Journal of Environmental Informatics 36(2) 93-106 (2020)

Journal of
Environmental
Informatics

www.iseis.org/jei

Predicting Ground Level Ozone in Marrakesh by Machine-Learning Techniques

J. Ordieres-Meré[1] [*], J. Ouarzazi[2], B. El Johra[3], and B. Gong[4]

[1] *Escuela Técnica Superior de Ingenieros Industriales, Universidad Politécnica de Madrid, Madrid 28006, Spain*
[2] *CNRST-URAC 20 Faculté des Sciences Semlalia Marrakech, Université Cadi Ayyad, Marrakech 40001, Morocco*
[3] *Direction de la Météorologie Nationale Casablanca 20240, Morocco*
[4] *Jülich Supercomputing Center Forschungszentrum Jülich GmbH, Jülich 52425, Germany*

Received 07 January 2016; revised 22 September 2016; accepted 27 January 2020; published online 13 July 2020

ABSTRACT. This study was undertaken to produce local, short-term, artificial intelligence-based models that estimate the ozone level with special attention to the relationship between diurnal and nocturnal ozone variations of some primary pollutants and meteorologyical parameters in the city of Marrakesh, Morocco. Hourly data has been collected from the three air-quality monitoring stations in the city. This paper seeks to analyze the main factors that are associated with ozone formation, including the generation of different daytime and nighttime scenarios. The present work extends existing publications about the region by developing ozone prediction models from meteorological variables and primary pollutants. Several experiments were conducted to verify properties of the produced models, thus making it possible not only to describe but also to predict ozone pollution in this geographical area. The findings facilitate 48 hour forecasts that have root mean square errors as low as 20 g/m^3. Our results highlight the importance of using such models for civil applications.

Keywords: ozone pollution, ozone diurnal concentration, ozone nocturnal concentration, machine learning, nonlinear models, Marrakesh

1. Introduction

Ozone is a secondary pollutant that is formed from chemical reactions of primary pollutants, such as nitrogen dioxide (NO_2) (Seinfeld, 1989; Derwent et al., 1998; Jenkin and Clemitshaw, 2000), nitrogen oxide (NO), and volatile organic compounds (VOCs) under certain weather conditions (Duan et al., 2008). High wind speeds, high temperatures, and low levels of relative humidity contribute to ozone formation (Ambroise and Grandvalet, 2001; Khatibi et al., 2013). The hydroxyl radical (OH) is one of the key factors in photochemical cycles that are responsible for ozone formation. This pollutant is recognized as being a significant contributor to global warming because of its positive radiative force (Myhre et al., 2013).

According to the United States Environmental Protection Agency (EPA), ozone is one of the six common air pollutants (known as “critical pollutants”) that have a strong and harmful effect on public health (Salazar-Ruiz et al., 2008; Khatibi et al., 2013). These chemical pollutants may even result in the death of people who are particularly sensitive to them (Matus et al., 2008; Matus et al., 2012). In addition, a high level of ozone can cause damage to agriculture, decrease crop yields, and lead to a great economic loss (Avnery et al., 2011). Therefore, predicting ozone accurately and timely has become an important issue. It has drawn the attention of researchers and related authorities to the need for alerting the public about the risk of exposure to high levels of the pollutant (Zhang et al., 2013; Riga et al., 2015).

Severe air pollution episodes could be intensified by certain weather conditions that are favorable for a weak dispersion of atmospheric pollutants, thereby impacting public health. This leads to elevated photochemical activities that are favorable to the production of ozone (Chan et al., 1998; Kommalapati et al., 2016). Strong radiation, abundant traffic, and calm winds combine to provide ideal conditions for the production and the accumulation of a high level of ozone in the city and its suburban areas. The wind transports the pollutants and influences the turbulence regime that is associated with their diffusion (Lin et al., 2012). Consequently, an association of meteorological factors with the observed rates of pollutants is essential to understand the behavior of tropospheric pollutants in urban areas (Lee et al., 2007; Salazar-Ruiz et al., 2008).

Ozone production depends on solar radiation, therefore, the daytime profile differs from the nighttime profile. Accordingly, we decided to separate the daytime and the nighttime data, taking account of the chemical reactions in the two periods and evaluating the importance of each parameter’s influence on the measured ozone concentration during the day and the night.

The aim of this paper is to provide an analysis that tran-

[*] Corresponding author. Tel.: +(34) 910677107; fax: +(34) 913363005.
E-mail address: j.ordieres@upm.es (J. Ordieres-Meré).

SSN: 1726-2135 print/1684-8799 online
doi:10.3808/jei.202000437

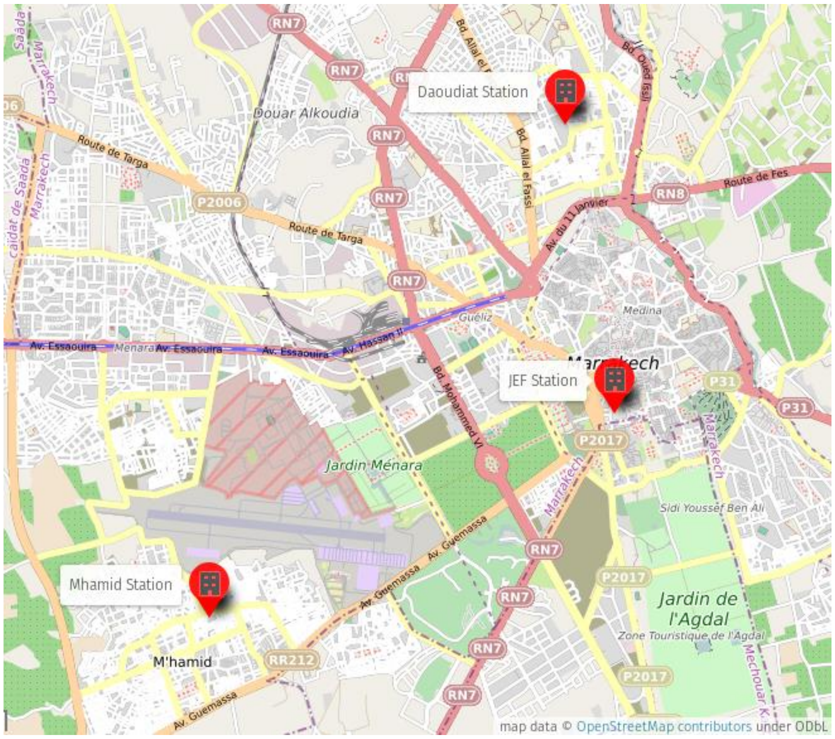

Figure 1. Surface of the pollutant-monitoring sites in Marrakesh (Courtesy of ©OpenStreetMap contributors under CC BY-SA license [https://www.openstreetmap.org/copyright]).

scends a descriptive approach and to contribute to the implementtation of active strategies when ozone is the targeted pollutant. Therefore, the expected contribution will be to demonstrate the capability of forecasting ozone levels several hours in advance by the use of advanced nonlinear models. It will be demonstrated through the analysis in a specific and relevant city in Morocco. Additionally, this will make it possible to implement mitigation policies and to identify differentiated mechanisms for ozone formation and, therefore, to understand the phenomenon better.

2. Study Area

2.1. Characteristics of the City

The more developed areas in Morocco, like most metropolitan areas of emerging countries, have experienced great urban, social, and industrial development during the last twenty years. They have experienced a significant increase in population, number of motor vehicles and industrial activities. This growth has generated jobs for hundreds of thousands of families; however, it has also contributed to the degradation of air quality in the region (Ouarzazi et al., 2003).

Although the importance of the problem provoked an official declaration from the government of Morocco, it is useful to show that a limited number of scientific analyses have been conducted. In all of the cases, the existing analyses were descriptive (Ouarzazi et al., 2003; Inchaouh et al., 2017).

The City of Marrakesh is situated at latitude of (31.54° N ~ 31.69° N) and longitude of (8° W ~ 7.84° W). It has a population of 1,070,000 inhabitants in an area of 230 km^2. Because of its distance from the Atlantic coast, Marrakesh has an arid climate characterized by strong seasonal and diurnal temperature variations. The city and the surrounding plains are affected by a major rainfall deficiency. The prevailing winds which blow through the area most of the year are from the Northwest and the West and are relatively calm. In contrast, the Chergui and Sirocco winds (observed during the summer) blow eastward and southward, respectively (Inchaouh et al., 2017). The pollutants that originate from industrial zones in the North and Northwest of Marrakesh disperse to the other parts of the city.

The City of Marrakesh receives a high level of solar radiation throughout the year (Sinha et al., 2014). However, there is no available information on the tropospheric ozone concentrations as only one descriptive study of air quality in Marrakesh has been conducted (Ouarzazi et al., 2003). Three permanent monitoring stations have been in operation, as presented in Figure 1. Although they have been operating only since 2009, huge effects of tropospheric ozone are foreseen for the near future (Lei et al., 2012).

