Daily pollution forecast using optimal meteorological data at synoptic and local scales

Ana Russo\textsuperscript{a}, Pedro G. Lind\textsuperscript{c,d}, Frank Raischel\textsuperscript{a,b}, Ricardo Trigo\textsuperscript{a}, Manuel Mendes\textsuperscript{e}

\textsuperscript{a}Center for Geophysics, IDL, University of Lisbon 1749-016 Lisbon, Portugal
\textsuperscript{b}Center for Theoretical and Computational Physics, University of Lisbon, Av. Prof. Gama Pinto 2, 1649-003 Lisbon, Portugal
\textsuperscript{c}TWIST - Turbulence, Wind energy and Stochastics, Institute of Physics, Carl-von-Ossietzky University of Oldenburg, DE-26111 Oldenburg, Germany
\textsuperscript{d}ForWind - Center for Wind Energy Research, Institute of Physics, Carl-von-Ossietzky University of Oldenburg, DE-26111 Oldenburg, Germany
\textsuperscript{e}Instituto de Meteorologia, Rua C-Aeroporto, 1749-077 Lisbon, Portugal

Abstract

We present a simple framework to easily pre-select the most essential data for accurately forecasting the concentration of the pollutant PM$_{10}$, based on pollutants observations for the years 2002 until 2006 in the metropolitan region of Lisbon, Portugal. Starting from a broad panoply of different data sets collected at several meteorological stations, we apply a forward stepwise regression procedure that enables us not only to identify the most important variables for forecasting the pollutant but also to rank them in order of importance. We argue the importance of this variable ranking, showing that the ranking is very sensitive to the urban spot where measurements are taken. Having this pre-selection, we then present the potential of linear and non-linear neural network models when applied to the concentration of pollutant PM$_{10}$. Similarly to previous studies for other pollutants, our validation results show that non-linear models in average perform as well or worse as linear models for PM$_{10}$. Finally, we also address the influence of Circulation Weather Types, characterizing synoptic scale circulation patterns and the concentration of pollutants.

1. Introduction

Air pollution is a global threat to public health and to the environment, although its effects are generally strongest in urban areas \cite{1,2,3,4}. Urban air pollution is a complex mixture of toxic components, which may induce acute and chronic responses from sensitive groups, such as children and people with previous heart and respiratory insufficiencies \cite{1,5,6}. Therefore, forecasting the temporal evolution of air pollution concentrations in urban locations emerges as a priority for guaranteeing life quality in urban areas \cite{1,3,4}. Modelling air pollution allows to describe the causal relationship between emissions, meteorology, atmospheric concentrations, deposition, and other factors, including the determination of the effectiveness of remediation strategies, and the simulation of future scenarios. Different types of approaches have been applied to characterize and forecast the dispersion of air pollutants, from the most simple approaches, such as box models \cite{7}, Gaussian plume models \cite{8}, persistence and regression models \cite{9}, to the most complex model systems, namely UAM-Urban Airshed Model \cite{10}, ROM-Regional Oxidant Model \cite{11}, CHIMERE \cite{12}, CMAQ-Community Multi-scale Air Quality Model \cite{13,14,15}.

Simpler models are used often as they can provide a fast overview. However they rely on significant simplifying assumptions and usually do not describe the complex processes and interactions that control the transport and chemical behavior of pollutants in the atmosphere \cite{13}.

For detailed characterization of atmospheric pollution more sophisticated models are needed, such as dispersion models which are driven by the objective quantification of chemical reactions and the physical transport of pollutants. In the last decades, significant progress has been made in air-quality dispersion models \cite{15}. However, being highly non-linear, they require large amounts of accurate input data and are computationally expensive \cite{16}.

Statistical models, such as Artificial Neural Networks (NN), can constitute a promising alternative to deterministic models \cite{17,18}, namely in what concerns air pollution problems \cite{16,19,20,22,23,24}. These models are usually regarded as a good compromise between simplicity and effectiveness, being capable of modeling the effect of non-linearities and fluctuations.

Although NN models may involve greater uncertainty than more complex models, the input data requirements are less strict. Several NN models were already tested, mostly for forecasting hourly averages \cite{1,25,26} or daily maxima \cite{27} of air pollutants. Some authors compared the potential of different approaches when applied to different pollutants and prediction time lags \cite{17,20,22,26}. Other authors have proven better forecasting results of NN over multiple linear regression (MLR) \cite{25,26,28}. More recently, some of us \cite{29} showed that, combining NN models and stochastic data analysis, allows to diminish the requirement of large training data sets often appearing when constructing a NN model.

Despite these improvements forecasting NN models still
present some caveats that need to be properly addressed. First, the time-lag in which air pollution prediction is performed should be as large as possible for enabling effectiveness of alert procedures in urban centers. Although hourly NN modeling has been frequently and successfully applied in air pollution studies, modeling daily concentration is more adequate to enable useful information to citizens [22].

Second, the construction of the best NN structure and the choice of input parameters constitutes another challenge for modelers [22]. Theoretically, any set of input data can be fed into any NN architecture for training and evaluation. However, the number of possible predictors and the number of ways they can be presented is too diverse to test all possible combinations. Here, we decided to use two of the most common architectures used in air quality modeling, a linear and non-linear NN. The linear NN model is based on a simple one-layer structure which produces the same results as a linear regression model [30] and the non-linear NN models are based on a feed-forward configuration of the multilayer perceptron that has been used by several authors [22, 23, 24]. Regarding the choice of the best input parameters, we argue that besides the intrinsic parameters describing air pollution, the interaction between pollutants and weather patterns should have the potential to significantly improve air quality forecasts.

Third, while several studies have been published establishing links between synoptic scale circulation patterns, usually named Circulation Weather Types (CWT) and air pollution [31, 33, 34, 35], the majority of the research focused on individual meteorological variables and non-automated procedures of variables selection. Weather is one of the factors that conditions air quality [31, 32], constraining the atmospheric processes that are associated to the occurrence of pollution episodes, namely, the processes of dilution, transformation, transport and removal of pollutants [32]. The relative importance of weather and climate for predicting the state of air quality has been investigated extensively over the last few decades [31, 33, 34, 35, 36, 37].

