GEOFIZIKA VOL. 30 2013

Original scientific paper
UDC 551.509.313.4

Comparative analysis of ozone level prediction models using gene expression programming and multiple linear regression

Saeed Samadianfard[1], *Reza Delirhasannia*[1], *Özgür Kişi*[2] *and Elena Agirre-Basurko*[3]

[1] University of Tabriz, Faculty of Agriculture, Department of Water Engineering, Tabriz, Iran

[2] Canik Basari University, Faculty of Architecture and Engineering, Department of Civil Engineering, Samsun, Turkey

[3] University of the Basque Country, School of Technical Industrial Engineering, Department of Applied Mathematics, Bilbao, Spain

Received 19 January 2013, in final form 11 March 2013

Ground-level ozone (O_3) has been a serious air pollution problem for several decades and in many metropolitan areas, due to its adverse impact on the human respiratory system. Therefore, to reduce the risks of O_3 related damages, developing, maintaining and improving short term ozone forecasting models is needed. This paper presents the results of two prognostic models including gene expression programming (GEP), which is a variant of genetic programming (GP), and multiple linear regression (MLR) to forecast ozone levels in real-time up to 6 hours ahead at four stations in Bilbao, Spain. The inputs to the GEP were meteorological conditions (wind speed and direction, temperature, relative humidity, pressure, solar radiation and thermal gradient), hourly ozone levels and traffic parameters (number of vehicles, occupation percentage and velocity), which were measured in the years of 1993–94. The performances of developed models were compared with observed values and were evaluated using specific performance measurements for the air quality models established in the Model Validation Kit and recommended by the US Environmental Protection Agency. It was found that the GEP in most cases gives superior predictions. Finally it can be concluded on the basis of the results of this study that gene expression programming appears to be a promising technique for the prediction of pollutant concentrations.

Keywords: air quality modeling, gene expression programming, multiple linear regression, ozone level forecasting, Bilbao area, Spain

1. Introduction

Analysis and forecasting of air quality parameters are important topics of atmospheric and environmental research today due to the health impact caused

by air pollution. As one of major pollutants, ozone, especially ground level ozone, is responsible for various adverse effects on both human being and foliage (Wang et al., 2003). Furthermore, ozone levels play an important role in damage to plant species and it can cause harmful effects in vegetation during the growing season. Ozone is unique among pollutants because it is not emitted directly into the air. This is the main reason why ozone is such a serious environmental problem that is difficult to predict and control. Ozone results from complex chemical reactions in the atmosphere (Abdul-Wahab and Al-Alawi, 2002). Therefore, to reduce the risks of O_3 related damages, developing, maintaining and improving short term ozone forecasting models is needed. Accordingly, several studies presented different statistical approaches to predict O_3 concentrations (Robeson and Steyn, 1990; Comrie, 1997; Chen et al., 1998; Hubbard and Cobourn, 1998; Cobourn and Hubbard, 1999; Prybutok et al., 2000; Gardner and Dorling, 2000; Ballester et al., 2002; Chaloulakou et al., 2003; Baur et al., 2004; Agirre-Basurko et al., 2006; Schlink et al., 2006; Al-Alawi et al., 2008; Omidvari et al., 2008; Pires et al., 2008, 2010, 2011; Ortiz-García et al., 2010). On the other hand, artificial neural network (ANN) systems are capable of representing highly nonlinear relationships between variables. Many ANN models have been successfully applied on ozone forecasting (Ruiz-Suarez et al., 1995; Yi and Prybutok, 1996; Comrie, 1997; Gardner and Dorling, 1998; Kolehmainen et al., 2001; Balaguer et al., 2002; Lu et al., 2002; Wang et al., 2003; Zolghadri et al., 2004; Ordieres et al., 2005; Agirre-Basurko et al., 2006; Sousa et al., 2007; Dudot et al., 2007; Al-Alawi et al., 2008; Tsai et al., 2009; Pires and Martins, 2011). There are also some studies that applied evolutionary computation to determine the model for predicting O_3 levels (Pires et al., 2010, 2011; Feng et al., 2011).

This study employs gene expression programming (GEP) which has been applied to a wide range of problems in artificial intelligence, artificial life, engineering and science, financial markets, industrial, chemical and biological processes, and mechanical models including symbolic regression, multi-agent strategies, time series prediction, circuit design and evolutionary neural networks. Research and application of evolutionary computing, over the years, have led to the independent development of five approaches, i.e., evolution strategies, evolutionary programming, classifier systems, genetic algorithms, and genetic programming.

GEP, a flavor of GP, can be successively applied to areas where (i) the interrelationships among the relevant variables are poorly understood (or where it is suspected that the current understanding may well be wrong), (ii) finding the size and shape of the ultimate solution is difficult and a major part of the problem, (iii) conventional mathematical analysis does not, or cannot, provide analytical solutions, (iv) an approximate solution is acceptable (or is the only result that is ever likely to be obtained), (v) small improvements in performance are routinely measured (or easily measurable) and highly prized, (vi) there is a large amount of data in computer readable form, that requires examination, classification, and integration, e.g., molecular biology for protein and DNA sequences,

astronomical data, satellite observation data, financial data, marketing transaction data, or data on the World Wide Web (Banzhaf et al., 1998).

In recent years, GEP have attracted researchers in many disciplines of science and engineering, since it is capable of correlating large and complex datasets without any prior knowledge of the relationships among them. Applications of GEP include those in the areas of splitting tensile strength of concrete (Özcan, 2012), cost prediction for highway construction (Lu et al., 2011) and statistical downscaling of watershed precipitation (Hashmi et al., 2011). In the present study, for the first time, a gene expression programming-based model was built to forecast O_3 levels in the Bilbao area. Furthermore, traffic variables were used as predictor variables in the developed models. The primary goal of the work was to build an accurate mathematical model to forecast O_3 levels k hours ahead in the Bilbao area ($k = 1, 2, ..., 6$). Two techniques were applied to build the models: the gene expression programming and multiple linear regression. Based on these techniques, six different models were designed, and comparisons between them established the most efficient performer as a forecasting tool.

2. Techniques applied in modelling

2.1. Multiple linear regression

The general form of a multiple linear regression could be written as:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_p X_{ip} + \varepsilon_i \tag{1}$$

where, for a set of i observations, Y_i is the predicted variable, β_0 is a coefficient, $\beta_1, \beta_2, ..., \beta_p$ are the coefficients of the $X_{i1}, X_{i2}, ..., X_{ip}$ independent variables (predictors) and ε_i is the residual error (difference between observations and predicted values).

The hypotheses required to apply multiple linear regression are: (i) the predictor variables must be independent, and (ii) the residual errors ε_i must be independent and they must be normally distributed, with 0 mean and σ^2 constant variance.

The observations $\{X_{i1}, X_{i2}, ..., X_{ip}, Y_i\}$, $i = 1, 2, ..., n$ are helpful in the estimation of the parameters β and they form the calibration set. The least square method is the usual technique used to estimate the parameters. Hence, the equation for the predicted value is:

$$\hat{Y}_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + ... + b_p X_{ip} \tag{2}$$

where, b_i are the estimations of the β_i parameters and $\hat{Y}_i$ is the predicted value.

The goal of the regression analysis is to determine the values of the parameters of the regression equation and then to quantify the goodness of the fit in respect of the dependent variable Y.

2.2. General overview of genetic programming

In this section, a brief overview of the GP and GEP is given. Detailed explanations of GP and GEP are provided by Koza (1992) and Ferreira (2006), respectively. GP was first proposed by Koza (1992). It is a generalization of genetic algorithms (GAs) (Goldberg, 1989). The fundamental difference between GA, GP, and GEP is due to the nature of the individuals. In the GA, the individuals are linear strings of fixed length (chromosomes). In the GP, the individuals are nonlinear entities of different sizes and shapes (parse trees), and in GEP the individuals are encoded as linear strings of fixed length (the genome or chromosomes), which are afterwards expressed as nonlinear entities of different sizes and shapes (Ferreira, 2001a,b). GP is a search technique that allows the solution of problems by automatically generating algorithms and expressions. These expressions are coded or represented as a tree structure with its terminals (leaves) and nodes (functions). GP applies GAs to a "population" of programs, i.e., typically encoded as tree-structures. Trial programs are evaluated against a "fitness function" and the best solutions selected for modification and re-evaluation. This modification-evaluation cycle is repeated until a "correct" program is produced.

There are five major preliminary steps for solving a problem by using GEP. These are the determination of (i) the set of terminals, (ii) the set of functions, (iii) the fitness measure, (iv) the values of the numerical parameters and qualitative variables for controlling the run, and (v) the criterion for designating a result and terminating a run (Koza, 1992).

A GEP flowchart improved by Ferreira (2001b) is presented in Fig. 1.

The automatic program generation is carried out by means of a process derived from Darwin's evolution theory, in which, after subsequent generations, new trees (individuals) are produced from old ones via crossover, copy, and mutation (Fuchs, 1998; Luke and Spector, 1998). Based on natural selection, the best trees will have more chances of being chosen to become part of the next generation. Thus, a stochastic process is established where, after successive generations, a well-adapted tree is obtained.

There are five major steps in preparing to use GEP of which the first is to choose the fitness function. The fitness of an individual program i for fitness case j is evaluated by Ferreira (2006) using:

$$If \quad E(ij) \le p, \ then \quad f_{(ij)} = 1; \ else \quad f_{(ij)} = 0 \tag{3}$$

where p is the precision and $E(ij)$ is the error of an individual program i for fitness case j. For the absolute error, this is expressed by:

$$E(ij) = \left| P_{(ij)} - T_j \right. \tag{4}$$

Again for the absolute error, the fitness f_i of an individual program i is expressed by:

$$f_i = \sum_{j=1}^{n} \left(R - \left| P_{(ij)} - T_j \right| \right) \tag{5}$$

where R is the selection range, $P_{(ij)}$ is the value predicted by the individual program i for fitness case j (out of n fitness cases) and T_j is the target value for fitness case j. The second major step consists of choosing the set of terminals T and the set of functions F to create the chromosomes. In this problem, the terminal set obviously consists of the independent variables. The choice of the appropriate function set is not so obvious. However, a good guess can always be helpful in order to include all of the necessary functions. In this study, four basic arithme-

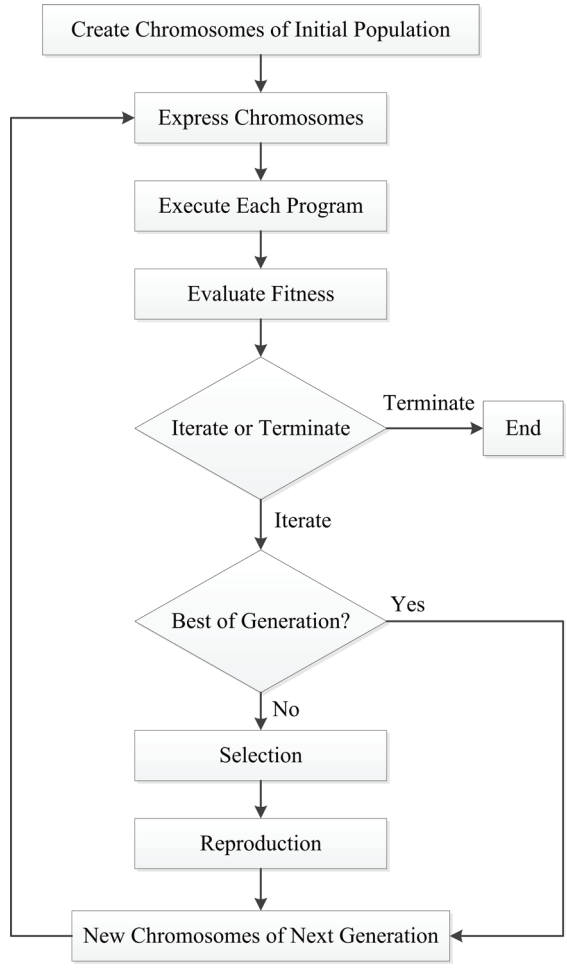

Figure 1. GEP flowchart.

tic operators, i.e., (+, –, ×, /) and some basic mathematical functions, i.e., (√, *Ln(x), exp, Power, Sin, Cosine, Arctangent*) were utilized. The third major step is to choose the chromosomal architecture, i.e., the length of the head and the number of genes. Values of the length of the head, $h = 10$, and four genes per chromosome were employed. The fourth major step is to choose the linking function. In this study, the sub-programs were linked by addition. Finally, the fifth major step is to choose the set of genetic operators that cause variation and their rates. A combination of all genetic operators, i.e., mutation, transposition and recombination, was used for this purpose.