To reduce air pollution in the metropolitan area there have been continuous efforts to improve public transport and to improve road conditions. Some initiatives for such reduction of air pollution include restructuring the urban arterial network by widening roads, using road signs more extensively, using one-way streets, restricting the use of vehicles in Old Marrakesh, and relocating the fruit and vegetable wholesale market away from the center of the city.

To protect human health, the Moroccan Directive (Kingdom of Morocco, 2010) established an ozone threshold of 110 μg/m^3 (the maximum daily 8 hours average). This threshold cannot be exceeded for more than three consecutive days. The ozone concentration that was established to protect vegetation is restricted to a daily average of 65 μg/m^3. The cost of the degradation of the air in Morocco is estimated by the World Bank at 3.6 billion dirhams per year, which is approximately 1.03% of the GDP-2002. A 4% annual growth rate in the number of automotive vehicles due to the fleet renewal process, a lack of maintenance, a lack of vehicle inspection, and inadequate means of control are the causes of this degradation (Secrétariat d’Etat de l’Eau et de l’Environnement, 2010).

2.2. Characteristics of the Stations

Both Daoudiat and Jamaa El Fna stations (Jef) are located in the center of the city. In contrast, the Mhamid is located in the southwest part of the city where there is less traffic. Specific details of their locations and their monitoring start dates can be found in Table 1 and Figure 1.

The variables that are available from those stations are related to the date and the time of measurement (DT) and certain meteorological parameters: relative humidity (RH), temperature (T), wind speed (WS), solar radiation (SR), and pollution-related factors. The latter include particulate matter with a diameter less than ten microns (PM_{10}), the ozone level (O_3), nitrogen dioxide (NO_2), carbon monoxide (CO) and sulfur dioxide (SO_2), even though the last two variables are not relevant to the ozone pollution level.

Table 1. Features of the Permanent Air Quality Monitoring Stations in Marrakesh

Station	Type	Latitude (N)	Longitude (W)	Began Operation
Mhamid	Urban	31.5962	8.0436	01/06/2009
Jef	Urban	31.6202	7.9888	01/06/2009
Daoudiat	Urban	31.6536	7.9956	01/03/2010

3. Methodology

As indicated in the introduction of the paper, the expected contribution will be to demonstrate the capability of forecasting ozone levels several hours in advance by the use of advanced nonlinear models. This means that the aim is to produce good local models that permit the forecasting of the ozone level based on potentially dependent variables, and to learn how different nonlinear-based models behave depending on their characteristic. The relevance of the contribution is not just at local level, but as far as Morocco’s air quality monitoring network is constituted of 29 fixed stations and 3 mobile stations across 15 towns. Processing of millions of air quality data from the national network with the prediction methods used in this paper could provide decision-makers with a realistic view thus constituting a rigorous decision support system for better air quality. These aspects are in line with interest from local authorities looking for a more integrated approach to spatial planning. Indeed, application to other places can be easily adopted.

Therefore, the selected methodology proposes to perform a first step focused on understand the relevant physical effects capable of increasing the quality of the forecasting activities. This activity was summarized in the next subsection.

A second aspect is to develop a better understanding of collected data, in order to know their characteristics and prepare them according to the requirements from the previous step, and looking to develop the modelling step. These two steps are presented in Subsections 3.2 and 3.3 respectively.

Additionally, quality assessment and sensitivity analysis are needed to understand the characteristics of the models, as well as their generalization capabilities. The discussion for these aspects will be provided in Section 4, after presenting the main characteristics of the used models.

3.1. Temporal Patterns

The maximum amount of ozone is obtained between four and six hours after the emission of the ozone precursors and at the beginning of ozone production (Gao, 2007). This explains why the ozone peaks are obtained far from the locations where the ozone precursors originate. The ozone, therefore, is presented in larger quantities in suburban and rural areas.

Based on details present in the supplementary material, it can be stated that there are two general ozone concentration profiles in Marrakesh. The first is during the months of renewal and low solar radiation. It is characterized by symmetrical peaks and similar values on both sides of the peak but becomes nil at night. The second is during the hot season when the broader peaks and higher ozone values characterize the profiles in the evening.

Such differentiated behavior can be associated to seasons. The first one can be related to autumn and winter, and the second to spring and summer. This effect can be ascribed to the discrepancies between ozone production and solar radiation patterns and implies the existence of other factors that contribute to the orientation of this amplitude.

On the entire spectrum of measured values, the region where the average 8-hour ozone concentrations are elevated is found where there is a significant amount of road traffic. However, the dispersion of ozone prevents its accumulation in this area and transports the pollutant out of the urban center. This explains the high maxima observed in the average of the 8-hour periods in Mhamid (see Supplementary material Table 1).

3.2. Dataset

The dataset includes samples for a period of a year and a half that begins in the middle of 2009. It is based on hourly data sampling, which explains the significant number of samples (see Table 2). Most stations have meteorological variables (temperature, relative humidity, solar radiation, wind speed), and other pollutants such as CO, NO_2, and PM_{10} (Daoudiat does not maintain a record of temperatures).

This study considers daytime and nighttime behaviors difervently. Therefore, specific datasets were prepared for the daytime, nighttime and total samples of each station. The error measurements that were considered were the adjusted R^2 for model fitness and the Nash-Sutcliffe Efficiency coefficient (NSE), the root mean square error (RMSE), and relative root mean square error (RRMSE) for residual characterization and correlation of real and predicted ozone values.

Table 2. Number of Hourly Samples per Dataset and Station Selected for Experimentation

Station	Segment	Daytime Dataset	Nighttime Dataset	Complete Dataset
Daoudiat	Total	1759	1351	3110
	Training	1501	1143	2644
	Testing	258	208	466
Mhamid	Total	4601	4438	9039
	Training	3902	3782	7684
	Testing	699	656	1355
Jef	Total	4302	3811	8113
	Training	3663	3234	6897
	Testing	639	577	1216

3.3. Data Modelling

Different alternatives to build models are possible, in the understanding that previous knowledge about pollution exists and that potential dependent variables have been also collected, which allow us to consider supervised machine learning based modelling. This means the data patterns will be used to build models by adjusting their outcome to the real ozone values observed for those patterns.

Auto-correlation between dependent variables has been tested by using the variance of inflation factors (VIF), which consists of metrics for the severity of multi-collinearity in the least square regression analysis, with a threshold of 5, which gives a maximum of the variance of 3.26 for temperature. These values support the idea of a very low correlation between dependent variables, as a VIF > 10 means high collinearity, making it possible to use generalized multilinear models as a way to provide reference levels for model accuracy.

After defining a reference for learning, based on multilinear models, different models of each station were configured by considering dependent variables, such as temperature, solar radiation, relative humidity, nitrogen oxide, carbon monoxide, wind speed, solid particulates, and the hour at which they were measured. The objective is to predict the ozone level at a specified time in advance, by using the selected supervised learning approach.

In the relevant literature, there are abundant statistical studies of the ozone production and prediction using data compiled over 24 hours (Salazar-Ruiz et al., 2008; He and Lu, 2012). Advanced models using ensemble architecture and other statistical technique have been used (Gong and Ordieres-Meré, 2016; Freeman et al., 2018; Gao et al., 2018), and they have been used to reconfigure measurement networks (Gong and Ordieres-Meré, 2017). Few studies, however, have been carried out on the meteorological parameters' influence on nocturnal ozone (Kovač-Andrić et al., 2009). Most studies have emphasized the lack of a relationship between SO_2 and O_3. Therefore, SO_2 was removed from the analysis (see Figure 2, where linear correlations are shown in the upper matrix, the lower matrix plots the pairing values to provide a feeling of their relative behavior, and the diagonal shows a histogram of the individual variables and their names).

As there is no consensus about which kind of machine learning model type should produce the best results, it was decided to select several of them in order to avoid bias coming from the type of model.

In order to avoid bias because of the learning dataset, the model training uses the tenfold cross-validation strategy, which is commonly adopted as a way to improve the robustness of the learning procedure. Tools for data training and testing were produced by using the R library (Team, 2014) and some of their packages (Limas et al., 2014; Kuhn et al., 2014).

4. Results and Discussions

4.1. Data Modelling

In this subsection a short introduction to all the different model types selected, either as reference learning or for advanced learning, is presented. Therefore, it will be possible to understand what the fundamentals of each of them are. Particular adaptation to the problem being addressed is also provided when needed.