Those studies revealed that certain weather parameters are relevant to model air pollutant concentrations, particularly, the temperature, wind speed and direction, relative humidity, cloud cover, dew point temperature, sea level pressure, precipitation and mixing layer height [22, 33]. However, the majority of the research focused on individual meteorological variables and non-automated procedures of variables selection. A specific application to pollution in the Iberian Peninsula was developed by Saavedra et al. (2012), who presented a very detailed description of the relationship between synoptic pressure patterns and high-ozone incidents in northwest Galicia. Nevertheless, there are to the best of our knowledge no studies in the literature focusing on the application over the Iberian Peninsula of objective automatic classification procedures of CWT as a predictor for air quality modelling. Bringing the insight of such previous studies and considering CWTs as possible input parameters for NN models could provide better forecasts, particularly for large time-lags. The proposed approach should also allow to ascertain how strongly do meteorological variables and CWT influence the concentration of pollutants.

In this paper, we address all these three issues, aiming on daily forecast, introducing a simple framework for automatically rank the set of variables used as input variables for training the NN model and, in particular, addressing the influence of CWT in air pollution evolution. More exactly, we study relations between weather and air pollution through a circulation-to-environment approach based on the analysis of the existence of links between meteorological parameters and daily air quality measurements. This study focuses on the development of air quality models within the greater urban area of Lisbon, Portugal, based on an approach that is able to capture the temporal evolution of air quality and to produce corresponding forecasts. We choose to address one single pollutant belonging to the group of particulate matter (PM), namely PM\textsubscript{10}. Atmospheric PM comprehends in general small particles of solid or liquid matter in suspension or associated with atmospheric gases. Usually this mixture of air and PM is called atmospheric aerosol. Typically, these cover a wide range of sizes and chemical characteristics, including PAHs, acid aerosols and diesel particulates [39]. PM\textsubscript{10} include respirable particulate matter sized 10 \( \mu \text{g} \) or less, which poses a major health risk [39]. Although pollutants’ emissions have decreased over the last two decades, this did not lead to a corresponding reduction of concentrations of PM\textsubscript{10} throughout Europe [2], which influenced the choice of this pollutant as target of this study. Additionally, evidence has accumulated during the last years that there is a direct association between daily variations in the concentrations of airborne particles and a range of health indicators, which include daily deaths, admissions to hospital for the treatment of both respiratory and cardiovascular diseases and symptoms amongst patients suffering from asthma [5, 6, 39]. For the particular case of PM\textsubscript{10} since 2005, European Union imposes a limiting value of 50 \( \mu \text{g} \) per cubic meter of air [40], with a number of exceeding values not more than 35 per year. Lisbon is located closely to the Atlantic ocean where most of the moisture affecting western Iberia arrives [41], particularly in winter months [42]. Despite this impact of the ocean that mitigates the effects of aerosols and pollution, Lisbon has been affected by several high pollution episodes in the last two decades, exceeding repeatedly the legal limits imposed for PM\textsubscript{10} [43]. Therefore, and facing such restrictive rules and the exceeding events that occurred on the last years, a good PM\textsubscript{10} prediction procedure within a sufficiently large time-lag to prevent the occurrence of exceeding concentrations is needed.

To perform a systematic study, we apply both linear and non-linear NN models to predict PM\textsubscript{10} daily average concentrations based on air pollution and weather historical information.

Comrie [44] and Cobourn et al. [18] have performed comparison studies between NN and regression models to forecast ozone concentrations, both showing that NN outcomes are only equal or slightly better than regression models for ozone prediction. In contrast, Gardner and Dorling (2000) [21] showed that there is a significant increase in performance when using non-linear models, although regression models allow to easily understand and interpret results in terms of the physical mechanisms between meteorological and air quality variables.

Three important components will be incorporated:
1. An important aspect addressed by us is periodicity. Since the factors mainly contributing to air pollution concentration are connected with source activity and periodic variations in nature, it is normal to expect periodic components in air quality time series [1]. Hence, following a similar approach to the study presented by Kolehmainen et al. (2001) [1], two modeling approaches are possible. One is to model the original and complete signal. Another is to model the residual component after the removal of a periodic component from the complete signal. Here, we address the relative importance of the weekly periodic component. The weekly component is mainly affected by traffic and weekly business and industrial fluctuations. The forecasting capabilities of the different approaches are compared.

2. Furthermore, we also focus on the advantages of the application of an automated procedure for the selection of variables [45], recently used by us [29][37]: the use of an automated procedure prior to NN modeling allows for substantial reduction in the number of input variables, which enables also to improve the quality and robustness of pollutant concentration forecasts. These are crucial properties when linking the forecast to alert systems.

3. Finally, we use data from several monitoring stations in Lisbon. Therefore, our predictions will allow us to define air pollution episode alerts with spatial variability, instead of a unique value representing the entire region of the urban center. All in all, even though performance indicators resulting from modeling daily concentration averages will be expectantly lower than those attained for hourly predictions, the methodological approach here presented can be relevant for daily surveillance and alert systems in the Lisbon area.

We start in Sec. 2 by describing the empirical data, comprising the different data sets in the city of Lisbon, Portugal (see Fig. 1). In Sec. 3 we briefly describe NNs models as well as the main points of the circulation-to-environmental approach used and in Sec. 4 the results are discussed in the light of predictive power measures and independent validation of our model is provided. Section 5 concludes the paper.

2. Data

2.1. Target data

We consider daily measurements of PM$_{10}$ concentrations measured by twelve monitoring stations in the agglomeration of Lisbon, between 2002 and 2006 (see bullets in Fig. 1).

The Lisbon agglomeration is covered by a conventional air quality monitoring network composed by traffic, industrial and background monitoring stations which record the atmospheric concentrations of major pollutants, such as gases like NO$_2$, NO and CO and aerosols like PM$_{10}$. This monitoring network for air quality is complemented by three meteorological monitoring stations (see diamonds in Fig. 1), located near the stations of Avenida da Liberdade (AL), Lavradio (L) and Olivais (O).