The parameters of the training of the GEP are given in Tab. 1.

Table 1. Parameters of the GEP model.

Parameter	Value
Function set	+, –, ×, /, √, *Ln(x)*, e^x, 10^x, *Power, Sin, Cosine, Arctangent*
Chromosomes	30
Head size	10
Number of Genes	4
Linking Function	Addition (+)
Mutation Rate	0.044
Inversion Rate	0.1
One-Point Recombination Rate	0.3
Two-Point Recombination Rate	0.3
Gene Recombination Rate	0.1
Gene Transposition Rate	0.1

3. Database

An air pollution network managed by the Basque Government since 1977 measures hourly meteorological parameters and air pollution variables at each station in Bilbao. In the same way, the traffic network managed by the Local Municipality of Bilbao measures two different and independent traffic variables at each station: the variable *NV* indicates the number of vehicles circulating every 10 min and the variable *OP* indicates the fraction of time for which the area of road is occupied by a vehicle. Both network measures are highly consistent. The data used in this work were hourly current (at time t) data and historical (at time $t-z$, $z = 1, 2, \ldots, 6$) data from the air pollution network and the traffic network of Bilbao during the years 1993–94. The data selected jointly

Table 2. Meteorological variables, air pollution variables and traffic variables used to develop the models.

Classification of variables	Variables	Notation
Meteorology	Wind speed ($m\,s^{-1}$)	V_x
	Wind direction (°)	V_y
	Temperature (°*C*)	TEM
	Relative humidity (%)	HUM
	Radiation ($cal\,cm^{-2}\,h^{-1}$)	RAD
	Thermal gradient (°*C*)	GRAD
Pollution	Ozone ($mg\,m^{-3}$)	O_3
Traffic	Number of vehicles (*vehicle / 10 min*)	NV
	Occupation percentage (%)	OP
	Velocity ($km\,h^{-1}\,100^{-1}$)	KH

reduced the study to four stations in Bilbao, namely Deusto, Elorrieta, Mazarredo and Txurdinaga. These four stations, located in the central area of Bilbao, are close to each other – the greatest distance between any of them is less than 5 kilometers. The selection of the variables of this study (Tab. 2) is based on earlier works (Ibarra-Berastegi et al., 2001a).

The meteorological variables considered were wind speed and direction, thermal contrast between Feria and Banderas (two stations located at sea level and 200 m above sea level, respectively), relative humidity, pressure, temperature and radiation. In the same way, O_3 levels measured at the four stations were used. All these variables were measured hourly. Finally, as several works have proven that traffic plays a significant role in the formation of ozone (Mayer, 1999; Borrego et al., 2000; Ibarra-Berastegi et al., 2001b), the database was completed with the mean hourly values of three traffic variables registered in Bilbao in the years of 1993–94: (i) the number of vehicles *NV*, (ii) the occupation percentage *OP*, and (iii) the variable $KH = (NV / OP)$, which gives an idea of the velocity.

Tabs. 3–6 represent the hourly statistical parameters of variables in four stations. In these tables, the terms X_{mean}, X_{min}, X_{max}, S_x, C_v and C_{sx} denote the mean, minimum, maximum, standard deviation, coefficient of variation and skewness coefficient, respectively. From theses tables, it is clear that the gradient has the maximum skewness for all the stations. Wind direction, relative humidity and radiation also show skewed distribution. Temperature and number of vehicles show normal distribution because they have significantly low skewness. Tabs. 7–10 show correlations between meteorological and traffic parameters in four mentioned stations. As it can be seen in Tabs. 7 and 8, humidity and

Table 3. Hourly statistical parameters of the observed data in Deusto station.

Variable	X_{mean}	X_{min}	X_{max}	S_x	C_V	C_{sx}
V_x	0.95	–5.94	10.79	2.26	2.37	0.29
V_y	–0.35	–8.17	3.89	1.54	–4.40	–1.20
TEM	15.41	–0.80	35.20	5.08	0.33	0.05
HUM	81.98	28.40	97.00	12.47	0.15	–1.40
RAD	0.23	0.00	1.50	0.35	1.54	1.72
GRAD	–2.50	–69.80	4.50	4.94	–1.98	–10.23
NV	400.72	10.00	835.74	248.51	0.62	0.10
OP	5.81	1.52	20.05	3.93	0.68	0.98
KH	0.42	0.03	0.94	0.15	0.36	–0.40
O_3	33.03	0.00	135.80	23.54	0.71	0.56

Note: The terms X_{mean}, X_{min}, X_{max}, S_x, C_v and C_{sx} denote the mean, minimum, maximum, standard deviation, coefficient of variation and skewness, respectively.

Table 4. Hourly statistical parameters of the observed data in Elorrieta station.

Variable	X_{mean}	X_{min}	X_{max}	S_x	C_V	C_{sx}
V_x	0.96	–5.94	10.79	2.23	2.31	0.36
V_y	–0.41	–8.17	3.89	1.60	–3.90	–1.23
TEM	15.29	0.30	35.20	5.01	0.33	0.09
HUM	82.29	28.40	97.00	12.14	0.15	–1.43
RAD	0.20	0.00	1.50	0.33	1.65	1.89
GRAD	–2.50	–69.80	4.50	4.94	–1.98	–10.23
NV	400.72	10.00	835.74	248.51	0.62	0.10
OP	5.81	1.52	20.05	3.93	0.68	0.98
KH	0.42	0.03	0.94	0.15	0.36	–0.40
O_3	28.69	1.00	137.00	22.76	0.79	1.06

Note: The terms X_{mean}, X_{min}, X_{max}, S_x, C_v and C_{sx} denote the mean, minimum, maximum, standard deviation, coefficient of variation and skewness, respectively.

solar radiation have higher correlations with ozone levels in comparison with other parameters for the stations, Deusto and Elorrieta. For the Mazarredo and Txurdinaga stations, however, wind speed and solar radiation have higher correlations with ozone level than those of the other variables. Number of vehicles, in general, has the lowest correlation.

Table 5. Hourly statistical parameters of the observed data in Mazarredo station.

Variable	X_{mean}	X_{min}	X_{max}	S_x	C_V	C_{sx}
V_x	0.90	–5.94	10.79	2.20	2.45	0.43
V_y	–0.39	–8.17	3.89	1.53	–3.91	–1.25
TEM	15.12	–0.10	34.90	4.93	0.33	0.02
HUM	82.91	28.40	97.00	11.85	0.14	–1.51
RAD	0.19	0.00	1.50	0.31	1.66	1.92
GRAD	–2.50	–69.80	4.50	4.94	–1.98	–10.23
NV	400.72	10.00	835.74	248.51	0.62	0.10
OP	5.81	1.52	20.05	3.93	0.68	0.98
KH	0.42	0.03	0.94	0.15	0.36	–0.40
O_3	36.36	0.00	182.5	31.71	0.87	0.85

Note: The terms X_{mean}, X_{min}, X_{max}, S_x, C_v and C_{sx} denote the mean, minimum, maximum, standard deviation, coefficient of variation and skewness, respectively.

Table 6. Hourly statistical parameters of the observed data in Txurdinaga station.

Variable	X_{mean}	X_{min}	X_{max}	S_x	C_V	C_{sx}
V_x	0.89	–5.94	10.79	2.22	2.49	0.44
V_y	–0.43	–8.17	3.89	1.56	–3.63	–1.26
TEM	15.03	–0.10	31.30	4.92	0.33	0.01
HUM	82.86	30.80	97.00	11.88	0.14	–1.43
RAD	0.20	0.00	1.50	0.33	1.63	1.85
GRAD	–2.50	–69.80	4.50	4.94	–1.98	–10.23
NV	400.72	10.00	835.74	248.51	0.62	0.10
OP	5.81	1.52	20.05	3.93	0.68	0.98
KH	0.42	0.03	0.94	0.15	0.36	–0.40
O_3	32.41	2.00	158.50	27.04	0.83	1.01

Note: The terms X_{mean}, X_{min}, X_{max}, S_x, C_v and C_{sx} denote the mean, minimum, maximum, standard deviation, coefficient of variation and skewness, respectively.

4. Methodology

Gene expression programming-based model (GEP) and multiple linear regression model (MLR) were developed using the current and past values of the indicated variables measured in the Bilbao air pollution and traffic networks

Table 7. Correlations between meteorological and traffic parameters in Deusto station.

	V_x	V_y	TEM	HUM	RAD	GRAD	NV	OP	KH	O_3
V_x	1.00									
V_y	0.25	1.00								
TEM	–0.05	0.16	1.00							
HUM	0.21	0.25	–0.39	1.00						
RAD	0.21	0.22	0.42	–0.48	1.00					
GRAD	–0.11	–0.05	–0.24	0.06	–0.15	1.00				
NV	0.16	0.15	0.15	–0.31	0.45	–0.11	1.00			
OP	0.08	0.10	0.09	–0.25	0.36	–0.07	0.89	1.00		
KH	0.17	0.09	0.16	–0.11	0.19	–0.10	0.27	–0.13	1.00	
O_3	0.18	–0.19	0.30	–0.47	0.42	–0.14	–0.02	–0.07	0.10	1.00

Table 8. Correlations between meteorological and traffic parameters in Elorrieta station.

	V_x	V_y	TEM	HUM	RAD	GRAD	NV	OP	KH	O_3
V_x	1.00									
V_y	0.24	1.00								
TEM	–0.05	0.17	1.00							
HUM	0.18	0.27	–0.37	1.00						
RAD	0.24	0.22	0.40	–0.45	1.00					
GRAD	–0.12	–0.06	–0.26	0.07	–0.19	1.00				
NV	0.18	0.14	0.14	–0.30	0.43	–0.11	1.00			
OP	0.10	0.10	0.08	–0.22	0.33	–0.07	0.89	1.00		
KH	0.15	0.07	0.16	–0.14	0.20	–0.10	0.27	–0.13	1.00	
O_3	0.23	–0.15	0.08	–0.35	0.26	–0.10	–0.10	–0.16	0.11	1.00

during the years of 1993–94. After introducing the appropriate inputs, the outputs of the models were the forecasted O_3 levels at time $t+k$, $k=1, 2,..., 6$. Two third of data were used to build the models and the residual one third data were used to test the models.

4.1. Building the models

The equation (6) represents parameters that have been used for building GEP and MLR models:

Table 9. Correlations between meteorological and traffic parameters in Mazarredo station.

	V_x	V_y	TEM	HUM	RAD	GRAD	NV	OP	KH	O_3
V_x	1.00									
V_y	0.21	1.00								
TEM	–0.08	0.15	1.00							
HUM	0.18	0.27	–0.37	1.00						
RAD	0.22	0.20	0.37	–0.44	1.00					
GRAD	–0.07	0.04	–0.03	0.12	–0.07	1.00				
NV	0.16	0.14	0.13	–0.29	0.44	–0.11	1.00			
OP	0.09	0.10	0.09	–0.23	0.35	–0.07	0.89	1.00		
KH	0.16	0.08	0.12	–0.12	0.19	–0.10	0.27	–0.13	1.00	
O_3	0.47	0.12	0.19	–0.23	0.37	–0.11	–0.01	–0.09	0.17	1.00

Table 10. Correlations between meteorological and traffic parameters in Txurdinaga station.