4.1.1. Linear Regression Modeling

Multiple Linear Regression (MLR) is one of the most commonly used methods to model ozone concentrations (dependent variable) according to meteorological parameters and different air pollutants (independent or explanatory variables). MLR is a model that creates a linear combination of input variables. The mathematical equation is presented in the Equation (1):

$$y_i = \beta_0 + \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2} + \cdots + \beta_n \cdot x_{in} \tag{1}$$

where y_i is the dependent value, $\forall i \in [1, m]$; β_j is the regression coefficient of input variable, $j \in [1, n]$; x_{ij} is the i^{th} value of the input variable j. The estimation of the parameters β_i is based on the least-squares method, which totals the squares of the differences between the observed and the predicted values.

Normalization and ridge regression were undertaken, as the latter shrinks coordinates with respect to the orthonormal basis that is formed by the principal components. Table 3 shows adjusted R^2 values per model and the Nash-Sutcliffe Efficiency coefficient (NSE), demonstrating that solar radiation (SR) has relevance in ozone production for diurnal models under the hypothesis of linearity. The wind speed is a dominant parameter during the night in the dispersion and the distribution of ozone in all three areas. An asterisk inside parentheses accounts for the coefficient's relevance when hypotheses about their significance were contrasted at $\alpha = 0.95$.

The R^2 values in Mhamid are lower than those in the two other areas, thereby justifying the difference in ozone behavior at this station. This station receives more ozone from down-

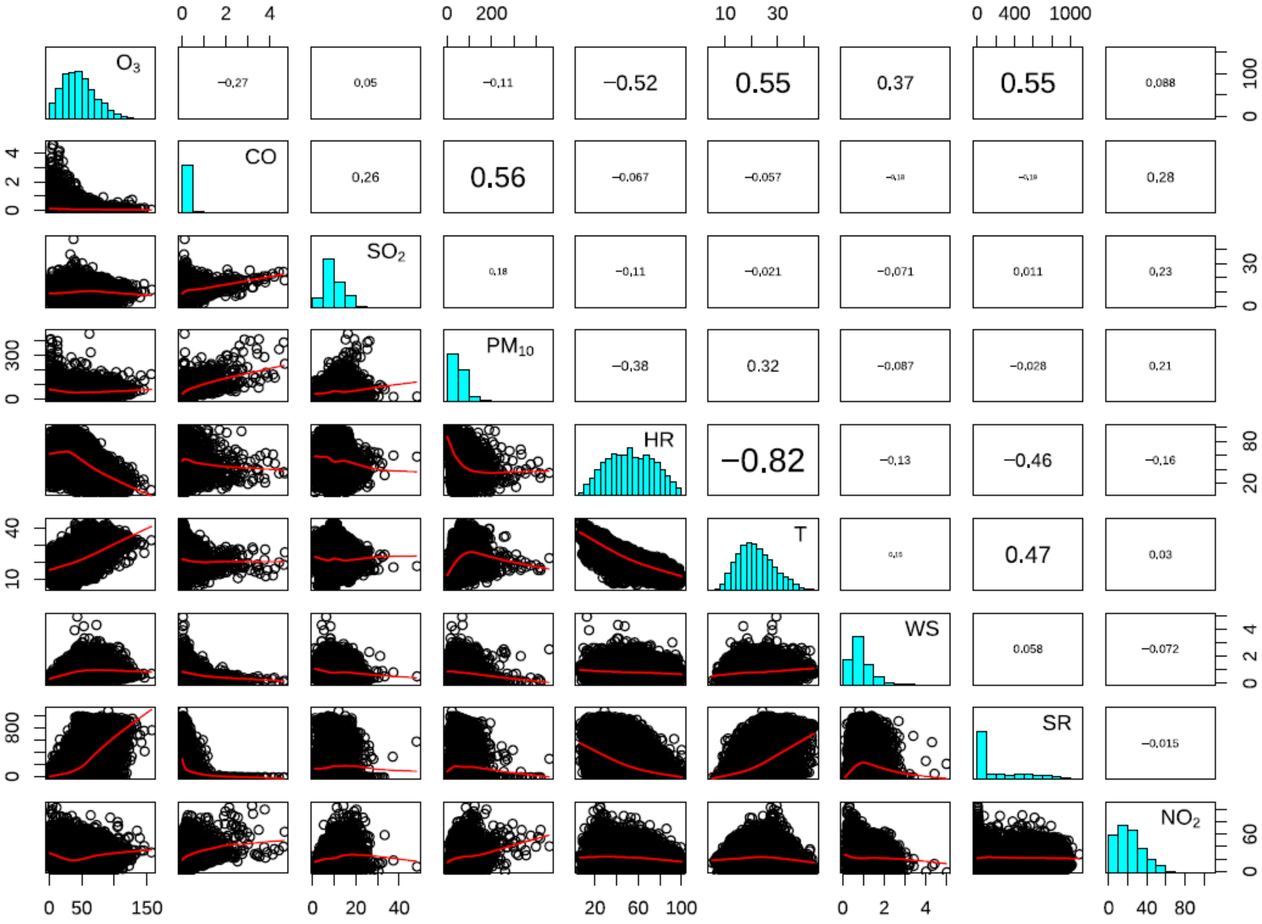

Figure 2. Linear correlation of variables for the Jef station.

town than by producing it, since it is the least affected by anthropogenic emissions.

Researchers wish to analyze different regressors (ANN, SVR, and RF) and their forecasting capability in order to reduce the risk of severe mistakes. The reference forecasting prediction of linear models will be considered the baseline for nonlinear techniques.

4.1.2.1. ANN Models

Artificial Neural Networks (ANN) are programming paradigms that seek to emulate the microstructure of the brain. Multilayer Perceptron (MLP) is an example of an ANN that can be seen as a function that transforms the input space into the output. It does this by processing every input signal by convenient weight into neurons that are located at the hidden layer (Gardner and Dorling, 1999; Dutot et al., 2007; Salazar-Ruiz et al., 2008).

The output O of a neural network (NN) can be defined as a function of input I and weight W in the form $O = \phi(I, W)$, in which ϕ represents the mapping function defined by the NN (ϕ: $R_m \rightarrow R$). The NN learning process consists of adjusting weights w_{ij} in order that a "good" mapping ϕ for the learning data is achieved, including local independence through bias component (B). Therefore, for learning data with an unknown mathematical relation, the NN provides a mapping. Based on an input I_i and its corresponding desired output O_i, the training process minimizes the energy function E, which is the mean quadratic error between the desired (d) and achieved outputs (o) for all elements in the training set (see Equation (2)):

$$E = \frac{1}{2} \sum\nolimits_{t \in Training\ Set} (d - o_t)^2 \tag{2}$$

An MLP trained with Back Propagation algorithm was used, varying the number of neurons in the hidden layer between 4 and 20. Also, different decay speeds and number of iterations and relative tolerances were used (see Figure 3).

The best architecture was a fully connected configuration, with linear output layer and logistic as activation functions and with the hidden neurons as indicated in Table 4.

One can learn from the architectures how behavior that differed from the others was selected in the case of Daoudiat station. At the latter station, the nighttime structure without temperature information reduces the incoming information through layers (6-5-1), instead of adding diversity by promoting larger number of neurons in the hidden layer than the input layers have.

Table 3. Linear Regression Modelling of Dependent Variable O_3 and the Resulting Regression Coefficients

	Jef	Jef	Jef	Daoudiat	Daoudiat	Daoudiat	Mhamid	Mhamid	Mhamid
	Total	Diurnal	Nocturnal	Total	Diurnal	Nocturnal	Total	Diurnal	Nocturnal
	$R^2 = 0.57$	$R^2 = 0.55$	$R^2 = 0.50$	$R^2 = 0.52$	$R^2 = 0.45$	$R^2 = 0.36$	$R^2 = 0.30$	$R^2 = 0.26$	$R^2 = 0.18$
	NSE = 0.58	NSE = 0.55	NSE = 0.50	NSE = 0.52	NSE = 0.45	NSE = 0.36	NSE = 0.30	NSE = 0.26	NSE = 0.18
Ind.	(*) 20.06	(*) 41.59	3.22	(*) 92.76	(*) 101.87	(*) 84.79	(*) 24.72	(*) 29.84	(*) 6.30
CO	(*) -5.08	(*) -12.46	-1.07	(*) -1.32	-0.86	(*) -1.58	(*) -32.20	(*) -44.03	(*) -25.93
NO_2	(*) 0.23	(*) 0.31	(*) 0.13	(*) -1.00	(*) -1.01	(*) -1.08	(*) 0.47	(*) 0.53	(*) 0.39
PM_{10}	(*) -0.16	(*) -0.15	(*) -0.14	(*) -0.05	-0.04	-0.03	0.002	0.006	-0.001
RH	(*) -0.20	(*) -0.43	0.001	(*) -0.40	(*) -0.45	(*) -0.36	(*) -0.18	(*) -0.28	(*) -0.15
WS	(*) 13.58	(*) 7.45	(*) 20.41	(*) 2.64	0.22	(*) 6.15	(*) 5.99	(*) 3.51	(*) 9.93
SR	(*) 0.03	(*) 0.02	-	(*) 0.02	(*) 0.02	-	(*) 0.03	(*) 0.03	-
T	(*) 1.06	(*) 0.81	(*) 1.04	-	-	-	(*) 0.39	(*) 0.51	0.03