To investigate the presence of seasonal cycles a preliminary data analysis is done, yielding the box-plot in Fig. 2(a), showing the monthly distribution of pollutant’s concentrations throughout the year for the entire studied period (2002-2006).
Variables (Lag=1 day)

- Mean concentration of NO$_2$, NO, CO, PM$_{10}$
- Maximum concentration of PM$_{10}$ (PM$_{10}$ m)
- Concentration of PM$_{10}$ at 0h UTC
- Daily circulation weather type (CWT)
- Boundary layers heights:
  - (BLH5) at 03:00 UTC
  - (BLH7) at 09:00 UTC
  - (BLH11) at 21:00 UTC
- Daily maximum temperature (Tmax)
- Daily mean wind direction ($V_d$) and intensity ($V_i$)
- Daily mean humidity (Hum) and radiance (Rad)

Table 1: Input parameters used for training the NN (see text).

and for all the monitoring stations. Figure 2(b) supplements the previous box-plot analysis just for Avenida da Liberdade (AL) monitoring station: no clear annual cycle can be drawn.

Daily legal limits were often exceeded during the 2002-2006 period in all the monitoring stations [43, 46], but the number of days with exceeding values is specially impressive for Avenida da Liberdade (AL) and Entrecampos (E) stations. It is worth mentioning that, in both stations two types of exceedences occur. On one hand, the daily legal limit (50 µg/m$^3$) is exceeded, but the number of times that the daily limit can be exceeded per year (35 exceedences/year) is also surpassed [43, 46].

2.2. Input data for NN training

The NN input data sets are shown in Table 1 and consist of daily concentration measurements of several pollutants besides PM$_{10}$ (the target), namely NO$_2$, NO, and CO. Additionally to the pollutant’s concentration measured on the previous day and at 0h UTC (Universal Time Coordinated), several meteorological variables measured in the 3 monitoring stations available were considered (Table 1).

In order to include information regarding the atmospheric stability and circulation, which is an important factor for the accumulation of pollutants near the surface, two other variables were considered, namely the boundary layer and the daily CWT. The boundary layer height (BLH) fields were retrieved from the ECMWF 40 years reanalysis [47] for the years 2002-2006.

Afterward, we extracted the 03:00 UTC (BLH5), 9:00 (BLH7) and 21:00 UTC (BLH11) data from the retrieved BLH fields. The CWT classification was determined for Portugal according to Trigo and DaCamara [48] as described in Sec. 3.1. Values for daily mean sea level pressure (SLP), relative humidity and temperature and geopotential height at the 1000 hPa level values were extracted from ERA Interim Reanalyses dataset [49] on a grid of 1° latitude by 1° longitude for Portugal (40W-30E, 20-70N). The period between 1981 and 2010 was used to perform a 30 year climatology that included the air quality period under analysis (2002-2006). Based on the large-scale fields, prevailing CWTs at regional scale were determined using the simple Geostrophic approximation according

Figure 2: (a) Monthly mean distributions of PM$_{10}$ concentrations for the years 2002 till 2006 in Lisbon; (b) PM$_{10}$ concentrations for the period 2002-2006 recorded at Avenida da Liberdade (AL) monitoring station. The light grey line represents PM$_{10}$ daily measures, the black line represents the 7 days moving average and the horizontal line refers to the PM$_{10}$ daily legal limit (50µg/m$^3$).
to the methodology proposed by Trigo and DaCamara [48]. The daily CWTs resulting from the classification procedure were then considered as an input variable.

In total there are 15 variables that are available as input data for the NN model. Table 1 summarizes the input training data.

Based on the available five years datasets, we constructed a collection of records, consisting of the input vector, which included the meteorological variables, air pollutant concentrations, and the corresponding target PM$_{10}$. All the data used refers to the period between 1/1/2002 and 31/12/2006. The first four years were used to construct the models and the year 2006 was used for independent evaluation (see Sec 4.3).

3. Methods

3.1. Circulation-to-environmental approach

The concentration of pollutants in the atmosphere are linked to the occurrence of certain synoptic weather conditions [33] and to the regional wind flow pattern induced by mesoscale meteorological processes (land-sea breezes) [33]. CWT dictates the long-range transport, linking a particular air mass to dispersion conditions and also to the mesoscale meteorological configuration that controls the regional transport of air pollution [33]. Considering the capabilities of this approach, these prevailing circulation patterns have witnessed a growing interest by the research community during the last decades [33, 48]. The aim of these studies varies considerably, ranging from applications to climatic variability, including trends and extreme years, to environmental purposes and also to access weather driven natural hazards.

CWTs objective classification has successfully been applied to Portugal mainland by Trigo and DaCamara [48], who linked CWTs to precipitation. Pereira et al. [50] and Ramos et al. [51] analysed the impacts of atmospheric circulation, respectively, on fire activity and on lightning activity over Portugal. There are other studies within the Iberian Peninsula, most of them focusing on climatic trends [52], associated to extreme events or to extreme years [53, 54].

The majority of CWT classification procedures are based on the application of statistical selection rules (e.g. cluster analysis, regression trees), but can also be based on the determination of physical parameters regarding the prevailing atmospheric circulation pattern. Furthermore, CWTs are generally specific to a given region, resulting from the examination of synoptic weather data (e.g. sea level pressure (SLP) or geopotential height at 500 hPa) [51].

In this paper, prevailing CWTs calculated according to Trigo and DaCamara [48] are considered as a potential predictor.

3.2. Predictors choice

A crucial step in the development of a forecast model is the choice of input parameters, the predictors [22]. Predictors can be fed into a model for training and evaluation in numerous ways. Usually, a number of statistical methods can be applied in order to choose the most appropriate set of predictors/inputs. Important methods in this scope are stepwise regression (SR), principal component analysis (PCA), cluster analysis and ARIMA [53]. These methods are pre-processing procedures, which allow reducing the number of input variables into the models, thus eliminating redundant information, instabilities and over-fitting.

Here, the selection of variables was made independently for each monitoring station through a forward stepwise regression (FSR), from which the best time lag for each input variable was also determined. During this procedure, which starts with the variable most correlated with the target, new variables are added which, together with the old one(s), most accurately predicts the target [45]. The procedure stops when any new variable does not significantly reduce the prediction error. Significance is measured by a partial F-test applied at 5% [53, 55].

3.3. The Neural Network framework

Neural networks (NN) are mathematical models inspired by the biological nervous system [13, 28, 53], since they are composed by a number of interconnected entities, the artificial neurons, which are similar in several ways to biological neurons.