	V_x	V_y	TEM	HUM	RAD	GRAD	NV	OP	KH	O_3
V_x	1.00									
V_y	0.20	1.00								
TEM	–0.10	0.15	1.00							
HUM	0.17	0.28	–0.37	1.00						
RAD	0.21	0.19	0.39	–0.47	1.00					
GRAD	–0.06	0.06	–0.01	0.12	–0.07	1.00				
NV	0.17	0.13	0.12	–0.28	0.47	–0.11	1.00			
OP	0.09	0.10	0.08	–0.22	0.38	–0.07	0.89	1.00		
KH	0.16	0.08	0.12	–0.13	0.21	–0.10	0.27	–0.13	1.00	
O_3	0.40	–0.03	0.12	–0.35	0.38	–0.12	–0.03	–0.10	0.15	1.00

$$O_3(t+k) = f\left(MET(t), TRAF(t), O_3(t-z)\right) \tag{6}$$

The $MET(t)$ variables are current values of temperature, pressure, wind, thermal gradient, relative humidity and global radiation. The $TRAF(t)$ variables are current values of the variables *NV, OP* and *KH*. The $O_3(t-z)$ are the current and historical ($z = 0, 1, 2, ..., 6$) values of O_3 in Deusto, Elorrieta, Mazarredo and Txurdinaga. These are the independent variables of the GEP and MLR models. $O_3(t+k)$, the forecasts of O_3 ($k = 1, 2, ... 6$), are the dependent variables.

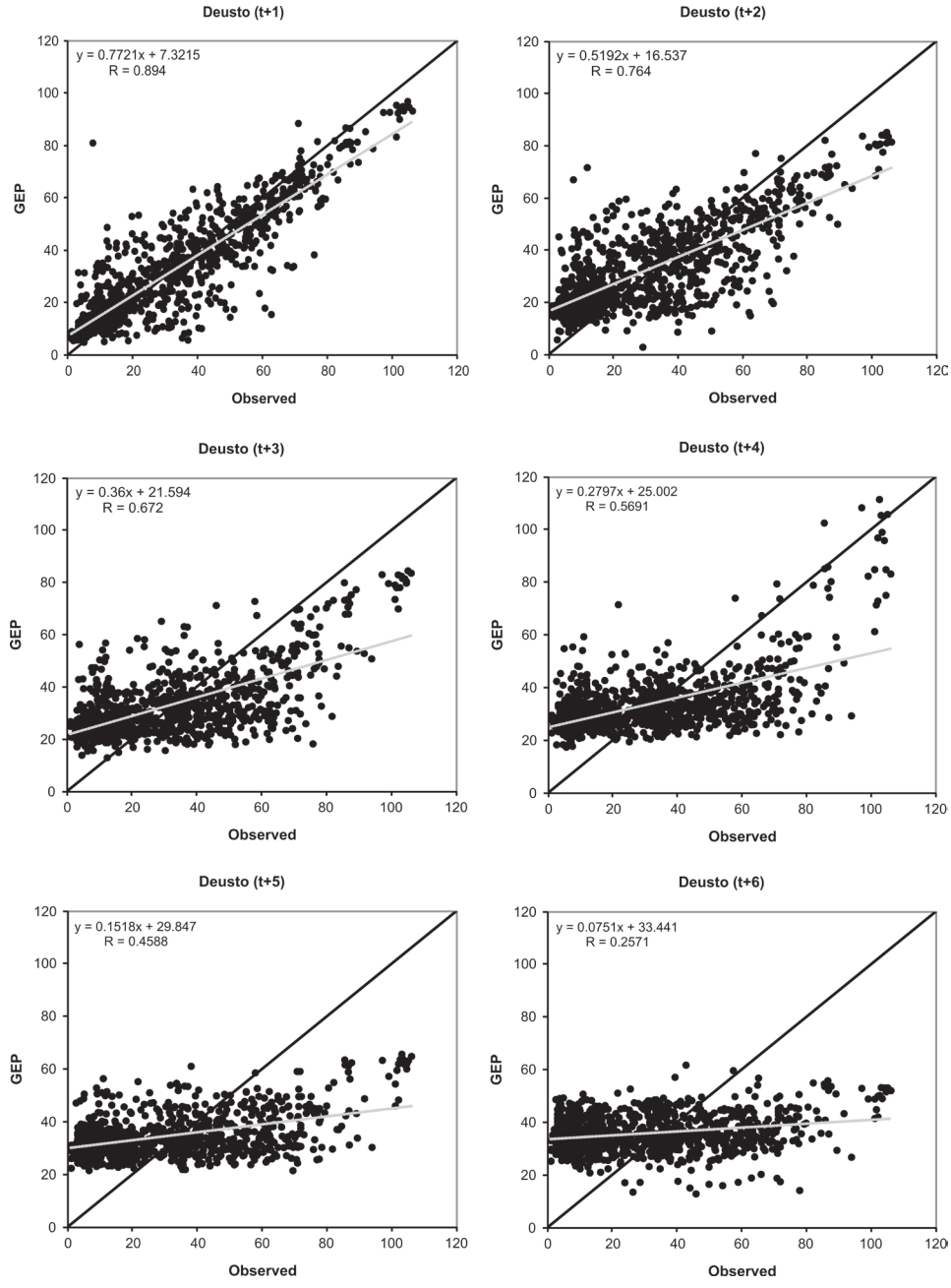

Figure 2. Scatter plots of observed values (x-axis) and forecasted values (y-axis) of $O_3(t+k)$, $k = 1, 2,\ldots, 6$ in Deusto station.

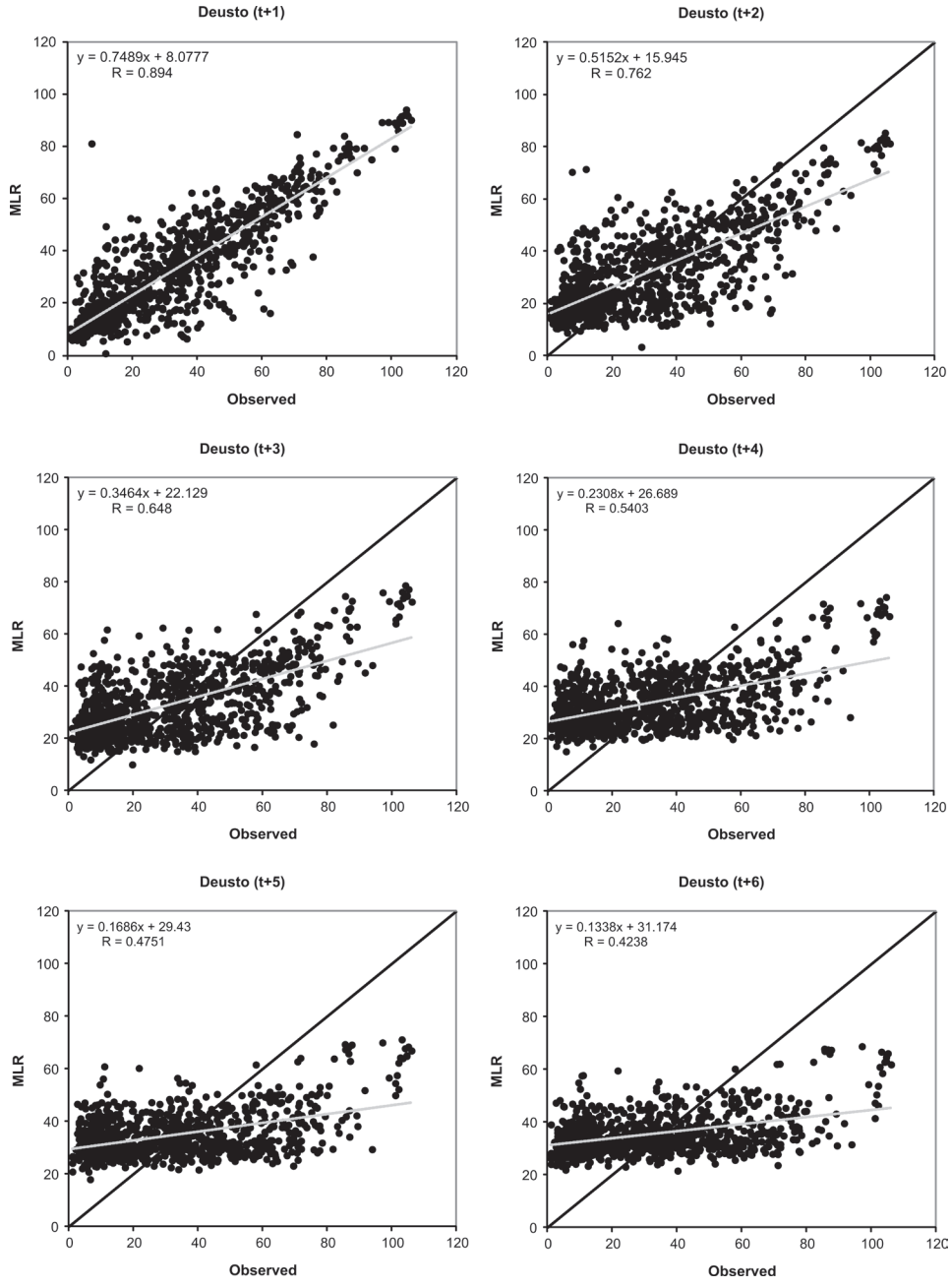

Figure 2. Continued.

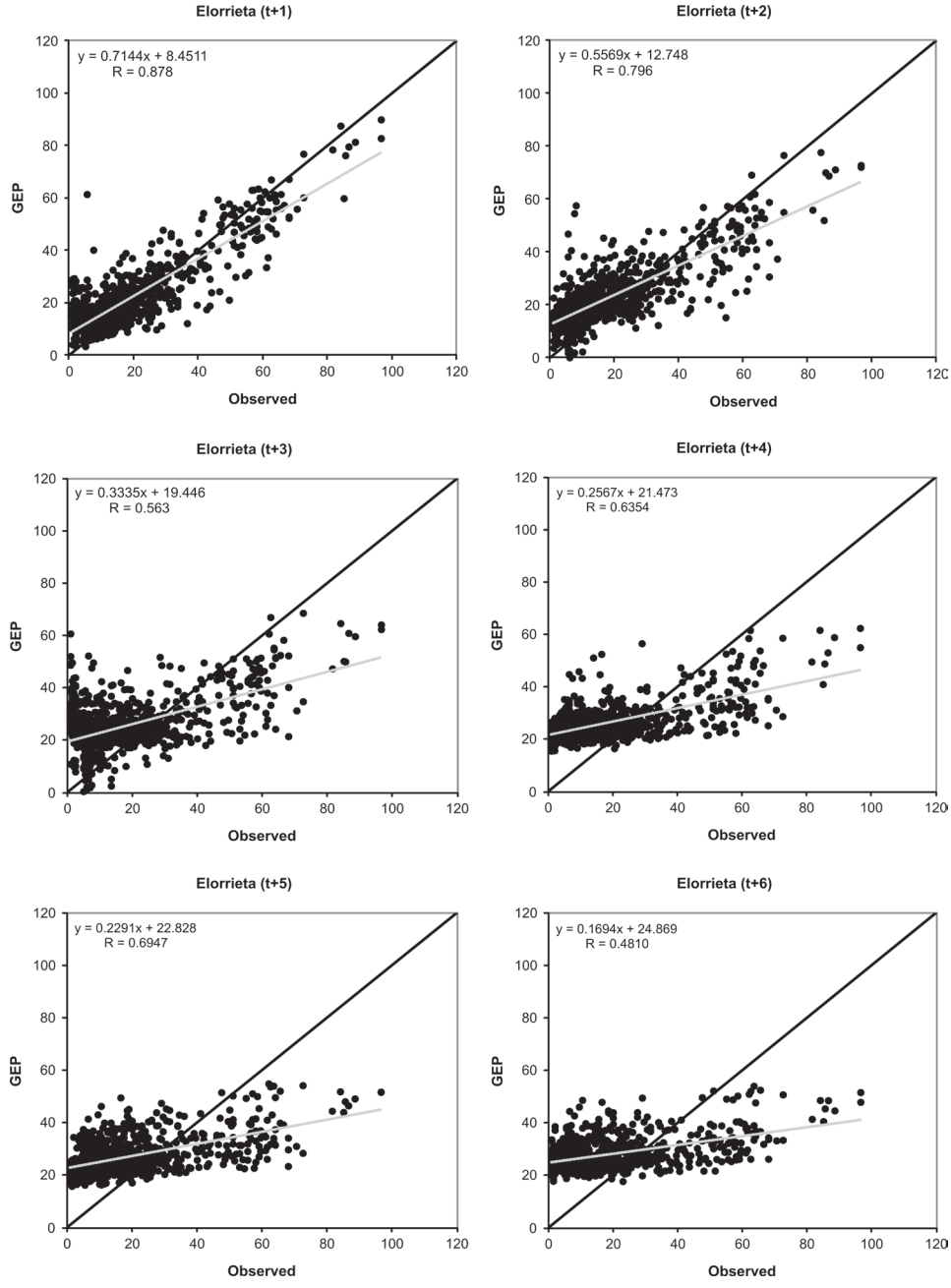

Figure 3. Scatter plots of observed values (x-axis) and forecasted values (y-axis) of $O_3(t+k)$, $k = 1$, 2,..., 6 in Elorrieta station.