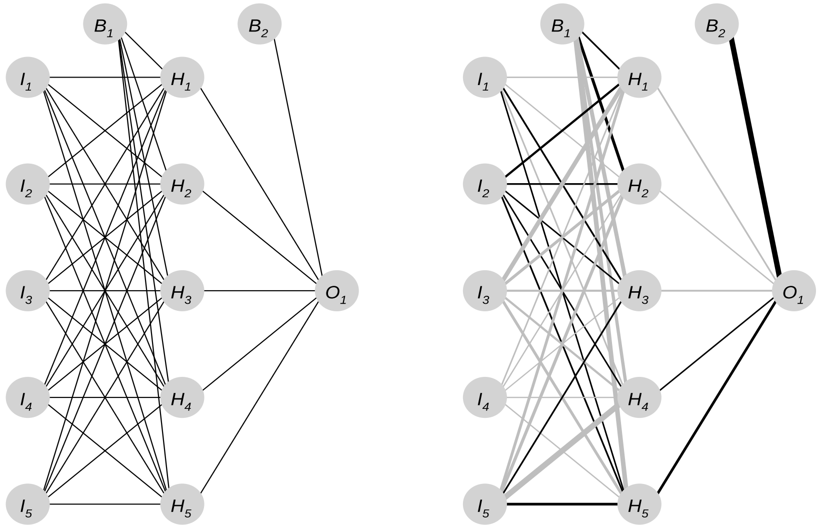

Figure 3. Proposed ANN topology for nightly prediction at the Daoudiat station. Left: Topology. Right: weighted connections relevance. *I*: Input neuron; *H*: Hidden neuron; *O*: Output neuron; *B*: Bias element.

After selecting the most suitable configuration from the tenfold cross-validation process, the model fitness from training data that was estimated as its adjusted R^2 coefficient can be seen in Table 5, where it outperforms the MLR by approximately 15%.

Table 4. Number of Hidden Neurons for the MLP Architecture (X)

Station	Daytime Dataset	Nighttime Dataset	Complete Dataset
Daoudiat (6-X-1)	13	5	4
Mhamid (7-X-1)	18	19	19
Jef (7-X-1)	18	19	13

Table 5. R^2 for the ANN / R^2 for the Linear Model

Station	Daytime Dataset	Nighttime Dataset	Complete Dataset
Daoudiat	0.48 / 0.45	0.38 / 0.35	0.53 / 0.52
Mhamid	0.46 / 0.26	0.49 / 0.18	0.53 / 0.30
Jef	0.601 / 0.55	0.58 / 0.50	0.67 / 0.58

4.1.2.2. SVR Models

The standard support vector regression (SVR) is a quadratic optimization where the goal is to find a function $f(x)$ that has no more than an ε deviation from the currently obtained targets for all of the training data, and also has the smallest possible slope (see Figure 4). It is helpful to introduce slack variables ξ_i, ξ^*_i to cope with the otherwise unfeasible constraints of the optimization problem (Cao et al., 2003). The constant $C > 0$ determines the trade-off between the flatness of f and the maximum acceptable amount of deviation that exceeds ε. Only the points that are outside the shaded region contribute to the cost insofar as the deviations are penalized in a linear fashion. To allow nonlinear SVR, we can preprocess the training patterns by a kernel Φ: $X \rightarrow F$ into some feature space F and then apply the standard SVR algorithm.

In SVR the objective function is to minimize a cost function called E, as presented in Equation (3):

$$E = \frac{1}{2}\left\|w\right\|^2 + C \cdot \sum_{i=1}^{n} \left|y_i - D(x_i, w)\right|_\varepsilon \tag{3}$$

where D is presented in Equation (4):

$$D(x,w) = sign(\sum\nolimits_{i=1}^{n} w\alpha_i \Phi^T(x_i)\Phi(x) + b) \qquad (4)$$

After selecting the most suitable configuration from the tenfold cross-validation process (see Table 6), the model fitness from training data, estimated as its adjusted R^2 coefficient and the NSE (presented in parentheses), can be seen in Table 7 where it outperforms the MLR by about 29% with less bias.

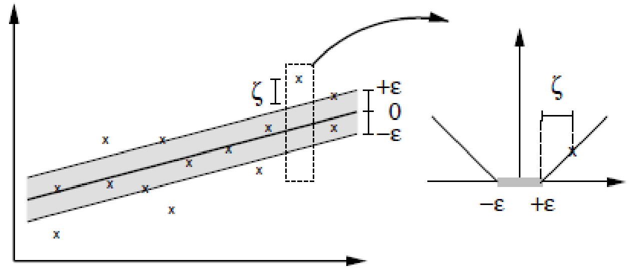

Figure 4. SVR soft margin concept.

Table 6. C, Gamma, Epsilon, and Number of Support Vectors Adopted for SVR Models

	Daytime Dataset	Nighttime Dataset	Complete Dataset
Daoudiat	8; 0.13; 0.1; 1205	2; 0.5; 0.1; 958	2; 0.5; 0.1; 2107
Mhamid	8; 0.5; 0.1; 3026	8; 0.5; 0.1; 3079	8; 0.5; 0.1; 5918
Jef	2; 0.5; 0.1; 2906	4; 0.5; 0.1; 2642	2; 0.5; 0.1; 5441

Table 7. R^2 for the SVR Model / Adjusted R^2 for the Linear Model

Station	Daytime Dataset	Nighttime Dataset	Complete Dataset
Daoudiat	0.58 / 0.45 (0.58)	0.58 / 0.35 (0.58)	0.69 / 0.52 (0.69)
Mhamid	0.72 / 0.26 (0.72)	0.64 / 0.18 (0.64)	0.69 / 0.30 (0.69)
Jef	0.85 / 0.55 (0.85)	0.76 / 0.50 (0.76)	0.83 / 0.58 (0.83)

4.1.2.3. Random Forest Models

Random Forest (RF) (Liaw and Wiener, 2002) is a predicttor ensemble with a set of decision trees that grow in randomly selected subspaces of data. They can be implemented quickly and easily. They produce highly accurate predictions and can handle a very large number of input variables without overfitting. In fact, RF is considered to be one of the most accurate general-purpose learning techniques that are available.

After selecting the most suitable configuration from the tenfold cross-validation process (see Table 8), the model fitness from the training data is estimated as its adjusted R^2 coefficient, which can be seen in Table 9 to outperform the MLR by about 38%.

It is also helpful to analyze the importance that these methods attribute to the different variables (see Table 10), as it appears that different nighttime and daytime mechanisms are at work. For daytime pollution, the major participants are T, NO_2, SR, and RH, as expected, because O_3 is a secondary pollutant. However, in the nighttime episodes, the relevance of WS becomes much greater where transport mechanisms used to be dominant.

Table 8. The Number of Variables per Tree and the Number of Trees in the Ensemble Adopted for the Random Forest Models

Station	Daytime Dataset	Nighttime Dataset	Complete Dataset
Daoudiat	5 ; 300	2 ; 500	3 ; 700
Mhamid	4 ; 500	4 ; 700	4 ; 300
Jef	4 ; 300	2 ; 700	3 ; 900

Table 9. R^2 for the Random Forest Model / R^2 for the Linear Model

Station	Daytime Dataset	Nighttime Dataset	Complete Dataset
Daoudiat	0.93 / 0.45 (0.93)	0.91 / 0.35 (0.91)	0.93 / 0.52 (0.91)
Mhamid	0.95 / 0.26 (0.95)	0.93 / 0.18 (0.93)	0.94 / 0.30 (0.94)
Jef	0.95 / 0.55 (0.95)	0.93 / 0.50 (0.93)	0.95 / 0.58 (0.95)

4.1.2.4. Other Techniques

In order to explore other nonlinear techniques, optimal selection of the model's parameters were carried out with the use of the caret meta-modeler package that is available within the R library for some other models. Those models were the followings. The relative performance of those models can be seen in Table 11, where the NSE is presented in parenthese.

- Stochastic Gradient Boosting (Gbm),
- Boosted Trees (Blackboost),
- Ridge Regression (Ridge),
- Project Pursuit Regression (Ppr),
- Multivariate Adaptive Regression Splines Models (Earth),
- Generalized Linear Models (Glm),
- Generalized Additive Models (Gam),
- Partial Least Squares with Kernel (Kernelpls) and
- Two ensembles - one defined by model stacking (ELM) and the other greedy approximated (Greedy) throughout all of the previously trained models.