One of the most common examples of architectures used is the multilayer perceptron [28, 56], where the artificial neurons can be organized following different types of architectures, composing a certain number of levels (Fig. 3) [28, 56]. In the zero level one has the set of independent variables, $X_i$, and a number of connections with a weight $\omega_{ij}$, joining the variables $X_i$ to neurons in the next level [55, 57]. In the first level (“input layer” in Fig. 3), each neuron computes a linear combination of the weighted inputs $\omega_{ij}$, including a bias term $b_j$: $Y_j = \sum \omega_{ij}X_i + b_j$. This sum is transformed using a linear or non-linear function, $W_j = f(Y_j)$. The weights can be initially randomly chosen, and are then properly tuned during the training of the NN as described below. The bias term is included in order to allow the activation functions to be offset from zero and it can be set randomly or set to a desired value, such as a dummy input with a magnitude equal to one.

The output $W_j$ obtained at the previous level is then passed as an input to other nodes in the following layer, usually named hidden layer. This procedure is performed repeatedly to better tune the weights until a certain accuracy threshold between the produced output and the target variable (empirical data) is reached. It is possible to use several hidden levels, successively. However, it is often advantageous to minimize the number of hidden nodes and layers, in order to improve the generalization capabilities of the model and also to avoid over-fitting [55].

There are several training procedures for estimating the weights and associate input and output. Here, we use a modified version of the back-propagation (BP) procedure, which is one of the most popular and common training procedures, see e.g. [56, 57]. As any other training algorithm, BP has drawbacks. One is that the convergence may be slow and the final weights may be trapped in local minima over the highly complex error surface [56, 57]. To overcome this shortcoming, numerically optimized techniques have been developed, such as the Levenberg-Marquardt method (LM) which are based on an approximation of the Gauss-Newton method. The LM method.
Application of the NN framework

The NN framework is applied to our data set in the following way. Consider an attribute \( Z(x, t) \), symbolizing the concentration of PM\(_{10} \), measured at a spatial location \( x \) at day \( t \), which yields a daily series of the pollutant’s concentration at each monitoring station.

3.4. Application of the NN framework

The NN framework is applied to our data set in the following way. Consider an attribute \( Z(x, t) \), symbolizing the concentration of PM\(_{10} \), measured at a spatial location \( x \) at day \( t \), which yields a daily series of the pollutant’s concentration at each monitoring station.

One considers then its decomposition into a periodic component \( M(x, t) \) and a residual \( R(x, t) \), yielding \( Z(x, t) = M(x, t) + R(x, t) \).

In particular we consider the periodic components \( M(x, 7) \), which is determined respectively by a 7 days moving average. Likewise, we take also the respective residuals \( R(x, 7) \) obtained from the removal of the correspondent periodic component.

We then apply the linear and non-linear models, i.e. MLR and NN models, to each monitoring station \( x \) in order to model both the complete signal, hereafter called TOT approach, and to model the residual components, hereafter called RES approach. The forecasting capabilities of the different approaches are compared in order to assess the potential improvement using non-linear NN in air quality modeling.

The NN models used here are based on a feed-forward configuration of the so-called multilayer perceptron \([29]\), sketched in Fig. 3 that has been used by several authors \([22, 23, 29, 57]\). We tested large number of architectures, each one with a given number of layers. The use of two layers was verified to be sufficient, since a superior number of layers does not improve the output.

For the linear model MLR, a perceptron with a linear activation function was used, while for the non-linear NN models, the log-sigmoid function was used, except for the single node in the output layer, for which we consider a linear transfer function.

For the linear regression model \( \text{MLR} \), a perceptron with a linear activation function was used, while for the non-linear NN models, the log-sigmoid function was used, except for the single node in the output layer, for which we consider a linear transfer function.

The MLR models are trained with the LMS rule and the NN models with the Levenberg-Marquardt method. In both cases, a cross-validation is applied with the available period being divided into four times and the calibration-validation procedure is completed four times independently, i.e. each time three years are used for construction and the remaining year is retained for validation. Further, a moving window is applied. Thus, the first run is performed using data for 2002–2004 to train the model, whereas data from 2005 is used for validation purposes. In the second run, data for 2003–2005 are used for training and data for 2002 for validating, and so on.

With such cross-validation procedures \([45]\), it is possible to account for the risk of over- or underfitting. Moreover, in this way, one is able to ascertain if the models are stable and if they are capable of generalizing correctly in forecast mode. After models calibration and validation with historical data (2002–2005), the models are used to produce forecasts for the daily average of PM\(_{10} \) concentration, during a period of one year. For this purpose an independent one-year sample, the year 2006, is left out in order to be used for evaluation of models performance (see Sec. 3.5) during the individual daily average predictions.

In the end, the forecasts are then compared with the actual observed pollutant values at the monitoring stations.

3.5. Performance indicators

Rigorous quantitative measures are required to perform models evaluation. Thus, in order to evaluate the efficiency and performance of the developed models three continuous performance indicators are used. The simplest measure is the the Pearson
correlation coefficient (PC):

\[ PC = \frac{\sum_{i=1}^{N}(y_i - \bar{y})(o_i - \bar{o})}{\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2 \sum_{i=1}^{N}(o_i - \bar{o})^2}}. \quad (1) \]

where \( y_i \) denotes the respective model forecast at time \( i \) while \( o_i \) denotes the real observed values at time \( i \), and \( \bar{y} \) and \( \bar{o} \) are the corresponding average values.

A quantity similar to PC, also related to correlation between series, is the root mean square error (RMSE) given by

\[ \text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N}(y_i - o_i)^2}. \quad (2) \]

Considering that correlation coefficients are not robust to deviations from linearity, its exclusive use to evaluate the quality of a model can lead to misleading results. Therefore we consider these quantities combined with other properties which present different abilities for accessing important aspects of the data such as outliers and average values.

The skill against persistence, SSP, which is interpreted as the percentage of improvement that our model can provide when compared with the persistence model \([45, 57]\), i.e. the model that yields the observed value of yesterday as the forecast for today. The score is quantitatively defined as

\[ \text{SSP} = \frac{1}{N} \sum_{i=1}^{N}(y_i - o_i)^2 - \frac{1}{N-1} \sum_{i=1}^{N}(o_{i+1} - o_i)^2 \quad (3) \]

The SSP is also used as a measure of the relative accuracy of the model.