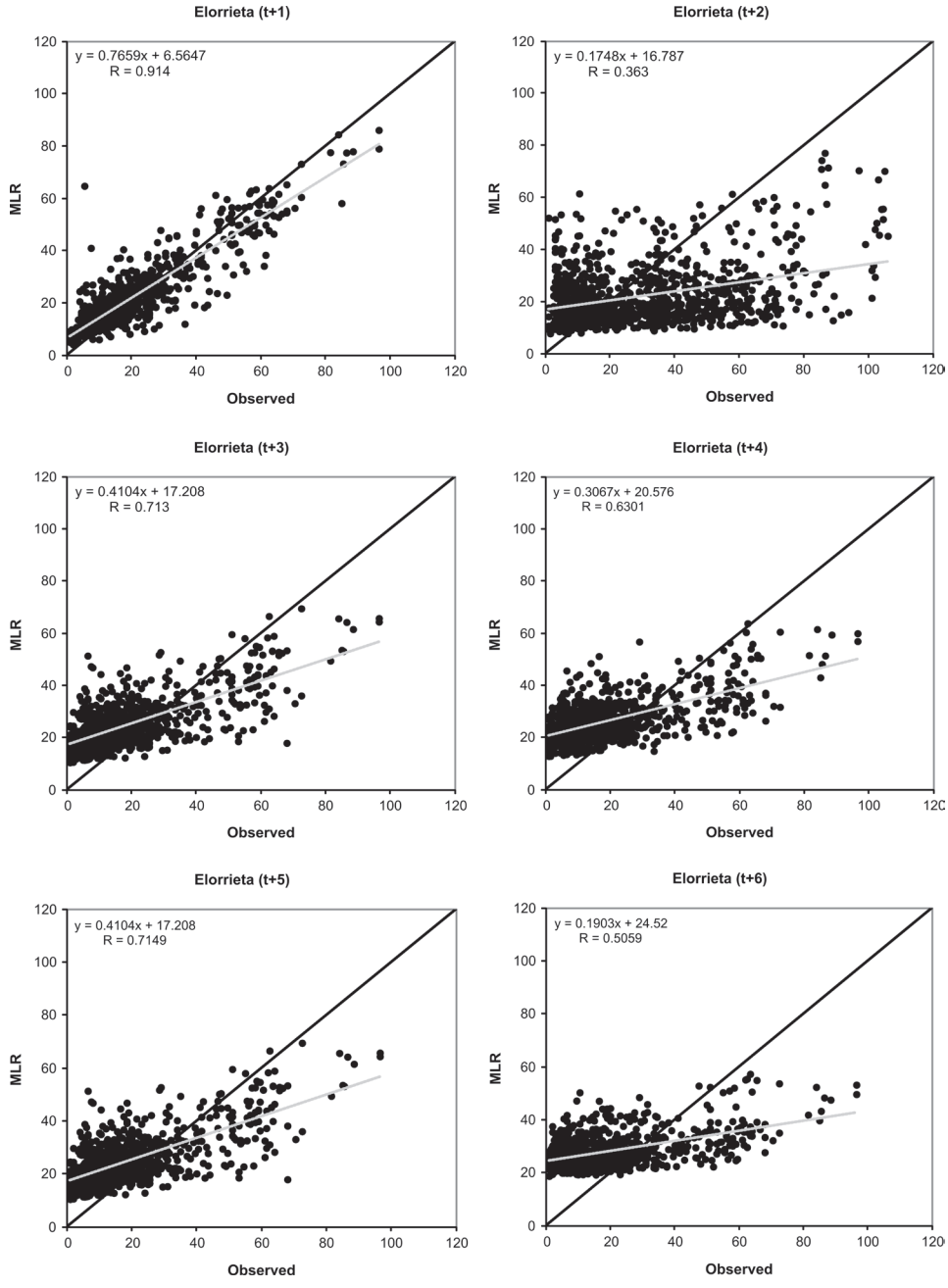

Figure 3. Continued.

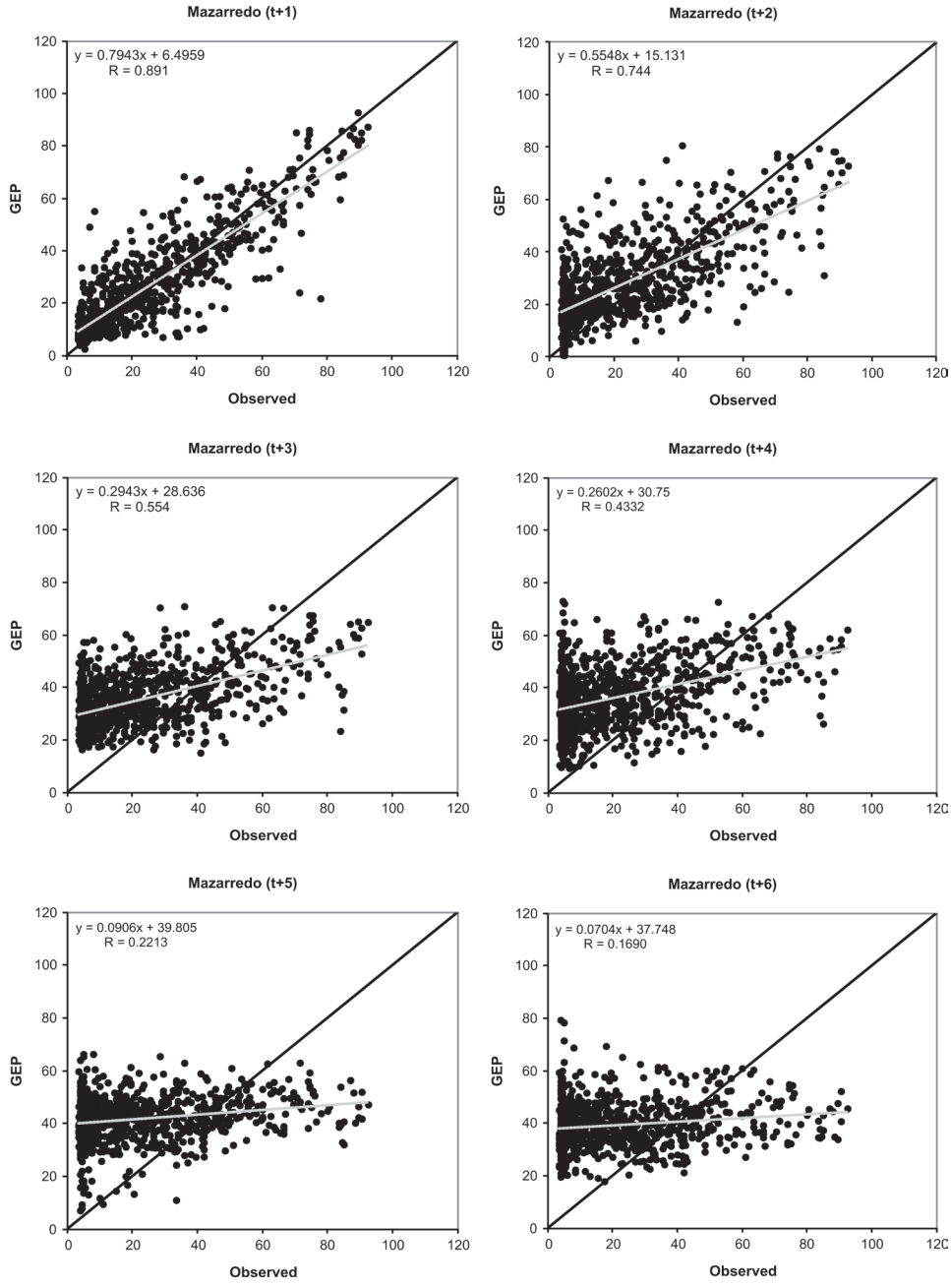

Figure 4. Scatter plots of observed values (x-axis) and forecasted values (y-axis) of $O_3(t+k)$, $k = 1, 2,\ldots, 6$ in Mazarredo station.

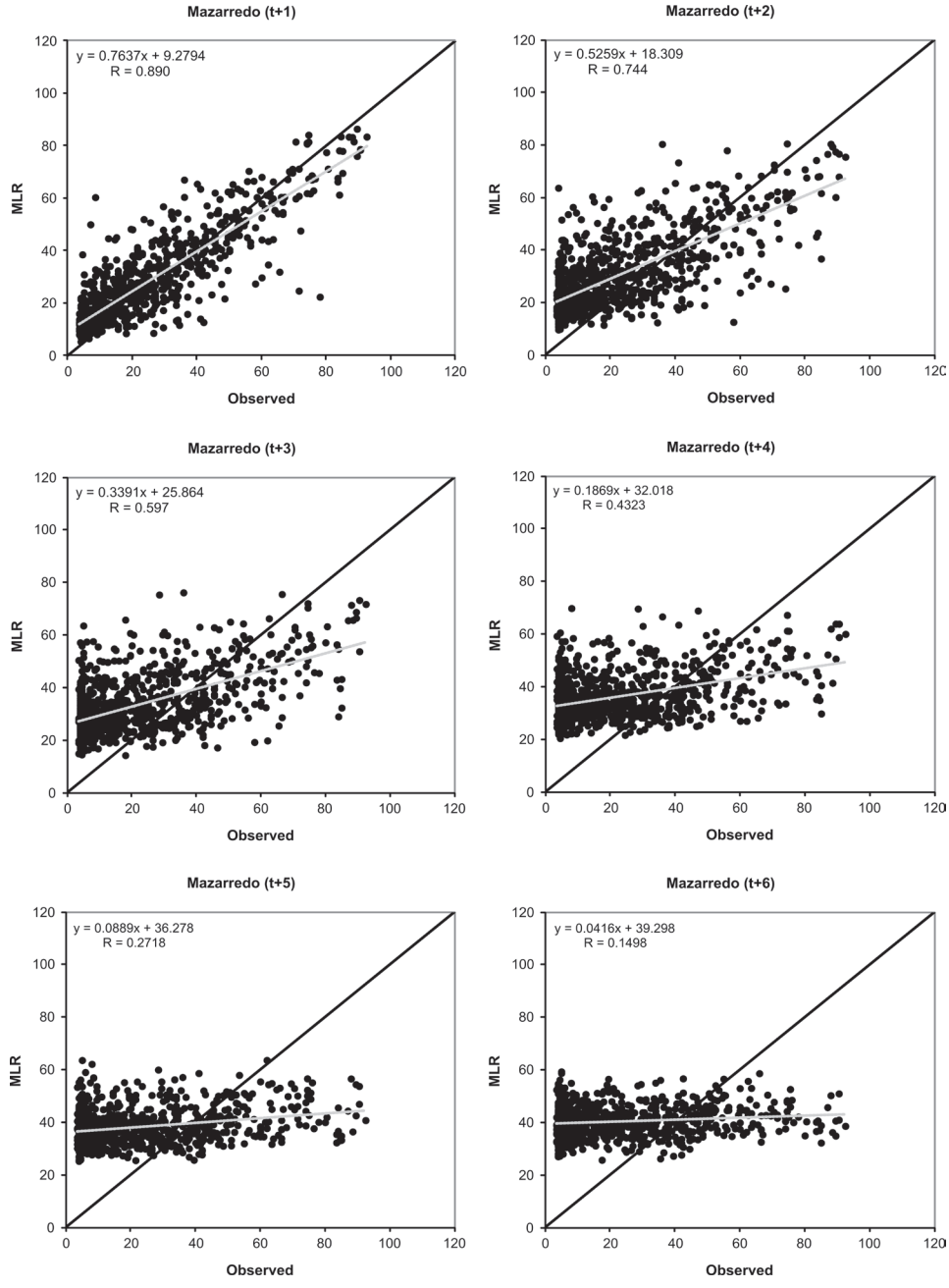

Figure 4. Continued.