Table 10. Relative Importance of Each Variable as Identified by the Random Forest Method. Importance Weighted Impurity Decreases Measured by the Gini Index

	Daoudiat Daytime Model	Daoudiat Nighttime Model	Daoudiat Complete Model	Mhamid Daytime Model	Mhamid Nighttime Model	Mhamid Complete Model	Jef Daytime Model	Jef Nighttime Model	Jef Complete Model
CO	9.23	12.00	8.94	7.45	13.31	8.43	4.90	10.99	6.00
NO_2	26.88	10.36	27.00	26.50	30.87	23.58	15.44	10.17	10.68
PM_{10}	9.57	10.79	7.79	5.84	18.23	7.52	7.72	13.78	8.43
RH	34.90	21.69	25.61	14.89	8.39	11.89	30.48	10.94	17.43
WS	5.80	45.15	6.61	5.01	12.75	6.90	4.66	34.12	12.56
SR	13.63	-	24.05	19.69	-	22.02	17.98	-	23.06
T	-	-	-	20.63	16.44	19.65	18.81	20.01	21.84

Table 11. R^2 of Additional Nonlinear Models throughout the Daoudiat Dataset

	Complete Dataset	Daytime Dataset	Nighttime Dataset
Gbm	0.78 (0.8)	0.79 (0.8)	0.71 (0.7)
Blackboost	0.78 (0.8)	0.78 (0.8)	0.66 (0.7)
Ridge	0.75 (0.8)	0.73 (0.7)	0.64 (0.6)
Ppr	0.78 (0.8)	0.77 (0.8)	0.68 (0.7)
Earth	0.78 (0.8)	0.78 (0.8)	0.64 (0.6)
Glm	0.72 (0.7)	0.73 (0.7)	0.64 (0.6)
Gam	0.76 (0.8)	0.75 (0.7)	0.65 (0.6)
Kernelpls	0.78 (0.7)	0.73 (0.7)	0.66 (0.7)
Greedy	0.82 (0.8)	0.82 (0.8)	0.72 (0.7)
ELM	0.82 (0.8)	0.82 (0.8)	0.72 (0.7)

4.1.3 Modelling Unbalanced Data

The analysis reveals clearly the improvement that is provided by combining different learners (ensemble modeling), even if they evolve into specialized learners. This arose previously in the context of RF models. This result also confirms what other studies (Mallet et al., 2009; Rahman et al., 2012; Silver et al., 2013; Debry and Mallet, 2014;) have found in different types of models.

Although the model learning processes look impressive, especially for the RF method (see Figure 5 for Daoudiat Station, where MLR, ANN, SVR, and RF model predictions are presented by rows; the first column reflects learning from the complete dataset, while the second column considers the nightly dataset only), these techniques are quite sensitive to unbalanced sets, because there is a slightly underestimated bias for higher pollution levels (Gong and Ordieres-Meré, 2016). To highlight this, we have plotted in Figure 5a dashed line with triangles embedded in it for the identity relationship and a dashed line with circles embedded in it for the linear regression of the predicted values. The closer together the two lines become, the greater the robustness against imbalanced datasets is. Again, the RF method outperforms all other methods. Further, it becomes clear how great the impact of lower data density is on imbalanced datasets, as nightly models do not perform as well as the diurnal ones or complete ones.

The biased learning due to an unbalanced dataset effect can be addressed by specific bootstrapping data preparation for such samples. Alternatively, we can accept it and simply change the model's prediction in a way that corrects any observed bias. The latter strategy was adopted in this work.

In 2009 the ozone hourly values regularly exceeded the threshold for several months, as presented in the supplementary material (see Figure S2). There were strong variations, reaching concentration levels of 100 to 150 μg/m^3, with some isolated events rising to 270 μg/m^3 in Mhamid. The hourly maximum reached 217 μg/m^3 in Daoudiat and 157 μg/m^3 in Jef, both of which are located downtown (see Figures S7 and S8 as well as Table S1 in the supplementary material). The same behavior was observed in 2010 as well.

In order to guarantee as much as possible the representativeness of the conclusions, datasets were prepared by randomly selecting 85% of the data to be used for training and the remaining 15% for final testing of the quality of the selected models.

Various types of models were presented in the previous section. The discussion of performance was based on cross-validation, including corrections to avoid a systematic bias. However, as discussed in Section 3.2, a small dataset was randomly selected and removed from the data that was used for cross-validation based learning processes. It is now possible to conduct an independent assessment of the model's quality with this new dataset.

4.2. Models Performance

To have confidence in the quality of models, a different experiment was conducted. To measure the forecasting capability of developed models, a subset of the existing Daoudiat station data was selected. This particular dataset had never been used previously, neither for training nor for validation. Therefore, the quality is not measured by the cross-validation technique, but by fresh data where developed models never saw before.

The RF models outperform all of the other types, even though SVR have only narrow differences (see Figure 6, where MLR, ANN, SVR, and RF models performance are presented by rows; the first column reflects validation from the complete dataset, while the second column considers the nightly dataset only). Consequently, the discussion will continue with reference to RF models, and the RMSE-measure error that

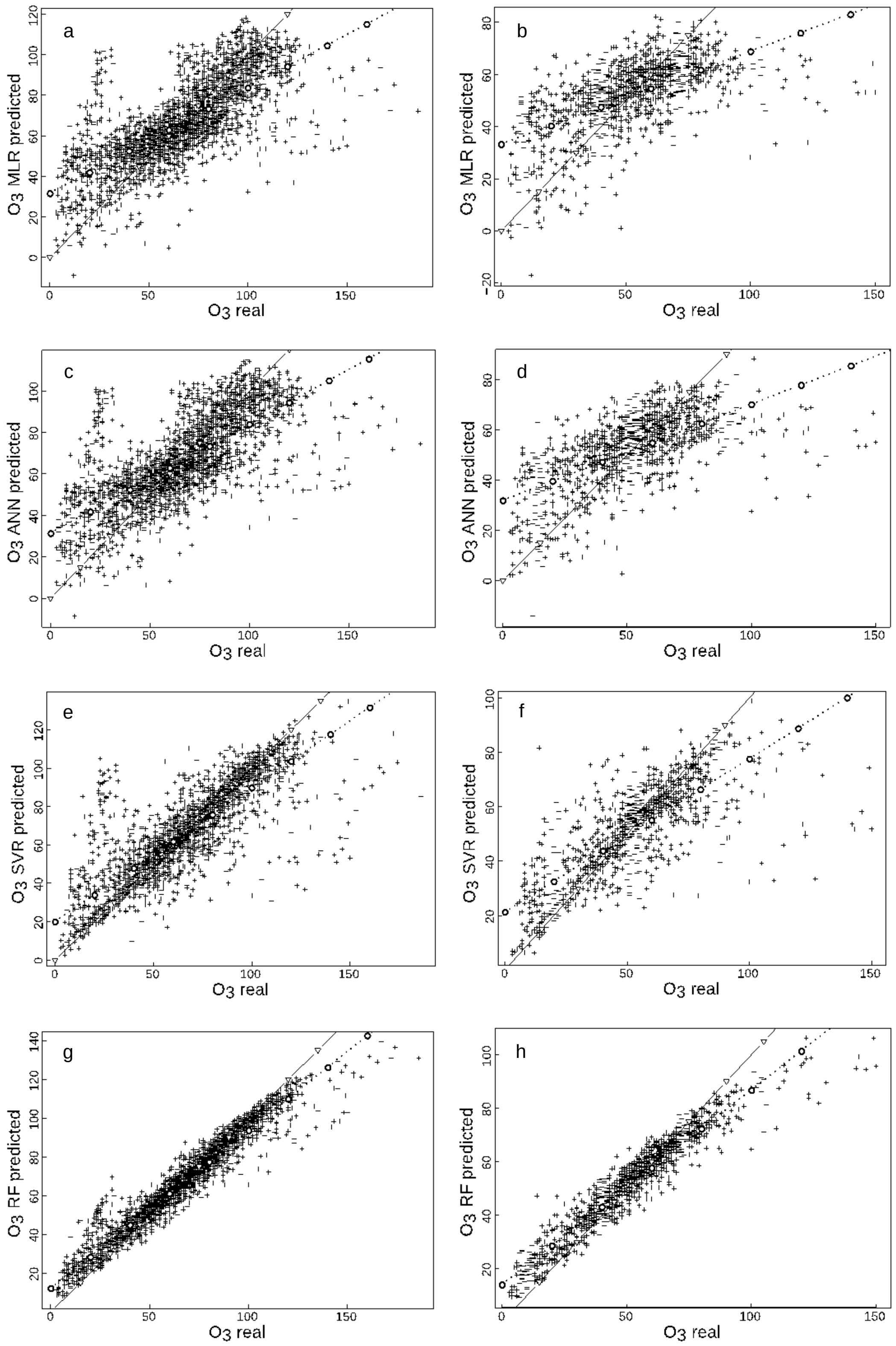

Figure 5. Bias of different models at Daoudiat Station. a - Linear regression model predictions for the whole dataset. b - Linear regression model predictions for the nighty dataset. c - ANN model predictions for the whole dataset. d - ANN model predictions for the nighty dataset. e - SVR model predictions for the whole dataset. f - SVR model predictions for the nighty dataset. g - RF model predictions for the whole dataset. h - RF model predictions for the nighty dataset.