Both linear and non-linear models will be compared with this persistence model, which is the simplest way of producing a forecast and assumes that the conditions at the time of the forecast will not change. Due to a certain level of memory that characterizes air pollutants, persistence corresponds to a benchmark model considerably more difficult to beat than climatology or randomness.\([33]\).

Additionally, four categorical measures are also considered, to ascertain if the models are able to predict exceedances.\([25]\). Traditional categorical metrics used in model evaluations assess the models ability to predict an exceedance which is defined by a fixed threshold. These metrics are defined by sets of observational forecasts that are paired together. Here, we used the false alarm rate (F), i.e. the proportion of non-occurrences incorrectly forecasted, and the proportion of correctness (PCS), i.e. the proportion of events properly forecasted. Both categorical measures, F and PCS, are applied against binary time series obtained with thresholds for poor air quality limit values (PM\(_{10}\), 50 \(\mu\text{g/m}^3\)), defined by the Portuguese National Environmental Agency. One should however notice that there is now considerable evidence that daily hospital admissions for cardiorespiratory diseases are linked to levels of PM\(_{10}\) not only on the same, but also on previous days\([5]\) and that association is positive for values lower than the legal thresholds. Thus, two additional categorical measures were introduced in order to assess if the models are able to perform correctly for a new threshold that corresponds to 50% of the legal limit value (F50 and PCS50).

### 4. Results and Discussion

#### 4.1. Selection of input variables

We first consider all 15 potential predictors for PM\(_{10}\) (see Table\([1]\)). The use of the FSR has reduced the complexity by retaining substantially less variables, namely only those marked in Table\([2]\).

We also found that adding time lags superior to one day do not provide relevant additional information. Therefore, only the one-day time lags for both meteorological and air quality variables are taken into account in the subsequent analysis.

Our analysis further revealed that the most significant variable in predicting PM\(_{10}\) for all the monitoring stations is the 0h UTC PM\(_{10}\) concentration.

Other variables that were retained for the majority of the sations under the TOT approach are the previous day average and maximum PM\(_{10}\) concentrations, the previous day average values of NO\(_2\), NO and CO concentrations, the maximum temperature, wind direction, humidity and BLH. The other variables retained for the majority of the sations under the RES-7 approach are the previous day average values of NO\(_2\) and CO concentrations, the maximum temperature, wind direction, humidity, CWT and BLH. Additionally, the RES-7 approach uses considerably less pollutant related predictors than the TOT approach, including also the CWT classification as one of the most important predictors in the majority of the monitoring stations.

While the dependence on the wind, relative humidity and BLH were also shown in previous works\([22, 33]\), the NO\(_2\) and CO dependence is present due to road traffic influence, as road traffic behaves as a local source of PM\(_{10}\)\([33]\).

Kukkonen and co-workers\([59]\) showed that the inclusion of meteorological variables for the day of prognosis improves the performance of NN models and that linear models perform significantly worse in this situation. However, we consider that these variables might unnecessarily increase the error associated to the prediction and choose not to include them at this stage.

#### 4.2. Comparison of methods

The validation tests presented here are based on the use of MLR and NN models in which all the retained predictor variables are incorporated according to Table\([2]\) framework. Validation results obtained with the MLR and NN models are shown in Table\([3]\). The numbers of hidden neurons applied are identified by the numeric index after NN, i.e., NN2 refers to a NN model with 2 neurons in the hidden layer. The choice of the number of hidden units was made iteratively. There are four main conclusions to be drawn from Table\([3]\):

1. All the models perform substantially better than persistence with SSP scores above 45%.
2. The proportion of correctness (PCS and PCS50) are quite high which indicate that the models are robust and able to correctly predict not only medium values but also events with high values.
3. The false alarm rate (F) is significantly low for high values, which indicates that the only a low percentage of pollution episodes are not correctly predicted relatively to the legal limit.

4. Weekly residuals (RES-7) models outperform the TOT models. Removing the weekly cycle appears to be a promising approach compared to the complete signal model (TOT).

5. The RES-MLR model performs approximately the same than RES-NN2 and RES-NN3, and considerably better than TOT-MLR models. The similitude between RES-NN and RES-MLR results that it looks doubtful that there is a significant advantage in using the NN model comparatively to the MLR in the present case.

These findings altogether indicate that there is no significant advantage in the use of NN against MLR. Henceforth, we will restrict the remaining analysis to the MLR approach.

From the operational point of view, the effectiveness of a prediction model should be judged according to its ability to forecast properly in order to be able to alert the population and the competent health authorities. However, the forecast models are known a priori to be imperfect, thus the alert threshold must be set below the critical level objectively identified, in order to allow for a margin of safety [18].

Still, the performance indicators presented here are superior to those obtained by Demuzere and co-workers [33] for the Netherlands, and are consistent with the results presented by Hooyberghs and colleagues [22] for Belgium. Demuzere et al. [33] presented a performance of $PC = 0.648$ for particulates and SSP lower values achieved through RES-MLR. Hooyberghs et al. [22] presented results of $PC$ between 0.65 and 0.80 for PM$_{10}$. Here we attain similar results, $0.75 < PC < 0.81$, by incorporating meteorological variables. Moreover, checking the performance results, one observes a tendency for higher performance for the independent validation, which is due to the favourable characteristics of year 2006, as we explain in the next section.

### 4.3. Forecast: Independent validation

The forecasts retrieved by the MLR models were compared with the actual observed pollutants values of the year 2006 at the monitoring stations. The scatter plots and correlation coefficients between observed and modelled values were computed for all monitoring stations. Figure 4 presents the aggregated scatter plots and correlation coefficient for all monitoring stations. The results for the independent sample show a very high average correlation ($PC > 0.84$) between the predicted and observed values.

