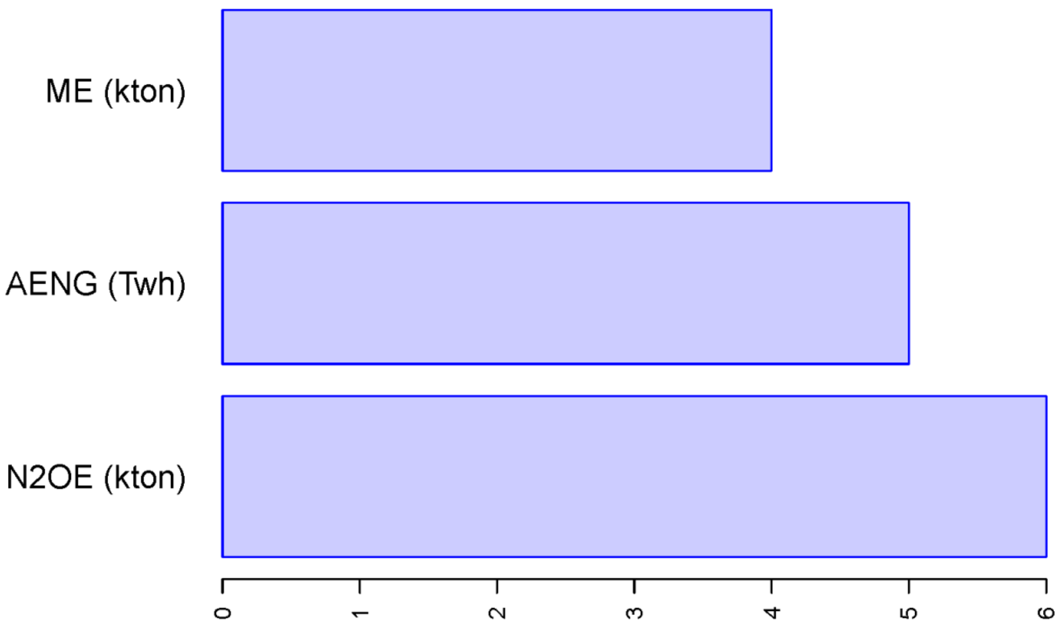


Fig. 7 Level of importance of the parameters in the DE/MARS CDE model in accordance with Nsubsets criterion

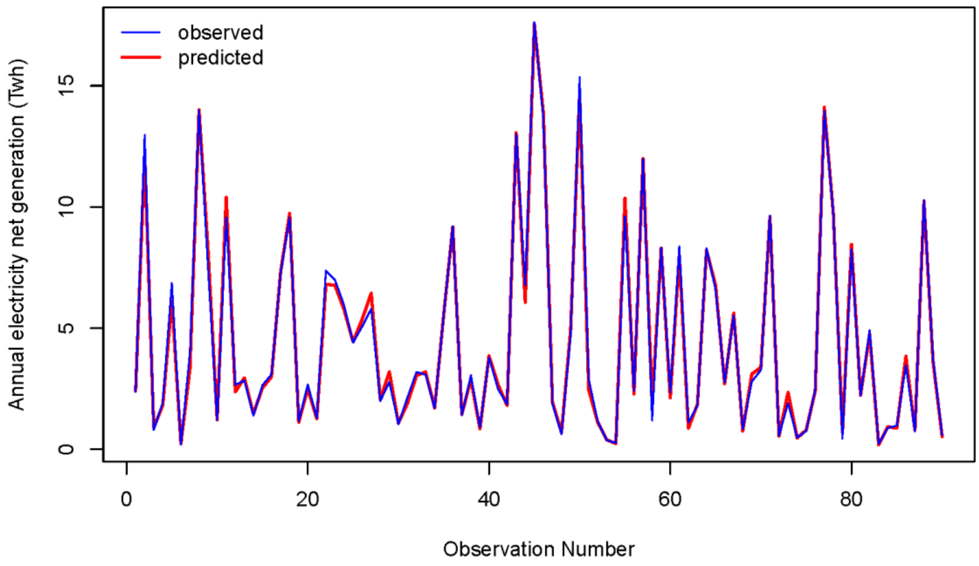


Fig. 8 Experimental and foretold NAEG values from the cross-breed DE/MARS model (R^2=0.9803)