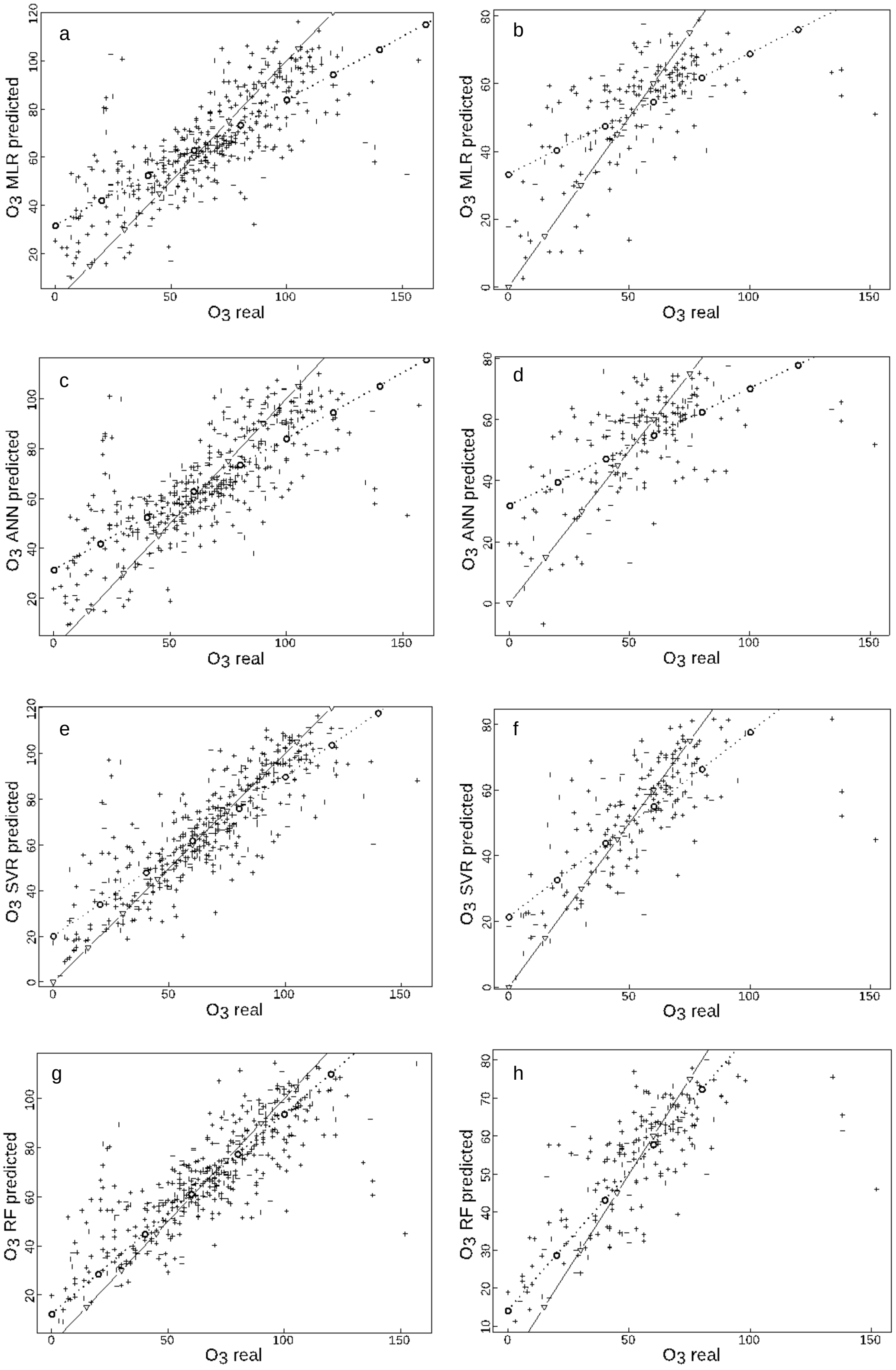

Figure 6. Performance of different models at Daoudiat Station. a - LR model performance for the whole dataset. b - LR model performance for the nighty dataset. c - ANN model performance for the whole dataset. d - ANN model performance for the nighty dataset. e - SVR model performance for the whole dataset. f - SVR model performance for the nighty dataset. g - RF model performance for the whole dataset. h - RF model performance for the nighty dataset.

was calculated at all stations for all of the datasets presented in Table 12, where the RRMSE values are presented in parentheses.

The ozone is a secondary pollutant that could reach a maximum due to a late reaction after sunrise (SR = 0) or lengthy transportation, depending on the wind speed. These factors can produce outliers (situations not well predicted by the specialized model as they are caused by a different physical mechanism, such as a local spill of VOC with fire, etc). They can be observed in Figure 6 where four outliers appear.

These values for RMSE are about half of those that were produced by the referenced linear regression model (33 μg/m^3 for Daoudiat, around 57 μg/m^3 for Mhamid and 24 μg/m^3 for Jef Stations). This evidence supports the contribution of non-linear modeling techniques in pollution forecasting, however, it also raises a clear signal that training models with cross-validation tend to be over-estimated by about 20% of the true RMSE for new and fresh data.

Table 12. RMSE in μg/m^3 for New Datasets when Random Forest Models are Used in Forecasting and are Bias-Corrected

Station	Daytime Dataset	Nighttime Dataset	Complete Dataset
Daoudiat	17.37 (0.22)	17.26 (0.33)	17.49 (0.26)
Mhamid	20.46 (0.39)	15.68 (0.41)	18.15 (0.40)
Jef	12.55 (0.22)	10.80 (0.30)	11.81 (0.26)

4.3. Models Sensitivity to the Ozone Production Mechanism

The authors were interested in determining how specific the models become when they were used at the same station, but to predict different datasets so several experiments were conducted when test datasets were being considered. They involved:

- predicting only daytime data with the model trained with the complete dataset;
- predicting only nighttime data with the model trained with the complete dataset;
- predicting daytime data with the model trained with the nighttime dataset; and
- predicting nighttime data with the model trained with the daytime dataset.

Table 13 (where the RRMSE values are presented in parentheses) summarizes the result. It is clear that non-linear models learned daytime and nighttime behaviors together, according to the RMSE values for cases (a) and (b), which are lower and similar for each station, and RRMSE which are smaller as well. Even when specific models outperform the predictions, the advantage is always moderated. However, it is also evident that mechanisms of ozone prediction differ for the daytime and nighttime datasets and when another model predicts them, the predictions become dramatically degraded, according to the values for cases (c) and (d), which are larger than the models trained with the whole dataset. The same happens with the RRMSE.

Table 13. RMSE in μg/m^3 for New Datasets when Random Forest Models are Used in Forecasting and are Bias-Corrected

Station	Case (a)	Case (b)	Case (c)	Case (d)
Daoudiat	17.27 (0.25)	17.27 (0.25)	20.18 (0.39)	28.54 (0.37)
Mhamid	19.94 (0.45)	15.53 (0.35)	19.78 (0.56)	30.15 (0.57)
Jef	12.58 (0.27)	10.81 (0.23)	14.43 (0.40)	24.18 (0.43)

4.4. Forecasting Time

The potential use of such models as a tool for policy decision makers to enforce regulatory measures requires the ability to forecast the ozone level several hours in advance. Within this time window, it will be possible to generate a proper advertisement and to fully communicate the actions being taken. The station selected was Mhamid and the family of models selected was RF, in accordance with the knowledge that was acquired in previous sections. Training sets were prepared by including the hour and the hourly ozone levels 8, 16, 24, 48 or 72 hours in advance. As usual, randomly-selected training and testing subsets were generated to analyze the performance. Figure 7 shows the capabilities of forecasting 72 hours in advance by models trained with the ensemble Random Forest technique are remarkable (left). They evolve logically as the standard deviation (SD) and inter-quartile ranges grow along with the forecast period (right).

The analysis of the importance of independent variables for the random forest model shows that the current ozone value is the most relevant factor. The second variable, with half of the relevance, is the current NO_2 concentration, followed by the wind speed, the hour of the day, the temperature, and the solar radiation. Other factors, such as relative humidity, are still relevant, although less so. Therefore, the evolution of ozone in this area is mainly affected by human behavior, with meteorology-cal factors having a less relevant, role. This means that pollution levels will have a strong relationship to anthropogenic combustion in car engines and related factors, thereby increasing the inhabitants and the transportation congestion that will negatively affect the ozone pattern. In the opposite way, policies that encourage increment renewal of cars, which will contribute to increasing their efficiency, will have a positive effect which can be quantified in limiting NO_2 levels. Because of the economic crisis currently affecting European citizens, the growth of tourism activities has been limited. This has contributed the most to moderating the pattern.