In Table 4 the correlation coefficients for each individual monitoring station for the calibration/validation period (2002-2005) and for the one-year independent sample (2006) are presented. These results show that MLR model generalizes well.
Table 3: Average performance indicators obtained for the PM$_{10}$ calibration/validation process, including the Pearson correlation coefficient (PC), the skill against persistence (SSp (%)), the coefficient of efficiency (CE), the root mean square error (RMSE (µgm$^{-3}$)), the false alarm rate (F (%)), the proportion of correctness (PCS (%)), and the 50% false alarm rate (F50 (%)) and the 50% proportion of correctness (PCS50 (%)). Each average performance indicator was determined based on the indicators of all the monitoring stations.

| Model     | PC   | SSp  | RMSE | F   | PCS | F50 | PCS50 |
|-----------|------|------|------|-----|-----|-----|-------|
| TOT-MLR   | 0.75 | 45.00| 12.85| 6   | 88  | 50  | 80    |
| RES-7-MLR | 0.81 | 54.41| 11.69| 12  | 86  | 62  | 89    |
| RES-7-NN2 | 0.81 | 54.30| 11.69| 11  | 85  | 64  | 90    |
| RES-7-NN3 | 0.81 | 54.20| 11.69| 11  | 85  | 63  | 90    |

Table 4: Correlation coefficients between observed and modelled PM$_{10}$ concentrations for each station considered and for the calibration/validation period (2002-2005) and for the independent forecast year (2006).

| Station | 2002-2005 | 2006 | %Δ  |
|---------|-----------|------|-----|
| E       | 0.83      | (0.78)| -5  |
| O       | 0.79      | (0.86)| 7   |
| AL      | 0.81      | (0.82)| 1   |
| L       | 0.83      | (0.86)| 3   |
| ESC     | 0.80      | (0.83)| 3   |
| R       | 0.79      | (0.87)| 8   |
| LAR     | 0.85      | (0.87)| 2   |
| LRS     | 0.83      | (0.87)| 4   |
| CC      | 0.75      | (0.78)| 3   |
| QM      | 0.83      | (0.86)| 3   |
| MM      | 0.82      | (0.86)| 4   |
| OD      | 0.85      | (0.82)| -3  |

5. Conclusions

In this paper we introduce a framework consisting in a preselection procedure of predictors which are then used as input data to train NN model. In order to assess the importance of the periodic and residual components present in pollutants time series, the application of linear (MLR) and non-linear (NN) models to each monitoring station was performed. The forecasting capabilities of the different approaches were then compared. The approach based on the removal of the weekly cycle presented the best results, comparatively to the use of the complete signal. Moreover, MLR and NN showed similar performances when evaluated by each of the above criteria. Therefore we find it reasonable to conclude that there is no significant advantage on the use of NN against MLR for the case studied.

Linear MLR and non-linear NN models designed to forecast daily average PM$_{10}$ concentrations in Lisbon, Portugal, were used to produce forecasts and hindcasts. The models were calibrated using air quality and meteorological data from 2002 until 2006 taken at 12 pollutant monitoring stations. Our framework enables to rank all given variables, and then select the highly ranked variables as predictors, which were chosen for each monitoring station separately. To rank the variables a forward stepwise regression was used. We found that the most significant variables in predicting PM$_{10}$ are pollutants for independent data and for each monitoring station.

In general, MLR techniques are known to underestimate peak levels. Interestingly, although the MLR model is built using the calibration dataset only, we can observe an increase in accuracy for the majority of the stations when in forecast mode. This may be explained by the characteristics of the historical data used to construct the models: The year 2005, which was included in the construction of the model, is considered an atypical meteorological year, with low wind and high temperatures and with a prolonged drought [53]. On the other hand, the PM$_{10}$ data sets used on this work comprehend the years from 2002 to 2006 in Lisbon. For this location, the years of 2003 and 2005 were particularly outstanding relatively to weather conditions, namely an exceptional heatwave that struck the entire western Europe, in 2003 [60] and one of the most severe droughts of the 20th century occurred in 2005 [53].

Moreover, air pollution is strongly influenced by shifts in the weather. Changes in the temperature, humidity and wind indeed induce changes in the transport, dispersion, and transformation of air pollutants at multiple scales [61]. Therefore, using all the years as individual calibration/validation samples, yields quite disparate skill values on one hand with an average that is significantly below the skill against persistence obtained when using these anomalous years for independent validation of 2006.
related to road traffic emissions and meteorological variables related to atmospheric stability. Particularly for the RES-7 approach, the most significant variables in predicting PM$_{10}$ are, in descending order of importance, the 0h UTC PM$_{10}$ concentration, the previous day average values of NO$_2$ and CO concentrations, the maximum temperature, wind direction, humidity, CWT and BLH. These results emphasize the importance of meteorological variables and of the circulation-to-environment approach to air quality forecast.

In particular, we found that for forecasting PM$_{10}$ in Lisbon, CTW should be taken as input data, though its rank is not particularly high compared with other meteorological data. However, we point out that the ranking of predictors varies considerably from one station to another, since it reflects the diversity of geographical and urban features, such as traffic, industries, distance to the coast. Therefore, a forthcoming approach to urban pollution would be to apply such procedure to a panoply of different pollutants and ascertain which ones are more sensible to synoptic scale circulation and meteorological constraints. Another issue to addressed in a forthcoming study is the interaction between stations.

All in all, the models presented here and the introduced framework are able to produce different results for each monitoring station, which allows a good spatial resolution for Lisbons urban area. Consistent with the performance measures, high pollutants peak values were reproduced in most cases by each model. The simplicity and cost efficiency of these models, associated with their performance capabilities, show to be very promising for urban air quality characterization, allowing further developments in order to produce an integrated air quality surveillance system for the area of Lisbon. Being a general numerical procedure for any given set of measurements, our finding can be easily adapted to other NN models in weather or geophysical forecast. An extension of this work to take into account the correlations between a higher number of measurement stations is planned.

Acknowledgments

The authors acknowledge the Instituto de Meteorologia and the Agência Portuguesa do Ambiente for the meteorological and environmental data, respectively. The authors thank DAAD and FCT for financial support through the bilateral cooperation DRI/DAAD/1208/2013. FR (SFRH/BPD/65427/2009) thanks Fundação para a Ciência e a Tecnologia for financial support, also with the support Ref. PEst-OE/FIS/UI0618/2011. PL thanks BMU (German Environment Ministry) under the project 41V6451.