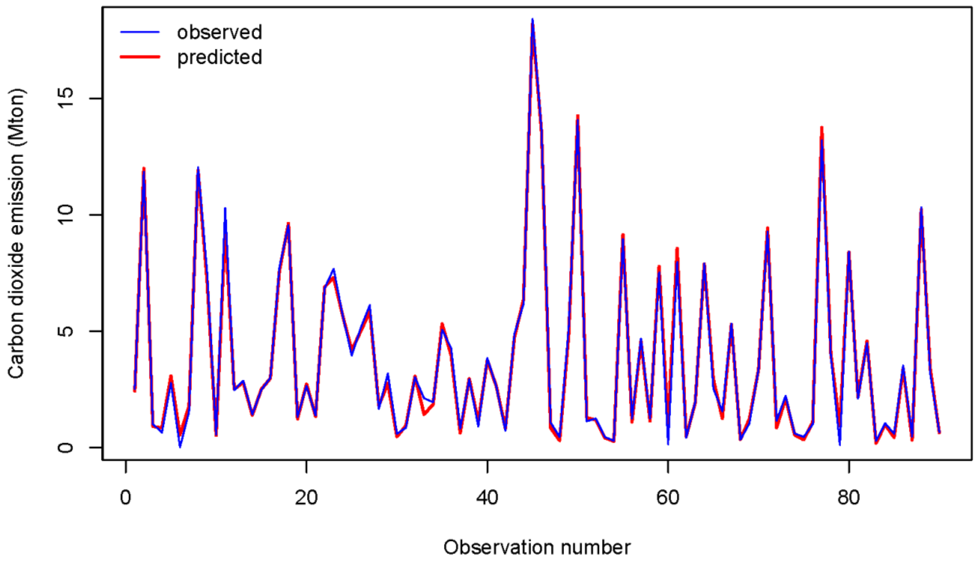


Fig. 9 Experimental and foretold Carbon dioxide (CO_2) emission values with the hybrid DE/MARS–based model (R^2=0.9895)

Table 1 Economic parameters used in this work, with their mean and standard deviation

Economic variables	Short name of the variable	Mean	Std
Book value of plant and land (millions of $)	BVPL	729.67	699.79
Annual production expenses (millions of $)	APE	164.07	152.23
Plant nameplate capacity (Mw)	PNC	1002.4	808.62
Annual number of employees	ANE	125.43	106.42
Annual revenue (millions of $)	AR	410.48	378.76
Net annual electricity generation (TWh)	NAEG	4.5214	4.1168

Table 2 Environmental variables in this work to investigate the power plant case study with their mean and standard deviation

Environmental variables	Short name of the variable	Mean	Std
Annual fuel consumed (MJ)	AFC	4038	8904.62
Carbon dioxide (CO_2) emissions (Mtons)	CDE	3.8511	3.85255
Nitrous oxide (N_2O) emissions (kton)	N2OE	17.798	20.699
Methane (CH_4) emissions (kton)	ME	8.0223	11.4688
Sulfur dioxide (SO_2) emissions (kton)	SDE	6.0031	9.5171
Nitrogen oxide (NO_x) emissions (kton)	NOxE	3.3789	3.48999

Table 3 Search for space in the DE tuning process for each of the MARS parameters

MARS hyperparameters	Lower limit	Upper limit
Maximum number of basis functions (MaxFuncs)	3	200
Interactions	1	6
Penalty parameter (d)	-1	4

Table 4 Optimal MARS model hyperparameters found with DE

Hyperparameter	Optimal value for AENG	Optimal value for CDE
Interactions	5	2
MaxFuncs	33	124
Penalty	2	0

Table 5 DE/MARS model basis functions and their coefficients c_i for the Net annual electricity generation (NAEG)

B_i	Definition	c_i
B_1	1	3.03492
B_2	h(4176–AFC)	–0.00035
B_3	h(CDE–2.47926)	0.15613
B_4	h(179.192–AR)	–0.00902
B_5	h(AR–179.192)	0.00592
B_6	ANE × h(CDE–2.47926)	0.00239
B_7	h(332.835–APE) × h(AR–238.603)	–0.000014
B_8	h(ANE–143) × h(AR–179.192)	–0.00003
B_9	BVPL × h(143–ANE) × h(AR–179.192)	0.00000002

Table 6 DE/MARS model basis functions and their coefficients c_i for the Carbon dioxide (CO_2) emission (CDE)

B_i	Definition	c_i
B_1	1	4.12415
B_2	h(7.464–N2OE)	–0.11589
B_3	h(N2OE–7.464)	0.11160
B_4	h(19.205–ME)	0.01381
B_5	h(ME–19.205)	0.09285
B_6	h(8.29257–AENG)	–0.42542
B_7	h(AENG–8.29257)	0.19642

Table 7 Cross-validation coefficients of determination (R^2) and correlation (r) for DE/MARS models used in this analysis to determine the NAEG and CDE

Variable	Coeff. of det. (R^2)/corr. coeff. (r)
AENG	0.9803 / 0.9901
CDE	0.9895 / 0.9947

Table 8

Results of the variables in relation with each variable to define their significance in the DE/MARS NAEG model

Variable	Nsubsets	*GCV*	RSS
AR	8	100.0	100.0
CDE	6	17.7	17.4
ANE	5	10.7	10.6
APE	4	9.1	9.0
AFC	3	7.8	7.5
BVPL	2	5.6	5.4

Table 9

Results of the variables in relation with each variable to define their significance in the DE/MARS CDE model

Variable	Nsubsets	*GCV*	RSS
N2OE	6	100.0	100.0
AENG	5	27.2	27.0
ME	4	17.6	17.4