4.5. Model Generalization Capability

The last relevant experiment for this research concerns the specificity of models by station. This matter is relevant as several authors have become interested in spatial pollution interpolation (Blanchard et al., 2014). Also, it could be relevant for alternative estimation by classical regional models and for exposure estimation (Fraser et al., 2013).

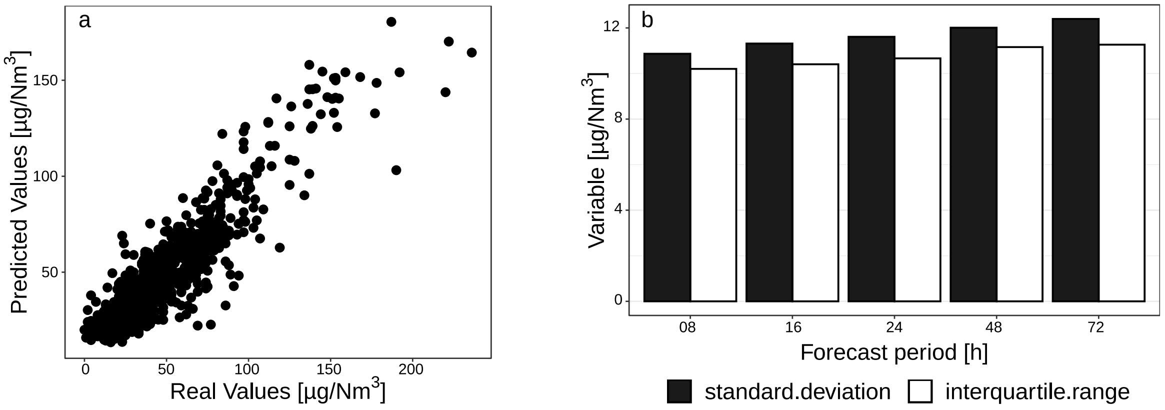

Figure 7. Performance of models at Mhamid station with different forecasting times.

Table 14. Error and Correlation of Data Predicted by the Models in Columns and Real Values of the Complete Dataset of Stations Defined by Row

Station	RMSE Daoudiat Model	RMSE Mhamid Model	RMSE Jef Model	Correlation Daoudiat Model	Correlation Mhamid Model	Correlation Jef Model
Mhamid	29.75	17.94	23.53	0.226	0.596	0.311
Jef	23.67	20.58	11.82	0.417	0.453	0.777

The experiment analyzes the forecasting of ozone at each station by using models from the other stations, whenever possible, as one station missed one of the relevant meteorological variables (T). In any case, it is still possible to evaluate the effect of substitution of specific models.

Table 14 presents the results of the experiment showing that replacement is not a good solution because performance losses range from 32% (17.94 of RMSE at Mhamid station against 23.53 by using the Jef model) to 99% (11.82 for RMSE at Jef station versus 23.67 by using Daoudiat model). This is relevant when inference is used to predict pollution in places where no information is available. However, it will not be possible to accept single behavior of the phenomena, which must be identified first in order to suggest an appropriate way to combine behaviors in the final forecasting model.

4. Conclusions

The main contributions of this paper can be understood from both a technical and an application point of view. From the technical point of view, one of the contributions of this paper is to show that nonlinear models for predicting hourly ozone levels outperform linear models by more than 50%, even though the latter makes it possible to understand the relevance of dependent variables. The study also shows that ensemble models have a greater forecasting accuracy than individual ones. To this end, it is worth remembering that RF is nothing but a class of ensemble technology. Indeed, it shows that boosted models outperform individual ones and that they can be compared to the ensemble-based ones, at least in the case of a limited number of variables. The learning capability of non-linear models based on the hidden structure of data was also emphasized, as switching models between daytime and nighttime datasets severely downgraded their performance. A large amount of data enables the construction of robust models, which makes it possible to extend the forecast time-window beyond the commonly used eight hours.

From the application point of view, it is relevant to highlight how such synthetic models can accurately forecast ozone pollution levels even when there is little information, not including details of wind direction, rainfall values, or traffic intensities. Therefore, the study contributes to policy makers' publicly declared interest in pollution monitoring by introducing characteristics of the area and pollution levels and also by developing specific and local models for short-term forecasting of air quality in an area of the world where there has been no such forecast. These instruments make it possible to warn the population of the days and periods that may pose a hazard and to contribute to the establishment of public policy for strategic decision-making procedures. The interest in Ozone emissions is motivated by two reasons: The first being that Morocco is an agricultural country and ozone at a daily concentration of 65 µg/m^3 has the effect of destroying the respiratory cells of plants and cereals. The second reason is that tourism is a significant source of economical revenue, therefore enhancing air quality is essential to maintaining and attracting tourists to Morocco. Indeed, by having better understanding of the local mechanisms responsible for the ozone pollution, different strategies can be assessed against what it could happen in the previous situation. Therefore, the 'what-if' scenario becomes easier to quantify, and it starts to be relevant as local authorities and elected officials are increasingly sensitive to the environment and ready to invest in the quality of life of citizens in a more integrated approach to spatial planning. In this way, alternatives such as diversifying infrastructure through one-way roads, multi-level roads, bypass routes and smart traffic lights that provide traffic flow with alternatives can be evaluated. Indeed, vigorously enforcing annual technical vehicles inspection can be measured in terms of real improvement by having a simulator of the previous context for a given meteorological context.

The ozone concentrations at the three stations differ in their specific nature, reflecting the economic characteristics of Marrakesh. That means that models are able to cope with physical behavior. Therefore, it was possible to identify the mechanisms that are responsible for the ozone levels that were found. Similarly, the separation of the ozone concentration data into two categories (daytime and nighttime), has underscored the importance of solar radiation in daytime ozone production, the importance of wind speed on the measured values of ozone concentration at night, and the influence of primary pollutants and relative humidity globally. To assess this approach, a specific study was carried out with the goal of extending the coverage of the models to other ranges where such exchanges increased the expected error levels.

Additionally, as far as Morocco's air quality monitoring network is constituted of 29 fixed stations and 3 mobile stations across 15 towns, the used methodology can be easily expanded to other cities in the country, as well as to different locations in the world.

Finally, this paper has shown that model interchange between geographical locations does not perform well without additional considerations as there are various mechanisms at work. An interesting field of research will be to develop forecasting ozone models that are based on local, short-term, data-oriented prediction models in a way that considers and employs different mechanisms.

Acknowledgments. The authors wish to thank the Centro de Supercomputación y Visualización de Madrid (CeSVIMa), and the Institut de Recherche pour le Développement for providing the solar radiation data. This work was completed within a framework of collaboration between the UCA and the DMN.

References

Ambroise, C. and Grandvalet, Y. (2001). Prediction of ozone peaks by mixture models, *Ecol. Model.*, 145(2-3), 275-289. https://doi.org/10.1016/S0304-3800(01)00399-4

Avnery, S., Mauzerall, D.L., Liu, J., and Horowitz, L.W. (2011). Global crop yield reductions due to surface ozone exposure: 1. Year 2000 crop production losses and economic damage, *Atmos. Environ.*, 45(13), 2284. https://doi.org/10.1016/j.atmosenv.2010.11.045

Blanchard, C.L., Tanenbaum, S., and Hidy, G.M. (2014). Spatial and temporal variability of air pollution in Birmingham, Alabama, *Atmos. Environ.*, 89, 382-391. https://doi.org/10.1016/j.atmosenv.2014.01.006

Cao, L.J., Chua, K.S., Chong, W.K., Lee, H.P., and Gu, Q.M. (2003). A comparison of PCA, KPCA and ICA for dimensionality reduction in support vector machine, *Neurocomputing*, 55(1-2), 321. https://doi.org/10.1016/S0925-2312(03)00433-8

Chan, L., Chan, C., and Qin, Y. (1998). Surface ozone pattern in Hong Kong, *J. Appl. Meteorol.*, 37(10), 1153-1165. https://doi.org/10.1175/1520-0450(1998)037<1153:SOPIHK>2.0.CO;2

Debry, E. and Mallet, V. (2014). Ensemble forecasting with machine learning algorithms for ozone, nitrogen dioxide and PM10 on the Prev'Air platform, *Atmos. Environ.*, 91, 71-84. https://doi.org/10.1016/j.atmosenv.2014.03.049