References

[1] Kolehmainen, M., Martikainen, H., Ruuskanen, J., 2001. Neural networks and periodic components used in air quality forecasting. Atmospheric Environment 35, 815-825.
[2] EEA - European Environment Agency, 2011. Air quality in Europe 2011 report. EEA Technical report, No 12/2011.
[3] EEA - European Environment Agency, 2012. Air quality in Europe 2012 report. EEA Technical report No 4/2012.
[4] EEA - European Environment Agency, 2013. Every breath we take: Improving Air Quality in Europe. EEA Signals 2013 Report. EEA, Copenhagen.
[5] Wong, C., Atkinson, R., Ross Anderson, H., Hedley, A., Ma, S., Chau, P., Tai-Hing, L., 2002. A Tale of Two Cities: Effects of Air Pollution on Hospital Admissions in Hong Kong and London Compared. In Environmental Health Perspectives, 110 (1), 19992009.
[6] Díaz, J., Linares, C., López, C., García-Herrera, R., Trigo, R.M., 2004. Relationship between Environmental Factors and Infant Mortality in Madrid, 1986-1997, Journal of Occupational and Environmental Medicine 6, 768-774.
[7] Middleton, D., 1997. A new model to forecast urban air quality: BOX-URB. Environmental Monitoring and Assessment 52, 31535.
[8] Reich, S., Gomez, D., Dawidowski, L., 1999. Artificial neural network for the identification of unknown air pollution sources. Atmospheric Environment 33, 3045-3052.
[9] Shi, J.P., Harrison, R.M., 1999. Regression modeling of hourly NO$_2$ and NO$_x$ concentrations in urban air in London, Atmospheric Environment 31 (24), 40814094.
[10] Morris, R.E., Meyers, T.C., 1990. Users guide for the Urban Airshed Model. In: Users manual for UAM (CB-IV). EPA-450/4-90-007A, vol. I. US Environmental Protection Agency, Research Triangle Park.
[11] Davis J.M, Nyckha, D., Bailey B., 2000. A comparison of regional oxidant model (ROM) output with observed ozone data Atmospheric Environment, Volume 34, Issue 15, 24132423.
[12] Monteiro, A., Vautard, R., Borrego, C., Miranda, A., 2005. Long-term simulations of photo oxidant pollution over Portugal using the CHIMERE model. Atmos Envir 39 (17) 30893101.
[13] Lueckcn, D.J., Hutzell, W.T., Gipsn, G.L., 2006. Development and analysis of air quality modeling simulations for hazardous air pollutants. Atmospheric Environment, 40 5087-5096.
[14] Sokhi, R.S., San José, R., Kittiwone, N., Fragkou, E., Pérez, J.L., Middleton, D.R., 2006. Prediction of ozone levels in London using the MM5CMAQ modeling system. Environmental Modelling & Software, 21 566576.
[15] Arasa R., Soler M., Ortega S., Olid M., Merino M., 2010. A performance evaluation of MM5/MNEQA/CMAQ air quality modelling system to forecast ozone concentrations in Catalonia. Tethys, 7, 11 - 23.
[16] Dutot, A.L., Rynkiewicz, J., Steiner F.E., Rude, J., 2007. A 24-h forecast of ozone peaks and exceedance levels using neural classifiers and weather predictions. Environmental Modelling and Software 22, 12611269.
[17] Yi, J., Pybizutok, V.R., 2002. A neural network model forecasting for prediction of daily maximum ozone concentration in an industrialized urban area. Environmental Pollution 92 (3), 349357.
[18] Cobourn, W., Dolcine, L., French, M., Hubbard, M., 2000. A comparison of nonlinear regression and neural network models for ground-level ozone forecasting. Journal of the Air & Waste Management Association 50 19992009.
[19] Lal, B., Tripathy, S.S., 2012. Prediction of dust concentration in open cast coal mine using artificial neural network. Atmospheric Pollution Research 3 211-218.
[20] Gardner, M., Dorling, S., 2000. Statistical surface ozone models: an improved methodology to account for nonlinear behaviour. Atmospheric Environment 34, 2134.
[21] Gardner, M., Dorling, S., 2000. Meteorologically adjusted trends in UK daily maximum surface ozone concentrations. Atmospheric Environment, 34, 171176.
[22] Hooyberghs, J., Mensink, C., Dumont, G., Fierens, F., Brasseur, O., 2005. A neural network forecast for daily average PM$_{10}$ concentrations in Belgium. Atmospheric Environment 39, 32793289.
[23] Papanastassiou, D. K., Melas, D., Kioutsioukos I., 2007. Development and Assessment of Neural Network and Multiple Regression Models in Order to Predict PM$_{10}$ Levels in a Medium-sized Mediterranean City, Water, Air and Soil Pollution 182 325334.
[24] Nejadkoorki, F., Baroutian, S., 2012. Forecasting extreme PM$_{10}$ concentrations using artificial neural networks. Int. J. Environ. Res. 6 277-284.
[25] Perez, P., Trier, A., Reyes, J., 2000. Prediction of PM$_{2.5}$ concentrations several hours in advance using artificial networks in Santiago, Chile. Atmospheric Environment 34 11891196.
[26] Kukkonen, J., Partanen, L., Karppinen, A., Ruuskanen, J., Junninen, H., Kolehmainen, M., Niska, H., Dorling, S., Chatterton, T., Foxall, R., Caw-

ley, G., 2003. Extensive evaluation of neural network models for the prediction of NO₂ and PM₁₀ concentrations, compared with a deterministic modeling system and measurements in central Helsinki. Atmospheric Environment 37, 45394550.

[27] Perez, P., 2001. Prediction of sulfur dioxide concentrations at a site near downtown Santiago, Chile. Atmospheric Environment 35 09294935.

[28] Aguirre-Basurko, E., Ibarra-Berasteig, G., Madariaga, I., 2006. Regression and multilayer perceptron-based models to forecast hourly O₃ and NO₂ levels in the Bilbao area. Environmental Modelling & Software 21 430446.

[29] Russo, A., Raischel, F., Lind, P.G., 2013. Air quality prediction using optimal neural networks with stochastic variables. Atmospheric Environment 79 822-830 (2013).

[30] Weisberg, S., Applied Linear Regression, (John Wiley & Sons. New York, 1985).

[31] Dayan, U. and Levy, I., 2002. Relationship between synoptic-scale atmospheric circulation and ozone concentrations over Israel J. Geophys. Res. 107 D24, 4813.

[32] Baumbach, G., 1996. Air Quality Control, Springer-Verlag.