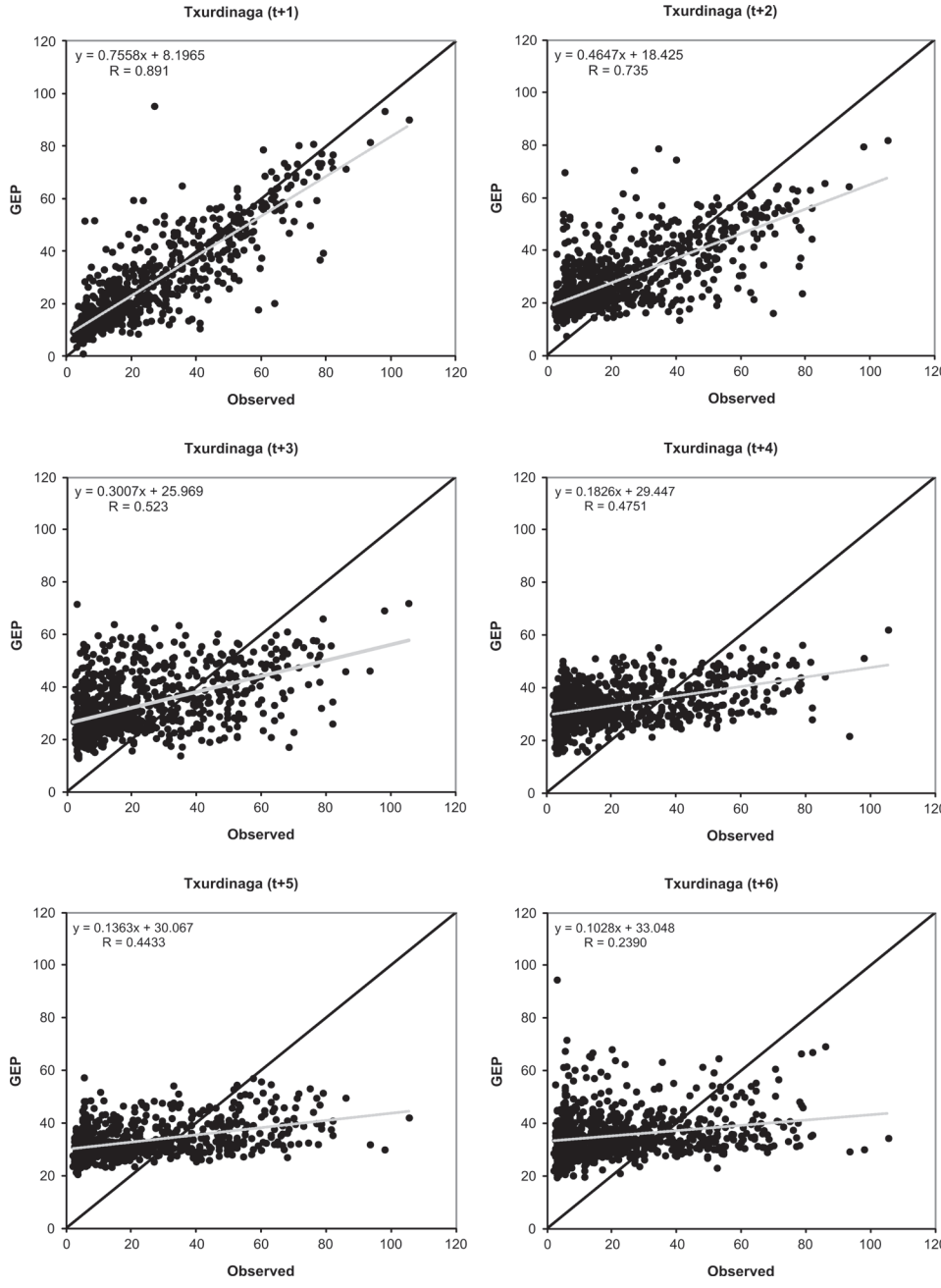

Figure 5. Scatter plots of observed values (x-axis) and forecasted values (y-axis) of $O_3(t+k)$, $k = 1, 2,\ldots, 6$ in Txurdinaga station.

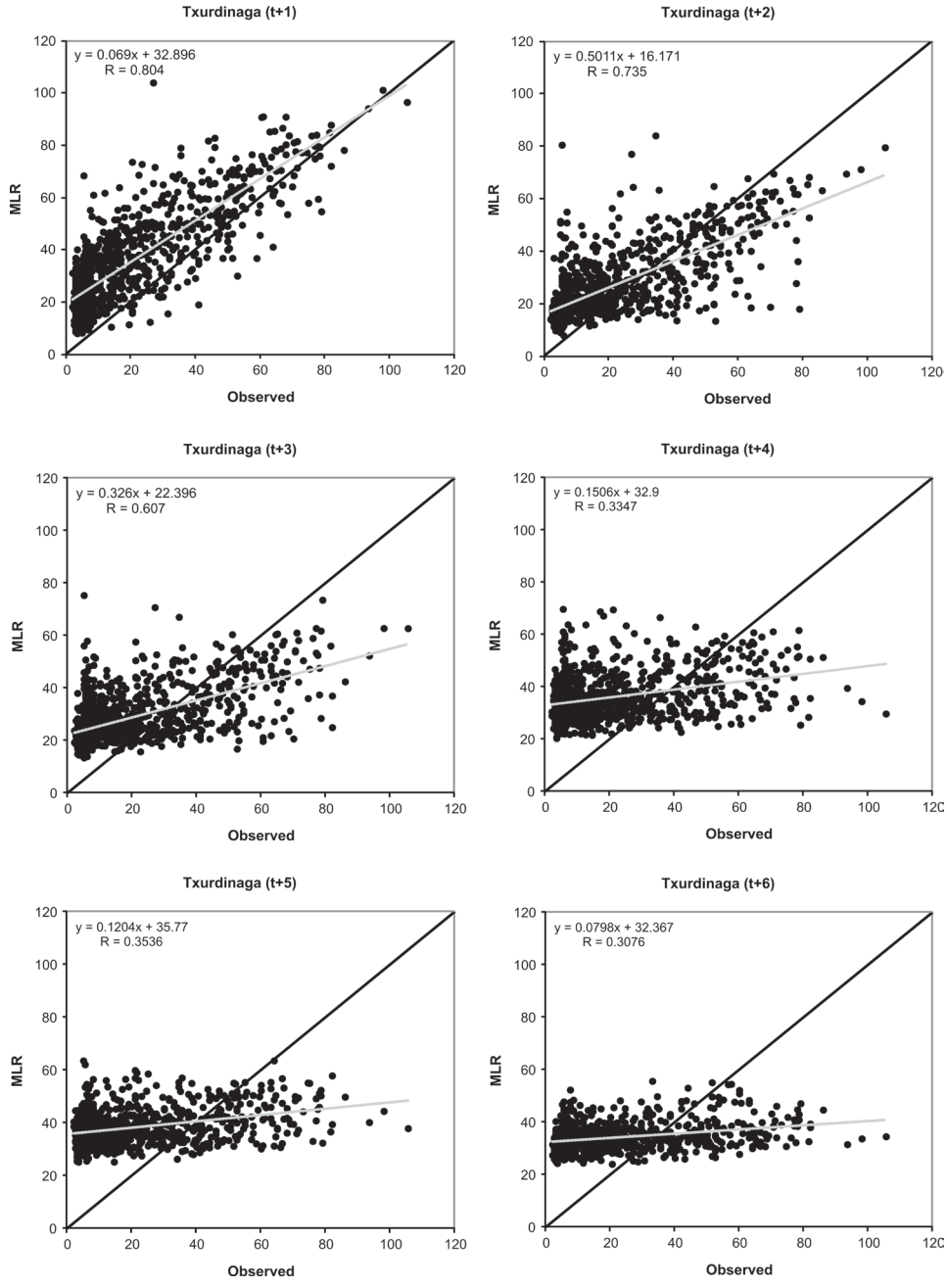

Figure 5. Continued.

4.2. Testing the models

To reduce the uncertainty of applying the appropriate statistics to choose the best model, in 1991 it was decided to initiate a series of workshops. These workshops were supported by COST 710 and COST 615 and the European Association for the Science of Air Pollution (EURASAP). In 1993 the workshop took place in Manno (Switzerland), and it was dedicated to the establishment of objective criteria for comparing different models. Consequently, a data processing package known as the Model Validation Kit (European Commission, 1994) was created, which was improved in the following workshop in Mol (Belgium) in 1994. The kit was formed by criteria based on a previous work (Hanna et al., 1991). Although these measures were thought to compare the performance of cause/effect models, their application to statistical models is immediate. These statistics allow the comparison of the performance of different models, where C_p are the forecasted values and C_o are the observed values, σ indicates the standard deviation and *Mean* is the mean value. The proposed measures in the Model Validation Kit are:

(i) The correlation coefficient between C_o and C_p, R, quantifies the global description of the model:

$$R = \frac{Mean\left[\left(C_o - Mean\left(C_o\right)\right)\left(C_p - Mean\left(C_p\right)\right)\right]}{\left(\sigma_{C_o}\right)\left(\sigma_{C_p}\right)} \tag{7}$$

(ii) The Normalized Mean Square Error, *NMSE*, is a version of the mean square error, but normalized with the object of establishing comparisons among different models:

$$NMSE = \frac{Mean\left(C_o - C_p\right)^2}{Mean\left(C_o\right)Mean\left(C_p\right)} \tag{8}$$

(iii) The factor of two, *FA2*, which gives the percentage of forecasted cases in which the values of the ratio C_o / C_p are in the range [0.5, 2]:

$$0.5 \le C_o/C_p \le 2 \tag{9}$$

(iv) The Fractional Bias, *FB*, is a normalized measure that allows the comparison of the mean of the observed values and the mean of the predicted values. A model with $FB = 0$ is a model that represents perfectly the measured mean value:

$$FB = 2 \times \frac{Mean\left(C_o\right) - Mean\left(C_p\right)}{Mean\left(C_o\right) + Mean\left(C_p\right)} \tag{10}$$

(v) The Fractional Variance, *FV*, is another normalized measure that allows the comparison of the difference between the predicted variance and the observed

Table 11. Values of the Model Validation Kit statistics for Deusto station.

Predicted	Model	NMSE	R	FA2	FB	FV
$O_3(t+1)$	GEP	0.121	0.894	0.972	–0.014	0.146
	MLR	0.124	0.892	0.971	–0.016	0.174
$O_3(t+2)$	GEP	0.245	0.764	0.914	–0.064	0.381
	MLR	0.251	0.762	0.935	–0.042	0.386
$O_3(t+3)$	GEP	0.326	0.672	0.896	–0.072	0.604
	MLR	0.340	0.648	0.901	–0.075	0.606
$O_3(t+4)$	GEP	0.384	0.569	0.883	–0.100	0.682
	MLR	0.404	0.540	0.878	–0.107	0.803
$O_3(t+5)$	GEP	0.446	0.458	0.862	–0.128	1.005
	MLR	0.436	0.475	0.859	–0.131	0.953
$O_3(t+6)$	GEP	0.506	0.257	0.860	–0.164	1.093
	MLR	0.455	0.423	0.847	–0.150	1.040

Table 12. Values of the Model Validation Kit statistics for Elorrieta station.