Derwent, R.G., Jenkin, M.E., Saunders, S.M., and Pilling, M.J. (1998). Photochemical ozone creation potentials for organic compounds in northwest Europe calculated with a master chemical mechanism, *Atmos. Environ.*, 32(14), 2429-2441. https://doi.org/10.1016/S1352-2310(98)00053-3

Duan, J., Tan, J., Yang, L., Wu, S., and Hao, J. (2008). Concentration, sources and ozone formation potential of volatile organic compounds (VOCs) during ozone episode in Beijing, *Atmos. Res.*, 88(1), 25-35. https://doi.org/10.1016/j.atmosres.2007.09.004

Dutot, A., 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. Model. Software*, 22(9), 1261-1269. https://doi.org/10.1016/j.envsoft.2006.08.002

Fraser, S., Marceau, D., De Visscher, A., and Roth, S. (2013). Estimating exposure by loose-coupling an air dispersion model and a geospatial information system, *J. Environ. Inf.*, 21(2), 84-92. https://doi.org/10.3808/jei.201300235

Freeman, B.S., Taylor, G., Gharabaghi, B., and Thé, J. (2018). Forecasting air quality time series using deep learning. *J. Air Waste Manage. Assoc.*, https://doi.org/10.1080/10962247.2018.1459956

Gao, H.O. (2007). Day of week effects on diurnal ozone/NOx cycles and transportation emissions in Southern California, *Transport. Res. D: Transp. Environ.*, 12(4), 292-305. https://doi.org/10.1016/j.trd.2007.03.004

Gao, M., Yin, L., Ning, J., (2018). Artificial neural network model for ozone concentration estimation and Monte Carlo analysis. *Atmos. Environ*., 184, 129-139. https://doi.org/10.1016/j.atmosenv.2018.03.027

Gardner, M. W., and Dorling, S. R. (1999). Neural network modelling and prediction of hourly NOx and NO2 concentrations in urban air in London, *Atmos. Environ.*, 33(5), 709-719. https://doi.org/10.1016/S1352-2310(98)00230-1

Gong, B. and Ordieres-Meré, J. (2016). Prediction of daily maximum ozone threshold exceedances by preprocessing and ensemble artificial intelligence techniques: Case study of Hong Kong, *Environ. Model. Software*, 84, 290-303. https://doi.org/10.1016/j.envsoft.2016.06.020

Gong, B. and Ordieres-Meré, J. (2017). Reconfiguring existing pollutant monitoring stations by increasing the value of the gathered information. *Environ. Model. Software*, 96, 106-122. https://doi.org/10.1016/j.envsoft.2017.06.034

He, H. and Lu, W. (2012). Decomposition of pollution contributors to urban ozone levels concerning regional and local scales, *Build. Environ.*, 49(0), 97-103. https://doi.org/10.1016/j.buildenv.2011.09.019

Inchaouh, M., Tahiri, M., Bouchra, E.J., and Abboubi, R. (2017). State of ambient air quality in Marrakech city (Morroco) over the period 2009-2012, *Int. J. GEOMATE*, 12(29), 99-106. https://doi.org/10.21660/2017.29.1254

Jenkin, M.E. and Clemitshaw, K.C. (2000). Ozone and other secondary photochemical pollutants: chemical processes governing their formation in the planetary boundary layer, *Atmos. Environ.*, 34(16), 2499-2527. https://doi.org/10.1016/S1352-2310(99)00478-1

Khatibi, R., Naghipour, L., Ghorbani, M.A., Smith, M.S., Karimi, V., Farhoudi, R., Delafrouz, H., and Arvanaghi, H. (2013). Developing a predictive tropospheric ozone model for Tabriz, *Atmos. Environ.*, 68, 286-294. https://doi.org/10.1016/j.atmosenv.2012.11.020

Kingdom of Morocco (2010). *Protection of air against Pollution*. January 21st, Bulletin Officiel du Maroc, 5806.

Kommalapati, R., Liang, Z., and Huque, Z. (2016). Photochemical model simulations of air quality for Houston-Galveston-Brazoria area and analysis of ozone-NOx-hydrocarbon sensitivity. *Int. J. Environ. Sci. Technol.*, 13(1), 209-220. https://doi.org/10.1007/s13762-015-0862-6

Kovač-Andrić, E., Brana, J., and Gvozdić, V. (2009). Impact of meteorological factors on ozone concentrations modelled by time series analysis and multivariate statistical methods, *Ecol. Inf.*, 4(2), 117-122. https://doi.org/10.1016/j.ecoinf.2009.01.002

Kuhn, M. and Contributions from Wing, Jed and the R Core Team (2014). *caret: Classification and Regression Training*.

Lee, S., Kim, Y., Kim, H., and Lee, H. (2007). Influence of dense surface meteorological data assimilation on the prediction accuracy of ozone pollution in the southeastern coastal area of the Korean Peninsula, *Atmos. Environ.*, 41(21), 4451. https://doi.org/10.1016/j.atmosenv.2007.01.050

Lei, H., Wuebbles, D.J., and Liang, X. (2012). Projected risk of high ozone episodes in 2050, *Atmos. Environ.*, 59(0), 567-577. https://doi.org/10.1016/j.atmosenv.2012.05.051

Liaw, A. and Wiener, M. (2002). Classification and Regression by Random Forest, *R news*, 2(3), 18-22.

Limas, M.C., Mere, J.B.O., Marcos, A.G., Ascacibar, Francisco Javier Martinez de Pison, Espinoza, A.V.P., Elias, F.A., and Ramos, J.M.P. (2014). *AMORE: A MORE Flexible Neural Network Package*, 2015.

Lin, M., Fiore, A.M., Horowitz, L., Cooper, O.R., Naik, V., Holloway, J., Johnson, B.J., Middlebrook, A.M., Oltmans, S.J., and Pollack, I.B. (2012). Transport of Asian ozone pollution into surface air over the western United States in spring, *J. Geophys. Res. Atmos*, 117(D21). https://doi.org/10.1029/2011JD016961

Mallet, V., Stoltz, G., and Mauricette, B. (2009). Ozone ensemble forecast with machine learning algorithms, *J. Geophys. Res. Atmos.*, 114(D5), 1984--2012. https://doi.org/10.1029/2008JD009978

Matus, K., Nam, K., Selin, N.E., Lamsal, L.N., Reilly, J.M., and Paltsev, S. (2012). Health damages from air pollution in China, *Global Environ. Change*, 22(1), 55-66. https://doi.org/10.1016/j.gloenvcha.2011.08.006

Matus, K., Yang, T., Paltsev, S., Reilly, J., and Nam, K. (2008). Toward integrated assessment of environmental change: air pollution health effects in the USA, *Clim. Change*, 88(1), 59-92. https://doi.org/10.1007/s10584-006-9185-4

Myhre, G., Shindell, D., Bréon, F., Pongratz, J. (2013). Anthropogenic and natural radiative forcing, *Clim. Change*, 658-740.

Ouarzazi, J., Terhzaz, M., Abdellaoui, A., Bouhafid, A., Nollet, V., and Dechaux, J. (2003). A descriptive study of atmospheric pollutants measurement in the Marrakech conurbation, *Pollut. Atmos.*, (177), 137-151.

Rahman, S.M., Khondaker, A.N., and Abdel-Aal, R. (2012). Self organizing ozone model for Empty Quarter of Saudi Arabia: Group method data handling based modeling approach, *Atmos. Environ.*, 59(0), 398. https://doi.org/10.1016/j.atmosenv.2012.05.008

Riga, M., Stocker, M., Rönkkö, M., Karatzas, K., and Kolehmainen, M. (2015). Atmospheric environment and quality of life information extraction from twitter with the use of self-organizing maps, *J. Environ. Inf.*, 26(1), 27-40. https://doi.org/10.3808/jei.201500311

Salazar-Ruiz, E., Ordieres, J.B., Vergara, E.P., and Capuz-Rizo, S.F. (2008). Development and comparative analysis of tropospheric ozone prediction models using linear and artificial intelligence-based models in Mexicali, Baja California (Mexico) and Calexico, California (US), *Environ. Model. Software*, 23(8), 1056-1069. https://doi.org/10.1016/j.envsoft.2007.11.009

Seinfeld, J.H. (1989). Urban air pollution: State of the science, *Science*, 243(4892), 745. https://doi.org/10.1126/science.243.4892.745

Silver, J. D., Ketzel, M., and Brandt, J. (2013). Dynamic parameter estimation for a street canyon air quality model, *Environ. Model. Software*, 47(0), 235. https://doi.org/10.1016/j.envsoft.2013.05.012

Team, R.C. (2014). *R: A Language and Environment for Statistical Computing*.

Zhang, W., Wang, J., Liu, X., and Wang, J. (2013), Prediction of ozone concentration in semi-arid areas of China using a novel hybrid model, *J. Environ. Inf.*, 22(1), 68-77. https://doi.org/10.3808/jei.201300246