[33] Demuzere, M., Trigo, R., Arellano, V., van Lipzig, N., 2009. The impact of weather and atmospheric circulation on O₃ and PM₁₀ levels at a rural mid-latitude site. Atmospheric Chemistry and Physics 9, 26952714.

[34] Carvalho, J., Raischel, F., Haase, M., Lind, P.G., 2011. Evaluating strong measurement noise in data series with simulated annealing method. J. Physics Conf. Ser. 285 012007.

[35] Pearce, J., Beringer, J., Nicholls, N., Hyndman, R.J., Uotila, P., Tapper, N.J., 2011. Investigating the influence of synoptic-scale meteorology on air quality using self-organizing maps and generalized additive modeling. Atmospheric Environment 451, 128-136.

[36] Saavedra, S., Rodriguez, A., Taboada, J.J., Souto, J.A., Casares, J.I., 2012. Synoptic patterns and air mass transport during ozone episodes in northwestern Iberia. Sci SET Environ. 441 97-110.

[37] Raischel, F., Russo, A., Haase, M., Kleinhans, M., Lind, P.G., 2012. Searching for optimal variables in real multivariate stochastic data. Physics Letters A 376, 2081.

[38] Dayan, U., Levy, I., 2004. The Influence of Meteorological Conditions and Atmospheric Circulation Types on PM₁₀ and Visibility in Tel Aviv. Journal of Appl. Meteorology 44 606-619.

[39] Studman, J.R., King, K., Holland, M., Walton, R., 2002. Quantification of the health effects of air pollution in the UK for revised PM₁₀ objective analysis. Report, DEFRA, AEAT/ENV/R/1162 Issue 1.

[40] Check [http://ec.europa.eu/environment/air/quality/standards.htm]

[41] Gimeno, L., Nieto, R., Trigo, R. M., Vicente-Serrano, S. M., Lopez-Moreno J.I., 2010. Where does the Iberian Peninsula moisture come from? An answer based on a Lagrangian approach. J. Hydrometeorol., 11, 1241-1236.

[42] Trigo, R., Osborn, T. J., Corte-Real, J., 2002. The North Atlantic Oscillation influence on Europe: Climate impacts and associated physical mechanisms. Climate Res. 20, 917.

[43] APA Agência Portuguesa do Ambiente, 2008. Report Evolution of the Portuguese air quality between the years 2001 and 2005.

[44] Comrie, A.C., 1997. Comparing neural network and regression models for ozone forecasting. Journal of the Air and Waste Management Association 47 653663.

[45] Wilks, D., Statistical Methods in the Atmospheric Sciences (2nd Ed.) No. 59 in International Geophysics (Academic Press, 2006).

[46] APA Agência Portuguesa do Ambiente, 2007. Report Relatório do Estado do Ambiente.

[47] See [http://data-portal.ecmwf.int/data]

[48] Trigo R., DaCamara, C., 2000. Circulation weather types and their impact on the precipitation regime in Portugal. International Journal of Climatology 20, 15591581.

[49] Dee, D.P., and 35 co-authors, 2011. The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. Quart. J. R. Meteorol. Soc. 137, 553-597.

[50] Pereira, M., Trigo, R.M., DaCamara, C., Pereira, J.M.C., Leite, S., 2005. Synoptic patterns associated with large summer forest fires in Portugal. Agricultural and Forest Meteorology 129, 1125.

[51] Ramos, A.M., Ramos, R., Sousa, P., Trigo, R.M., Janeira, M., Prior, V. 2011. Cloud to ground lightning activity over Portugal and its association with Circulation Weather Types. Atmospheric Research 101, 84-101.

[52] Lorenzo, M.N., Tabouda, J.J., Gimeno, L., 2008. Links between circulation weather types and teleconnection patterns and their influence on precipitation patterns in Galicia (NW Spain). Int J Climatol 28 11, 14931505, doi:10.1002/joc.1646.

[53] Garcia-Herrera, R., Paredes, D., Trigo, R., Trigo, I., Hernandez, E., Barriopedro, D., Mendes, M., 2007. The outstanding 2004/05 drought in the Iberian Peninsula: Associated atmospheric circulation. Journal of Hidrometeorology 8, 483498.

[54] Vicente-Serrano, S.M., Trigo, R.M., López-Moreno, J.I., Liberato, M., Lorenzo-Lacruz, J., Begueria, S., Morán-Tejeda, E., Kenawy, A., 2011. Extreme winter precipitation in the Iberian Peninsula in 2010: anomalies, driving mechanisms and future projections. Clim Res 46, 5165, doi: 10.3354/cr00977.

[55] Gardner, M.W., Dorling, S.R., 1998. Artificial neural networks (the multilayer perceptron) - a review of applications in the atmospheric sciences, Atmospheric Environment 32 (14/15), 26272636.

[56] Haykin, S., Neural Networks - A Comprehensive Foundation 2 Ed. MacMillan (College Publishing Company, New York, 1999).

[57] Trigo, R., Palutikof, J., 1999. Simulation of daily temperatures for climate change scenarios over Portugal: a neural network model approach. Climate Research 13 4539.

[58] Legates, D.R., McCabe, G.J., 1999. Evaluating the use of goodness-of-fit Measures in hydrologic and hydroclimatic model validation. Water Resources Research, 35, 1, 233241.

[59] Kukkonen, J., Partanen, L., Karrpinen, A., Ruuskanen, J., Junninen, H., Kolehmainen, M., Niska, H., Dorling, S., Chatterton, T., Foxall, R., Cawley, G., 2003. Extensive evaluation of neural network models for the prediction of NO₂ and PM₁₀ concentrations, compared with a deterministic modeling system and measurements in central Helsinki. Atmospheric Environment 37, 45394550.

[60] Trigo R.M., Pereira J.M.C., Pereira M.G., Mota B., Calado M.T., DaCamara C.C., Santo F.E., 2006. Atmospheric conditions associated with the exceptional fire season of 2005 in Portugal. International Journal of Climatology 26, 13, 1741-1757.

[61] Dias, D., Tchepel, O., Carvalho, A., Miranda, A.I., Borrego, C., 2012. Particulate matter and health risk under a changing climate: assessment for Portugal. Scientific World Journal 2012, 409546.