Predicted	Model	NMSE	R	FA2	FB	FV
$O_3(t+1)$	GEP	0.214	0.878	0.765	–0.190	0.205
	MLR	0.157	0.914	0.781	–0.140	0.176
$O_3(t+2)$	GEP	0.337	0.796	0.511	–0.265	0.353
	MLR	0.834	0.363	1.401	0.312	0.700
$O_3(t+3)$	GEP	0.581	0.563	0.648	–0.384	0.511
	MLR	0.455	0.713	0.649	–0.346	0.539
$O_3(t+4)$	GEP	0.571	0.635	0.619	–0.410	0.849
	MLR	0.554	0.630	0.619	–0.408	0.690
$O_3(t+5)$	GEP	0.560	0.694	0.586	–0.442	0.853
	MLR	0.454	0.714	0.647	–0.346	0.539
$O_3(t+6)$	GEP	0.685	0.48	0.590	–0.476	0.950
	MLR	0.672	0.505	0.589	–0.476	0.907

variance. A model with $FV=0$ is a model whose variance is equal to the variance of the observed values:

$$FV = 2 \times \frac{\sigma_{C_o} - \sigma_{C_p}}{\sigma_{C_o} + \sigma_{C_p}} \tag{11}$$

Table 13. Values of the Model Validation Kit statistics for Mazarredo station.

Predicted	Model	NMSE	R	FA2	FB	FV
O_3 (t + 1)	GEP	0.193	0.891	0.837	–0.111	0.110
	MLR	0.207	0.890	0.719	–0.204	0.155
O_3 (t + 2)	GEP	0.408	0.745	0.694	–0.268	0.290
	MLR	0.441	0.744	0.627	–0.360	0.343
O_3 (t + 3)	GEP	0.698	0.554	0.548	–0.481	0.612
	MLR	0.632	0.597	0.574	–0.479	0.551
O_3 (t + 4)	GEP	0.812	0.433	0.548	–0.569	0.499
	MLR	0.793	0.432	0.541	–0.564	0.793
O_3 (t + 5)	GEP	1.025	0.221	0.483	–0.700	0.838
	MLR	0.907	0.272	0.519	–0.620	1.014
O_3 (t + 6)	GEP	0.987	0.169	0.516	–0.646	0.824
	MLR	0.991	0.149	0.499	–0.668	1.131

Table 14. Values of the Model Validation Kit statistics for Txurdinaga station.

Predicted	Model	NMSE	R	FA2	FB	FV
O_3 (t + 1)	GEP	0.206	0.891	0.749	–0.170	0.163
	MLR	0.574	0.804	0.499	–0.476	0.026
O_3 (t + 2)	GEP	0.462	0.734	0.625	–0.357	0.450
	MLR	0.437	0.735	0.670	–0.300	0.378
O_3 (t + 3)	GEP	0.708	0.523	0.568	–0.500	0.540
	MLR	0.596	0.607	0.617	–0.403	0.603
O_3 (t + 4)	GEP	0.765	0.475	0.550	–0.536	0.889
	MLR	0.897	0.335	0.522	–0.612	0.759
O_3 (t + 5)	GEP	0.782	0.443	0.558	–0.528	1.059
	MLR	0.942	0.353	0.486	–0.668	0.984
O_3 (t + 6)	GEP	0.920	0.239	0.539	–0.592	0.797
	MLR	0.857	0.308	0.550	–0.562	1.176

In this study, the calculation of the statistics of the Model Validation Kit on the test set determined the goodness of the fit of the GEP and MLR models in a quantitative manner. These results were compared with the values of the statistics corresponding to observations.

5. Results and discussion

Tabs. 11–14 show the values of the statistics included in the Model Validation Kit for the observation, GEP and MLR on the test set in Deusto, Elorrieta, Mazarredo and Txurdinaga stations, respectively. The best forecast has *NMSE*, *FV* and *FB* values equal to zero and the corresponding values of *R* and *FA2* equal to unit.

In the case of $O_3(t+1)$ forecast, the lowest values of *NMSE*, *FB* and *FV* were obtained with the GEP model, being lower than the corresponding values obtained by the MLR model in all stations except the Elorrieta. Also, the *R* and *FA2* values of the GEP model are higher than those of the MLR model for the Deusto, Mazarredo and Txurdinaga stations. For the Elorrieta, however, the MLR model has higher *R* and *FA2* than the GEP model. All these statistics indicate that the GEP model generally performs better than the MLR model in forecasting one-hour ahead ozone levels.

In the case of $O_3(t+2)$ forecast, the lowest values of *NMSE* were obtained with the GEP model, being lower than the corresponding values obtained by the MLR model in all stations except the Txurdinaga. The *R* values of the GEP model are higher than those of the MLR model for the Deusto, Mazarredo and Elorrieta stations. For the Txurdinaga, however, MLR model has higher *R* and *FA2* than the GEP model. From these statistics it can be said that the GEP model performs better than the MLR model in forecasting two-hour ahead ozone levels in three out of four stations.

Different trend was seen in the case of $O_3(t+3)$ forecast. The lowest values of *NMSE* and *FB* were obtained with the MLR model in Elorrieta, Mazarredo and Txurdinaga stations. In all stations, the MLR performs better than the GEP model in respect to *FA2*. In Txurdinaga station, GEP model with $FV = 0.540$ provides closer variance to the variance of the observed values than the MLR. In overall, the MLR performs better than the GEP model in three stations in forecasting three-hour ahead ozone levels.

In the case of $O_3(t+4)$ forecast, the lowest values of *NMSE* and *FB* were obtained with the GEP model, being lower than those of the MLR model in Deusto and Txurdinaga stations. Also, the *R* and *FA2* values of the GEP model are higher than those of the MLR in all stations. For the Elorrieta and Mazarredo stations, however, the MLR model has lower *NMSE* and *FB* than the GEP model.

Different trends were seen in the cases of $O_3(t+5)$ and $O_3(t+6)$ forecasts, In the case of five-hour ahead ozone level forecast, the highest *R* and *FA2* values were obtained by the MLR model in Deusto, Elorrieta and Mazarredo stations. GEP model seems to be better than the MLR in only Txurdinaga station. In the case of six-hour ahead ozone level forecast, the MLR model has the lowest *NMSE*, *FB* and *FV* values and the highest *R* value in Deusto and Elorrieta stations. For

Table 15. Mathematical expressions of GEP model in Deusto station.

Predicted	Mathematical expression of the model
$O_3(t+1)$	$O_3(t)-\sqrt{O_3(t)}-\arctan\left[4.96353-O_3(t-2)-O_3(t)\times RAD\times TEM\right]+$ $+\arctan\left[RAD^{1/8}\left(O_3(t-6)+TEM\right)\left(3.8844+GRAD-V_x\right)^3\right]-$ $\sqrt{O_3(t-2)+\cos\left[O_3(t-2)\right]}-KH\,\hat{}\,\left(\left(O_3(t-4)\times O_3(t-5)-RAD^3\right)\right)+$ $6.57785\times KH+RAD$
$O_3(t+2)$	$O_3(t)+0.162622\times\left(-O_3(t-1)+TEM-1.96857\times V_y+\left(O_3(t)/\arctan\left[OP\right]\right)\right)+$ $0.279741\times\left(-O_3(t)+\cos\left[\left(-O_3(t)+RAD\right)^3\arctan\left[GRAD\times V_y\right]\right]\right)-$ $\arctan\left[\left(-7.43402+TEM\right)\times V_x+\cos\left(Exp\left[O_3(t-2)\right]\right)\right]^3$
$O_3(t+3)$	$2.50568+10\,\hat{}\left(\sin\left[\left(O_3(t)+\sin\left[O_3(t)\times\left(O_3(t-1)-RAD\right)\right]\right)^{1/3}\right]\right)+$ $10\,\hat{}\left(\arctan\left[\sin\left[\sin\left[O_3(t-1)^{1/3}\right]\right]\right]\right)+10\,\hat{}\left(\sin\left[10^{OP^{(-2.50568)}}\right]^{O_3(t-2)}\right)+$ $+KH^{1/3}\times O_3(t)-V_x$
$O_3(t+4)$	$\sqrt{HUM}+\sqrt{O_3(t-6)}-V_x+\cos\left[3.1778-\tan\left[V_x\right]\right]+$ $\cos\left[12.6476+KH-\sin\left[O_3(t)^{1/3}\right]-\tan\left[KH\right]\right]^{V_y}$
$O_3(t+5)$	$TEM+\sqrt{O_3(t)\times\arctan\left[O_3(t)\times\arccos\left[\sin\left[\tan\left[\sin\left[HUM\right]\right]\right]\right]\right]}+\sqrt{10^{\cos[GRAD]}+O_3(t)}+$ $9.93207-V_y+\sqrt{\sqrt{1-KH^2}+O_3(t-6)}+\sin\left[O_3(t-3)^3-\arccos\left[OP\right]-\sin\left[HUM\right]\right]+$ $TEM\left(\cos\left[Exp\left[0.141502V_x\right]-\frac{1}{\sqrt{1+O_3(t-1)^2}}\right]^2\right)$
$O_3(t+6)$	$\sqrt{O_3(t)}+6.04062+arc\csc\left[\left(NV+\cos\left[0.179027HUM\right]\right)^{1/3}\right]+TEM-V_y+$ $\sqrt{O_3(t-6)}+TEM\times\sin\left[\sin\left[OP^{GRAD}\right]\right]+\cos\left[\arctan\left[0.512497OP+V_y\right]^3\right]+$ $\cos\left[\left(\frac{1}{HUM^{4.71314}}\right)^{RAD}+V_x+\cos\left[RAD\right]\right]$

Table 16. Mathematical expressions of GEP model in Elorrieta station.

Predicted	Mathematical expression of the model
$O_3(t+1)$	$O_3(t)+RAD-0.146123\times\left(O_3(t)+\cos\left[NV^{RAD}\times O_3(t-5)\right]\right)-V_y+$ $RAD+Ln\left[KH^2\right]+Ln\left[KH\left(10^{GRAD}+Exp\left[V_x\right]+2HUM+9.69299V_y\right)\right]+$ $\sin\left[3RAD-V_x\right]+\left(23.3266/O_3(t-2)\right)$
$O_3(t+2)$	$8.31232-V_y+Ln\left[KH\right]^3+\left(HUM-\sqrt{O_3(t-2)(2OP-GRAD)}+TEM\right)^{1/3}+$ $\left(O_3(t)/\arctan\left[O_3(t-3)\right]\right)+\arctan\left[1.15851+GRAD+\cos\left[\cos\left[O_3(t-4)\right]\right]\right]$
$O_3(t+3)$	$Ln\left[\frac{(HUM+V_y)\arctan\left[OP\right]^3}{O_3(t)^2}\right]^2+\left(\frac{\left(NV+O_3(t-4)^2\times O_3(t-5)+O_3(t-6)^3\right)^{1/3}}{O_3(t-2)}\right)+$ $10\,\hat{}\left(\left(O_3(t)^3-(TEM-4.9621)V_x\right)^{0.037037}\right)+Ln\left[\arctan\left[\arctan\left[\arctan\left[KH\right]\right]\right]\right]^3$
$O_3(t+4)$	$V_x-2\times V_y+\sqrt{O_3(t-6)}-\sqrt{O_3(t-1)}+\arccos\left[\sin\left[GRAD+RAD-\cos\left[V_y\right]\right]\right]$ $-\sqrt{KH\times RAD\times O_3(t-3)}+RAD+16.3994+O_3(t)\times\arctan\left[KH\right]-$ $\cos\left[\tan\left[O_3(t-3)\right]\right]+\sin\left[OP\right]-\sqrt{4.66785-V_y}$
$O_3(t+5)$	$\sqrt{O_3(t)-\arctan\left[V_x\right]}+\sqrt{O_3(t-6)}-V_y+\arctan\left[GRAD\times V_x\right]-\tan\left[\cos\left[TEM\right]\right]+7.51242+$ $\sqrt{O_3(t)-Exp\left[\arctan\left[\left(-7.80743+O_3(t)\right)^3\times TEM^3\times arc\csc\left[O_3(t-6)\right]^3\right]\right]}+$ $\arccos\left[\sin\left[V_x\left(\sqrt{\cos\left[\sin\left[3.59003\times O_3(t-1)\right]\right]}-RAD\right)\right]\right]-$ $\arctan\left[HUM-NV+0.959076KH\times O_3(t-2)\right]+$ $\sqrt{O_3(t)-OP-\left(-4.11584+\arctan\left[O_3(t)\right]\right)^3}$
$O_3(t+6)$	$\sqrt{O_3(t-6)}-1.98069\arctan\left[V_y\right]+\left(HUM+O_3(t)\right)\times arc\csc\left[0.108378\left(HUM-arc\csc\left[OP\right]\right)\right]-$ $+Log\left[\left(O_3(t-5)+\left(NV+\left(O_3(t)\times O_3(t-5)\right)^{-3.62079+V_x}\right)^{1/3}\right)^3\right]+arc\csc\left[O_3(t-2)\right]+$ $\arccos\left[\sin\left[\cos\left[O_3(t)\right]\right]\right]-\sin\left[V_x\right]+6.16837\tan\left[KH\right]-$ $\arctan\left[O_3(t-2)\times V_x\right]+\cos\left[\tan\left[\sqrt{\arctan\left[O_3(t-3)\right]}\right]\right]$

Table 17. Mathematical expressions of MLR model in Deusto station.

Coefficient	O_3(t+1)	O_3(t+2)	O_3(t+3)	O_3(t+4)	O_3(t+5)	O_3(t+6)
A	1.5400	1.2400	2.0200	6.1100	11.1000	15.6000
B	0.9240	0.7640	0.6000	0.4590	0.3350	0.2310
C	–0.1120	–0.1390	–0.1290	–0.1150	–0.1010	–0.0522
D	–0.0318	–0.0389	–0.0439	–0.0459	–0.0100	–0.0120
E	–0.0086	–0.0194	–0.0272	0.0028	–0.0044	0.0271
F	–0.0135	–0.0237	0.0067	–0.0021	0.0296	–0.0074
G	–0.0120	0.0116	–0.0031	0.0192	–0.0227	0.0002
H	0.0524	0.0732	0.0938	0.0841	0.1170	0.1250
I	–0.1520	–0.4290	–0.6680	–0.8420	–0.8850	–0.8010
J	–0.3870	–0.7840	–0.9210	–0.9310	–0.9530	–0.8690
K	4.4700	6.2800	4.8300	1.9700	–0.5300	–2.8600
L	0.0217	0.0844	0.1650	0.2690	0.3730	0.4580
M	0.0056	0.0350	0.0412	0.0333	0.0296	0.0258
N	0.0014	0.0545	0.0955	0.1460	0.1780	0.1960
O	–0.0092	–0.0145	–0.0239	–0.0249	–0.0223	–0.0207
P	0.4060	0.6760	1.2300	1.2400	1.0300	0.8770
Q	11.5000	19.1000	26.7000	26.0000	18.1000	9.7500

the Mazarredo station, however, GEP model has a better accuracy than the MLR with respect to *NMSE*, *R*, *FA2*, *FB* and *FV* statistics. In the case of $O_3(t+1)$ forecast, GEP model performs significantly better than the MLR in Txurdinaga station from the *NMSE*, *FA2* and *FB* viewpoints. In the case of $O_3(t+2)$ forecast, significant differences between GEP and MLR models are seen for the Elorrieta station. In Elorrieta, GEP model considerably performs better than the MLR from the *NMSE*, *R*, *FB* and *FV* viewpoints. In the case of $O_3(t+3)$ forecast, the MLR shows significantly better accuracy than the GEP model in the Txurdinaga station from the *NMSE* and *FB* viewpoints. In the case of $O_3(t+4)$ forecast, there is significant differences between GEP and MLR models in Elorrieta and Mazarredo stations with respect to *FV* statistics. In the case of $O_3(t+5)$ forecast, the GEP model considerably performs better than the MLR model in the Elorrieta and Mazarredo stations from the *FV* viewpoint. In the case of $O_3(t+6)$ forecast, there is significant difference between GEP and MLR models in Mazarredo and Txurdinaga stations with respect to *FV* criterion.

Table 18. Mathematical expressions of MLR model in Elorrieta station.

Coefficient	O_3(t+1)	O_3(t+2)	O_3(t+3)	O_3(t+4)	O_3(t+5)	O_3(t+6)
A	3.6200	6.5100	10.5000	16.6000	25.1000	32.9000
B	0.8710	0.6910	0.5380	0.3920	0.2670	0.1900
C	–0.0699	–0.0683	–0.0783	–0.0745	–0.0440	–0.0196
D	–0.0098	–0.0354	–0.0443	–0.0258	–0.0071	–0.0258
E	–0.0270	–0.0374	–0.0187	–0.0027	–0.0213	0.0092
F	–0.0116	0.0025	0.0136	–0.0085	0.0188	0.0055
G	0.0153	0.0245	0.0007	0.0257	0.0097	0.0549
H	0.0248	0.0372	0.0696	0.0736	0.0945	0.0638
I	0.0180	–0.0090	–0.059	–0.0380	–0.0440	0.0570
J	–0.5850	–1.0500	–1.1100	–1.0300	–0.8420	–0.8300
K	3.5500	4.4100	2.6700	0.0100	–2.4700	–3.7700
L	–0.0415	–0.0529	–0.0656	–0.4780	–0.0269	0.0292
M	–0.0114	–0.0274	–0.0624	–0.1130	–0.1660	–0.1870
N	0.0435	0.0701	0.0667	0.0455	0.0153	0.0063
O	–0.0074	–0.0146	–0.0244	–0.0284	–0.0239	–0.0139
P	0.2970	0.6780	1.2400	1.5100	1.2500	0.6680
Q	11.7000	22.9000	33.6000	37.3000	32.1000	17.7000

Figs. 2–5 demonstrate the scatter plots of one-, two-, …, six-hour ahead forecasts and observed ozone level values for the test period for Deusto, Elorrieta, Mazarredo and Txurdinaga stations, respectively. Significantly overestimations are clearly seen for the MLR model in Txurdinaga station in the case of $O_3(t+1)$ forecast. Increasing forecast horizon considerably decreases models accuracy. Both GEP and MLR models significantly overestimate low values and underestimate high values in two-, three-, four-, five-, and six-hour ahead forecasting cases. As can be clearly seen from the Figs. 2–5, too much scattered estimates were obtained from the both models in the case of four-, five- and six-hour ahead ozone level predictions.

One of the advantages of GEP in comparison with other soft computing techniques is producing analytical formula for determination of output parameter. Tabs. 15 and 16 summarize the GEP mathematical equations for Deusto and Elorrieta stations. In these tables, $O_3(t)$ and $O_3(t-k)$, $k = 1, 2, \ldots, 6$ are the current and past data of ozone levels.

Table 19. Mathematical expressions of MLR model in Mazarredo station.

Coefficient	O_3(t+1)	O_3(t+2)	O_3(t+3)	O_3(t+4)	O_3(t+5)	O_3(t+6)
A	–14.0000	–20.0000	–22.6000	–21.3000	–13.5000	–6.0700
B	0.9300	0.7660	0.6220	0.4320	0.3150	0.2230
C	–0.0937	–0.0931	–0.1540	–0.1060	–0.0913	–0.0795
D	–0.0026	–0.0770	–0.0373	–0.0323	–0.0305	0.0089
E	–0.0761	–0.0358	–0.0350	–0.0341	0.0075	–0.0436
F	0.0404	0.0331	0.0221	0.0486	–0.0168	–0.0193
G	–0.0060	–0.0096	0.0187	–0.0399	–0.0318	0.0481
H	0.0214	0.0489	0.0437	0.0951	0.1360	0.0948
I	0.0470	0.1750	0.2690	0.3480	0.6070	0.8000
J	–0.7180	–0.8970	–0.9400	–0.8630	–0.8120	–0.9530
K	7.6800	12.2000	13.5000	13.3000	9.2200	7.3900
L	0.2160	0.3850	0.5430	0.6800	0.8430	0.9820
M	0.1340	0.2070	0.2540	0.2810	0.2690	0.2710
N	0.3330	0.5190	0.6450	0.7480	0.7300	0.7380
O	–0.0103	–0.0191	–0.0286	–0.0390	–0.0362	–0.0350
P	0.6880	1.2000	1.7000	2.1400	1.7400	1.3800
Q	15.9000	27.9000	36.7000	39.5000	29.6000	17.0000

Also, mathematical equations of MLR for prediction of $O_3(t+k)$ ($k=1,2,\ldots,6$) in all stations are presented in Tabs. 17–20. It is necessary to note that typical equation for MLR is:

$$\begin{aligned} O_3(t+k) &= A + B\times O_3(t) + C\times O_3(t-1) + D\times O_3(t-2) \\ &+ E\times O_3(t-3) + F\times O_3(t-4) + G\times O_3(t-5) + H\times O_3(t-6) \\ &+ I\times V_x + J\times V_y + K\times RAD + L\times TEM + M\times HUM \\ &+ N\times GRAD + O\times NV + P\times OP + Q\times KH \end{aligned} \tag{12}$$

6. Conclusion

The management of ozone control and public protection activities requires accurate forecasts. Although many ozone prediction models have been developed and some of them are in use, there is a pressing need for accurate models capable of determining the relative importance of environmental variables. Therefore,

Table 20. Mathematical expressions of MLR model in Txurdinaga station.

Coefficient	O_3(t+1)	O_3(t+2)	O_3(t+3)	O_3(t+4)	O_3(t+5)	O_3(t+6)
A	–6.4300	–9.3200	–8.1100	–2.0800	7.3100	16.5000
B	0.9910	0.7970	0.6420	0.4530	0.3300	0.2420
C	–0.1960	–0.1580	–0.1970	–0.1370	–0.0971	–0.0718
D	0.0410	–0.0334	0.0031	0.0077	0.0086	0.0072
E	–0.0749	–0.0333	–0.0242	–0.0121	–0.0094	–0.0281
F	0.0443	0.0420	0.0464	0.0346	0.0029	0.0164
G	–0.0018	0.0079	–0.0031	–0.0275	–0.0117	–0.0182
H	0.0126	0.0151	0.0264	0.0578	0.0669	0.0835
I	0.0350	–0.0410	–0.1280	–0.0070	0.3010	0.5540
J	–0.5770	–0.0814	–0.9720	–1.0600	–1.2900	–1.6100
K	5.3600	7.6500	7.4900	5.0400	0.6600	–1.0300
L	0.0025	–0.0070	0.0150	0.0790	0.2040	0.3180
M	0.0754	0.1250	0.1410	0.1320	0.1020	0.0802
N	0.2040	0.3530	0.3740	0.3460	0.2420	0.1590
O	–0.0156	–0.0293	–0.0366	–0.0331	–0.0234	–0.0140
P	0.9650	1.7900	2.2400	1.9600	1.2200	0.4090
Q	16.1000	30.7000	37.8000	34.6000	23.3000	8.7800

an ozone forecasting system using gene expression programming and multiple linear regression were developed to predict hourly concentrations in Bilbao area, Spain. The system forecasts ozone levels in the near future are based on the current data of meteorological parameters and past data of ozone levels. A study of the values obtained from the statistics of the model validation kit showed that gene expression programming-based models performed better than the multiple linear regression method. The proposed gene expression programming model possesses some merits. Firstly, it can provide better predicting results with explicit mathematical formulation. Secondly, the model is extensible and reproducible. It can be used in the areas with similar environmental features so that the expenses can be reduced as well. At the end, we can conclude that the gene expression programming can be used in modelling and predicting the ground ozone levels. Clearly, this study has indicated the potential of the gene expression programming method for capturing the non-linear interactions between ozone and other factors and for the identification of the relative importance of these factors.

References

Abdul-Wahab, S. A. and Al-Alawi, S. M. (2002): Assessment and prediction of tropospheric ozone concentration levels using artificial neural networks, *Environ. Modell. Softw.*, **17**, 219–228.

Agirre-Basurko, E., Ibarra-Berastegi, G. and Madariaga, I. (2006): Regression and multilayer perceptron-based models to forecast hourly O3 and NO2 levels in the Bilbao area, *Environ. Modell. Softw.*, **21**, 430–446.

Al-Alawi, S. M., Abdul-Wahab, S. A., Bakheit, C.cS. (2008): Combining principal component regression and artificial neural networks for more accurate predictions of ground-level ozone, *Environ. Modell. Softw.*, **23**, 396–403.

Balaguer, E., Camps-Valls, G., Carrasco-Rodriguez, J. L., Soria Olivas, E., del Valle-Tascon, S. (2002): Effective 1-day ahead prediction of hourly surface ozone concentrations in Eastern Spain using linear models and neural networks, *Ecol. Model.*, **156**, 27–41.

Ballester, E. B., Valls, G. C., Carrasco-Rodriguez, J. L., Soria Olivas, E. and Valle-Tascon, S. L. (2002): Effective 1-day ahead prediction of hourly surface ozone concentrations in eastern Spain using linear models and neural networks, *Ecol. Model.,* **156**, 27–41.

Banzhaf, W., Nordin, P., Keller, R. E., Francone, F. D. (1998): Genetic programming:An introduction: On the automatic evolution of computer programs and its applications. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 470 pp, ISBN: 1-55860-510-X.

Baur, D., Saisana, M. and Schulze, N. (2004): Modelling the effects of meteorological variables on ozone concentration e a quantile regression approach, *Atmos. Environ.*, **38**, 4689–4699.

Borrego, C., Tchepel, O., Barros, N., and Miranda, A. I. (2000): Impact of road traffic emissions on air quality of the Lisbon region, *Atmos. Environ.*, **34**, 4683–4690.

Chaloulakou, A., Assimakopoulos, D. and Kekkas, T. (2003): Forecasting daily maximum ozone concentrations in the Athens Basin, *Environ. Monit. Assess.*, **56**, 97–112.

Chen, J. L., Islam, S. and Biswas, P. (1998): Nonlinear dynamics of hourly ozone concentrations: nonparametric short term prediction, *Atmos. Environ.*, **32**, 1839–1848.

Cobourn, W. G. and Hubbard, M. C. (1999): An enhanced ozone forecasting model using air mass trajectory analysis, *Atmos. Environ.,* **33**, 4663–4674.

Comrie, A. C. (1997): Comparing neural networks and regression models for ozone forecasting, *J. Air Waste Manage.*, **47**, 653–663.

Dudot, A. L., Rynkiewicz, J., Steiner, F. E. and Rude, J. (2007): A 24-h forecast of ozone peaks and exceedance levels using neural classifiers and weather predictions, *Environ. Modell. Softw.*, **22**, 1261–1269.

European Commission (1994): The evaluation of models of heavy gas dispersion. Model evaluation group seminar. Office for official publications of the European communities, L-2985, Luxemburg.

Feng, Y., Zhang, W., Sun, D. and Zhang, L. (2011): Ozone-concentration forecast method based on genetic algorithm optimized back propagation neural networks and support vector machine data classification, *Atmos. Environ.,* **45**, 1979–1985.

Ferreira, C. (2001a): Gene expression programming in problem solving. *6th Online World Conf. on Soft Computing in Industrial Applications* (invited tutorial), 22 pp, accessed in January 2013 at http://www.gene-expression-programming.com/webpapers/GEPtutorial.pdf

Ferreira, C. (2001b): Gene expression programming: A new adaptive algorithm for solving problems, *Compl. Sys.*, **13**, 87–129.

Ferreira, C. (2006): *Gene expression programming: Mathematical modeling by an artificial intelligence*. Springer, Berlin, 478 pp.

Fuchs, M. (1998): Crossover versus mutation: An empirical and theoretical case study, in *Proceedings of the Third Annual Genetic Programming Conference*, Morgan-Kauffman, San Mateo, CA, USA, 78–85.

Gardner, M. W. and Dorling, S. R. (1998): Artificial neural networks (the multilayer perceptron) – A review of applications in the atmospheric sciences, *Atmos. Environ.,* **32**, 2627–2636.

Gardner, M. W. and Dorling, S. R. (2000): Statistical surface ozone models: an improved methodology to account for non-linear behavior, *Atmos. Environ.,* **34**, 21–34.

Goldberg, D. E. (1989): *Genetic algorithms in search, optimization, and machine learning*. Addison-Wesley, Reading, MA 412 pp.

Hanna, S. R., Strimaitis, D. G. and Chang, J. C. (1991): *Hazard response modeling uncertainty (a quantitative method)*. User's Guide for Software for Evaluating Hazardous Gas Dispersion Models. American Petroleum Institute, Washington, 334 pp.

Hashmi, M. Z., Shamseldin, A. Y. and Melville, B. W. (2011): Statistical downscaling of watershed precipitation using Gene Expression Programming (GEP), *Environ. Modell. Softw.*, **26**, 1639–1646.

Hubbard, M. and Cobourn, G. (1998): Development of a regression model to forecast ground-level ozone concentration in Louisville, KY, *Atmos. Environ.,* **32**, 2637–2647.

Ibarra-Berastegi, G., Elias, A., Agirre, E. and Uria, J. (2001a): Short-term, real-time forecasting of hourly ozone, NO2 and NO levels by means of multiple linear regression modeling, Gate to EHS, pp. 1–7, DOI: http://dx.doi.org/10.1065/ehs2001.06.009.

Ibarra- Berastegi, G., Madariaga, I., Elias, A., Agirre, E. and Uria, J. (2001b): Long-term changes of ozone and traffic in Bilbao, *Atmos. Environ.,* **35**, 5581–5592.

Kolehmainen, M., Martikainen, H. and Ruuskanen, J. (2001): Neural networks and periodic components used in air quality forecasting, *Atmos. Environ.,* **35**, 815–825.

Koza, J. R. (1992): *Genetic programming: On the programming of computers by means of natural selection*. MIT Press, Cambridge, MA, 819 pp, ISBN 0-262-11170-5.

Lu, W. Z., Fan, H. Y., Leung, A. Y. T. and Wong, J. C. K. (2002): Analysis of pollutant levels in central Hong Kong applying neural network method with particle swarm optimization, *Environ. Monit. Assess.*, **79**, 217–230.

Lu, Y., Luo, X. and Zhang, H. (2011): A gene expression programming algorithm for highway construction cost prediction problems, *J. Transp. Sys. Eng. Inf. Technol.,* **11**(6), 85–92.

Luke, S. and Spector, L. (1998): A revised comparison of crossover and mutation in genetic programming. In *Proceeding of the Third Annual Genetic Programming Conference*, edited by Koza, J. R., Banzhaf, W., Chellapilla, K., Deb, K., Dorigo, M., Fogel, D. B., Garzon, M. H., Goldberg, D. E., Iba, H., and Riolo, R., Morgan-Kauffman, Madison, San Mateo, CA, USA, 208–213.

Mayer, H. (1999): Air pollution in cities, *Atmos. Environ.,* **33**, 4029–4037.

Omidvari, M., Hassanzadeh, S. and Hosseinibalam, F. (2008): Time series analysis of ozone data in Isfahan, *Physica A*, **387**, 4393–4403.

Ordieres, J. B., Vergara, E. P., Capuz, R. S. and Salazar, R. E. (2005): Neural network prediction model for fine particulate matter (PM 2.5) on the US–Mexico border in El Paso (Texas) and Ciudad Juarez (Chihuahua), *Environ. Modell. Softw.*, **20**, 547–559.

Ortiz-García, E. G., Salcedo-Sanz, S., Pérez-Bellido, Á. M., Portilla-Figueras, J. A. and Prieto, L. (2010): Prediction of hourly O3 concentrations using support vector regression algorithms, *Atmos. Environ.,* **44**, 4481–4488.

Özcan, F. (2012): Gene expression programming based formulations for splitting tensile strength of concrete, *Constr. Build. Mater.*, **26**, 404–410.

Pires, J. C. M., Alvim-Ferraz, M. C. M., Pereira, M. C. and Martins, F. G. (2011): Prediction of tropospheric ozone concentrations: application of a methodology based on the Darwin's theory of evolution, *Expert. Syst. Appl.*, **38**, 1903–1908.

Pires, J. C. M., Alvim-Ferraz, M. C. M., Pereira, M. C. and Martins, F. G. (2010): Evolutionary procedure based model to predict ground-level ozone concentrations, *Atmos. Pollut. Res.*, **1**, 215–219.

Pires, J. C. M. and Martins, F.G. (2011) Correction methods for statistical models in tropospheric ozone forecasting, *Atmos. Environ.,* **45**, 2413–2417.

Pires, J. C. M., Martins, F. G., Sousa, S. I. V., Alvim-Ferraz, M. C. M. and Pereira, M. C. (2008): Selection and validation of parameters in multiple linear and principal component regressions, *Environ. Modell. Softw.*, **23**, 50–55.

Prybutok, V. R., Yi, J. and Mitchell, D. (2000): Comparison of neural network models with ARIMA and regression models for prediction of Houston's daily maximum ozone concentrations, *Eur. J. Oper. Res.*, **122**, 31–40.

Robeson, S. M. and Steyn, D. G. (1990): Evaluation and comparison of statistical forecast models for daily maximum ozone concentrations, *Atmos. Environ.,* **2**, 303–312.

Ruiz-Suarez, J. C., Mayora-Ibarra, O. A., Torres-Jimenez, J. (1995): Short-term ozone forecasting by artificial neural networks, *Adv. Eng. Softw.*, **23**, 143–149.

Schlink, U., Herbarth, O., Richter, M., Dorling, S., Nunnari, G., Cawley, G. and Pelikan, E. (2006): Statistical models to assess the health effects and to forecast ground-level ozone, *Environ. Modell. Softw.*, **21**, 547–558.

Sousa, S. I. V., Martins, F. G., Alvim-Ferraz, M. C. M. and Pereira, M. C. (2007): Multiple linear regression and artificial neural network based on principal components to predict ozone concentrations, *Environ. Modell. Softw.*, **22**, 97–103.

Tsai, C. H., Chang, L. C. and Chiang, H. C. (2009): Forecasting of ozone episode days by cost-sensitive neural network methods, *Sci. Total. Environ.*, **407**, 2124–2135.

Wang, W., Lu, W., Wang, X. and Leung, A. Y. T. (2003): Prediction of maximum daily ozone level using combined neural network and statistical characteristics, *Environ. Int.*, **29**, 555–562.

Yi, J. S. and Prybutok, V. R (1996): A neural network model forecasting for prediction of daily maximum ozone concentration in an industrialized urban area, *Environ. Pollut.*, **92**, 349–357.

Zolghadri, A., Monsion, M., Henry, D., Marchionini, C. and Petrique, O. (2004): Development of an operational model-based warning system for tropospheric ozone concentrations in Bordeaux, France, *Environ. Modell. Softw.*, **19**, 369–382.

SAŽETAK

Usporedna analiza modela za prognozu koncentracija ozona pomoću evolucijskog programiranja gena i višestruke linearne regresije

Saeed Samadianfard, Reza Delirhasannia, Ozgur Kisi i Elena Agirre-Basurko

Zbog štetnog utjecaja na dišni sustav prizemni ozon (O_3) već nekoliko desetljeća predstavlja ozbiljan problem u mnogim onečišćenim urbanim područjima. Kako bi se smanjili rizici od oštećenja uzrokovanih ozonom, potrebno je razvijati, održavati i poboljšavati modele kratkoročne prognoze ozona. Ovaj rad prikazuje rezultate dvaju prognostičkih modela, evolucijskog programiranja gena (GEP), koje je varijanta genetskog programiranja (GP), te prognoziranje razina ozona u realnom vremenu višestrukom linearnom regresijom (MLR) do šest sati unaprijed na četiri postaje u Bilbau u Španjolskoj. Ulazni podaci za GEP su meteorološki uvjeti (brzina i smjer vjetra, temperatura, relativna vlažnost zraka, tlak, sunčevo zračenje i termički gradijent), satne razine ozona i parametri prometa (broj vozila, udio vremena zauzetosti ceste vozilima i njihova brzina), koji su izmjereni u razdoblju 1993–1994. Performanse razvijenih modela ocijenjene su usporedbom s mjerenjima te upotrebom alata za validaciju modela koje je predložila američka Agencija za zaštitu okoliša. Utvrđeno je da GEP u većini slučajeva daje bolje prognoze. Na kraju je zaključeno da je evolucijsko programiranje gena obećavajuća tehnika za prognozu koncentracija onečišćujućih tvari.

Ključne riječi: modeliranje kvalitete zraka, evolucijsko programiranje gena, višestruka linearna regresija, prognoziranje razina ozona, područje Bilbaa, Španjolska

Corresponding author's address: Saeed Samadianfard, tel: +98 91 41 101 845, e-mail: s.samadian@Tabrizu.ac.ir